frontiers | Research Topics

Big data and machine learning in sociology

Edited by
Heinz Leitgöb, Tobias Wolbring and Dimitri Prandner

Published in
Frontiers in Sociology
Frontiers in Big Data

ISSN 1664-8714
ISBN 978-2-8325-2514-2
DOI 10.3389/978-2-8325-2514-2

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Big data and machine learning in sociology

Topic editors

Heinz Leitgöb — Leipzig University, Germany

Tobias Wolbring — University of Erlangen Nuremberg, Germany

Dimitri Prandner — Johannes Kepler University of Linz, Austria

Citation

Leitgöb, H., Wolbring, T., Prandner, D., eds. (2023). *Big data and machine learning in sociology*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-2514-2

Table of contents

frontiers | Frontiers in Sociology

TYPE Editorial
PUBLISHED 09 May 2023
DOI 10.3389/fsoc.2023.1173155

OPEN ACCESS

EDITED AND REVIEWED BY
Scott Schaffer,
Western University, Canada

*CORRESPONDENCE
Heinz Leitgöb
heinz.leitgoeb@uni-leipzig.de

RECEIVED 24 February 2023
ACCEPTED 13 April 2023
PUBLISHED 09 May 2023

CITATION
Leitgöb H, Prandner D and Wolbring T (2023) Editorial: Big data and machine learning in sociology. *Front. Sociol.* 8:1173155. doi: 10.3389/fsoc.2023.1173155

Editorial: Big data and machine learning in sociology

Heinz Leitgöb[1,2]*, Dimitri Prandner[3] and Tobias Wolbring[4]

[1]Institute of Sociology, Leipzig University, Leipzig, Germany, [2]Institute of Sociology, University of Frankfurt, Frankfurt, Germany, [3]Institute of Sociology, University of Linz, Linz, Austria, [4]Institute of Labour Market and Socioeconomics, University of Erlangen-Nuremberg, Nuremberg, Germany

KEYWORDS

big data, machine learning, computational social science, digitalization, artificial intelligence, social science research methodology

Editorial on the Research Topic
Big data and machine learning in sociology

Introduction

The dawn of the digital age, aptly characterized by "computers everywhere" (Salganik, 2018, p. 3), has shaped modern societies and, thus, the lives of individuals worldwide in unique ways. The ubiquity of the internet, in conjunction with the mass distribution of a variety of affordable internet-enabled digital devices, has created new possibilities for collecting, storing, linking, sharing, and exchanging information. Also, the massive progress in computer performance regarding processing capacities and computational speed has paved the way for advances in programming which culminated in the recent progress in artificial intelligence (AI) research, referred to as the recent AI spring (for a brief outline of the history of AI research, see, e.g., Mitchell, 2019). Its results are—among others—the deep-learning-induced successes in speech and object recognition that enable processes as complex as simultaneous translation or autonomous driving. The societal consequences range from the emergence of new professions, business fields, leisure activities, behavioral cultures, and associated lifestyles to new social inequalities (digital divide), dependencies (digital and data literacy gaining relevance as key competencies), and forms of deviant/criminal activity (e.g., cyberbullying and -crime, online hate speech, crimes organized/executed through the internet).

This digital revolution affects the social sciences in various ways. First, social processes experience fundamental change and adaption that require extensive scientific elaboration. Second, the steadily increasing application of digital technologies generates an enormous mass of finely granulated data in various forms and formats. It is not just that enormous amounts of data can now be easily accessed and analyzed. Digital innovations have allowed the collection of data in various formats that were previously difficult to compile (e.g., georeferenced data, tracking or process data, intensive longitudinal data, social media text data; Golder and Macy, 2014; Leitgöb and Wolbring, 2021). This digitization and datafication of society have shaped empirical social science research fundamentally in recent years and will continue to do so. Third, the increasing computational power and the maturation of software environments have promoted the development of algorithmic solutions for complex statistical problems. It paved the way for the nascent field of computational social science (CSS; e.g., Lazer et al., 2009, 2020; Edelmann et al., 2020; Engel et al., 2022a,b) at the intersection of the social sciences, statistics, informatics, and mathematics.

The future viability of the empirical social sciences will largely depend on their ability to adapt to the conditions associated with the ongoing digitization of society (Wolbring, 2020). While new digital technologies have provided empirical social research with unique opportunities for data generation and analytical processing, they also impose new methodological challenges that shape research designs, theoretical foundations, and the methods used. For example, using digital process data for scientific purposes requires the development of tailored data and measurement theories, quality criteria, and corresponding quality assurance procedures to establish quality standards comparable to those from survey methodology. Also, this shift in perspectives afflicts the way the obtained data are typically analyzed, raising the question of how to transfer the relevant advancements from computer science to social science methodology (Törnberg and Uitermark, 2021; Jarvis et al., 2022).

Against this backdrop, the Research Topic covers two core elements of CSS, (i) *big data* and (ii) *machine learning*. While this editorial focuses on the big picture, highlighting some key aspects in both areas without purporting to represent a comprehensive review, the research papers published in this Research Topic provide detailed insights into the unfolded content area. We organize the remaining part of the editorial according to the three perspectives typically addressed in the discussion of the impact of digitalization on social science research: the epistemological perspective (Section 2), the data perspective (Section 3), and the data analytical perspective (Section 4).

Epistemological consequences of digitalization

There are multiple competing epistemological concepts in the discussions about CSS (e.g., Törnberg and Uitermark, 2021). While the relevance of data and the potential consequences of "big data" for the social sciences were first addressed long before societal digitization, it was the digitalization wave of the late 20th century that brought the discussion to a broader part of the scientific communities. At the beginning of the millennium, both social scientists and statisticians stated that it is necessary to discuss the impact which computer science had on the emerging CSS and reflect on the consequences of analyzing social phenomena through a "computational paradigm of society" (Törnberg and Törnberg, 2018).

As a naïve starting point, it can be assumed that digital data and their traces are *true* and, thus, exact representations of social processes. As such, digital data would be naturally emerging data representing the real underlying structure of society and social interactions. This view mirrors how computer scientists not necessarily capture but often handle digital trace data in practice: Pursuing a data and performance-driven research agenda, they focus primarily on the algorithmic optimization of predictions by specifying models that are superior to others concerning predictive accuracy but with few concerns regarding the included measures (e.g., by considering selection effects and measurement error). While trying to trace the complex networks and data flows that shape modern societies and economies in much greater detail and to establish causal inferences beyond traditional methods, they tend to be less preoccupied with the data generating process, including aspects of research design or protocol (Allen et al., 2017).

In contrast, a more realistic view would neglect the idea of natural data. As Lazer et al. (2014, p. 1203) highlighted: "quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data". All digital platforms are designed by humans within certain societal constraints to measure and often even monetize social interactions, resulting in structures that potentially manipulate individuals (Mayer-Schönberger and Cukier, 2013; van Dijck, 2014; Couldry and Mejias, 2021). Research has shown that empirical studies can disadvantage minorities or groups of low social status unless they adhere to a strict definition of fairness and justice (e.g., Mitchell et al., 2021) and theoretical reasoning (Mullainathan and Spiess, 2017; Molina and Garip, 2019). Accordingly, big data and AI-driven research need to be embedded into theoretical frameworks and enable transparent discussions about how data are biased. Algorithms can also be sensitive to contextually problematic conceptualizations and depend on interactional settings. This can be highly impactful for the generation and reproduction of social inequalities as "one of the core competencies—and responsibilities—of the social sciences" (Gerdon et al., 2022, p. 2; see also Section 4).

Nevertheless, scholars pursuing these ideas certainly see much benefit in the increased amount of available data, the rich granularity, and new types of measures. Likewise, they are eager to integrate new data sources and methods into their theoretical work, but they will interpret their results more carefully and reflected and deal critically with the limitations of their data. Developing and expanding a social scientific perspective (e.g., Blei and Smyth, 2017) on the implementation of big data and AI-driven analysis into the research processes is an essential complement to the more technical focus of disciplines such as informatics and mathematics, which sociology and related social science disciplines can contribute to the fields of CSS and data science. In the context of this Research Topic, such issues are also at the forefront of several articles examining how good or fair automated classification and decision-making processes can be. The studies of Kuppler et al. (in this volume) and Seewann et al. (in this volume) examined how new methods and techniques could support social scientific work but also expressed their concerns about ethics and limits attached to such methods.

Digitalization and the big data era

The datafication of society is a consequence of the digital revolution. In contemporary societies, individuals leave digital traces in numerous processes, such as communication, mobility, shopping, banking, dating, working, and learning (Lazer et al., 2009; for a review see Golder and Macy, 2014). These digital behavioral data (DBD) increase at an exponential rate (Jarvis et al., 2022, p. 35). Typically, they are collected and processed by institutions such as public administration, non-governmental organizations, and commercial companies. They differ in some relevant respects from scientifically produced data in quantitative social research, such as survey data and experimental data.

First, they differ in size. DBD are available in incredible quantity, allegorized as "data deluge". Second, DBD are

omnipresent, often generated continuously and available in real time. Third, DBD do not represent some homogeneous data type, but differ considerably in form, format and complexity (e.g., dimensionality and structuredness). Their diversity ranges from social media text and respective metadata (Hadler et al.; Schünemann et al.; Schwitter et al. in this volume), social network and interaction data, data from webpages (Seewann et al. in this volume), online consumer behavior data, geocoding (Nguyen et al. in this volume) and time references, physical condition and mobility data, internet search engines results, to information extracted from images and videos. Accordingly, DBD fall under the minimal definition of "big data," typically characterized by the three Vs: (i) huge in *volume*, (ii) high in *velocity*, and (iii) diverse in *variety* (Laney, 2001; Beyer and Laney, 2012).

The systematic use of DBD and other digitalized mass data (e.g., contextual data from ecological systems, large-scale digitalized register, administrative and official statistical data) for scientific purposes marks the beginning of a big data era (e.g., Kitchin, 2014; Connelly et al., 2016) in the social sciences. Many advantages of this development are obvious (for overviews, see, e.g., Golder and Macy, 2014; Adams and Brueckner, 2015; Cesare et al., 2018). Foremost, a tremendous amount of data containing fine-grained and often high-dimensional information about social phenomena at different societal levels, which are impossible to collect with traditional non-digital procedures, is potentially accessible now. What once was a rare commodity in science is now ubiquitous (Golder and Macy, 2014; Salganik, 2018) and is often systematically stored in massive social data archives. However, Connelly et al. (2016, p. 1) argue that it is "not the size or quantity of these data that is revolutionary. The revolution centers on the increased availability of new types of data which have not previously been available for social science research". This allows under-addressed research questions to be answered. And the systematic linkage of DBD, also with various other data sources (e.g., survey, register, official statistics and contextual data, e.g., Christen et al., 2020; Klumpe et al., 2020; Stier et al., 2020), entails additional analytical boost. For example, see the contributions of Hadler et al., Nguyen et al., and Schünemann et al., in this volume. Furthermore, DBD are expected to be less prone to errors induced by reactivity because they are often collected unobtrusively in the background without social interaction with others (e.g., Harari et al., 2017; Salganik, 2018; Diekmann, 2020; Keusch et al., 2022).

However, the scientific use of DBD is also associated with various challenges. DBD are typically produced for administrative, commercial, or other purposes outside the academic field or as the by-product of everyday digital processes. Thus, DBD do not necessarily meet scientific quality standards (Salganik, 2018), and their application in a research context presupposes the critical evaluation of—among others—conceptual fit (Do the observed variables adequately map the theoretical constructs of interest?), measurement quality, and representation to avoid bias that invalidates the conclusions. However, while well-established (missing) data and measurement theories, error models, and relevant quality criteria are readily available for scientific data, this is usually not the case for DBD. The first important contributions to this topic were provided by Hsieh and Murphy (2017), Amaya et al. (2020), Biemer and Amaya (2021), and Sen et al. (2021). Furthermore, rigorous inferences from empirical data greatly benefit from systematically implemented research designs that determine the data-generating process (e.g., Wolbring, 2020). For example, causal effects cannot simply be learned from a joint distribution of observed variables (Pearl, 2010). It also requires theoretical elaboration and a research design that rules out threats to internal validity, such as confounding, endogeneity, and systematic selection. In other words, "design trumps analysis" (Rubin, 2008, p. 808) in causal effect identification. However, the generative process of DBD does not, in principle, rely on such design considerations, limiting their usability for the causal inference task and frequently resulting in very noisy data (e.g., Silver, 2012).

Finally, it is also worth noting that progress in portable digital and sensor technologies offers unique opportunities in academic research to collect DBD about individuals' everyday practices and routines. App-based survey tools allow for the active and passive collection of DBD and their systematic combination with online survey data (e.g., Jäckle et al., 2019; Keusch et al., 2019; Kreuter et al., 2020). For participant recruiting, non-probability samples particularly online access panels are expected to play a decisive role and require extensive investigation (e.g., Cornesse et al., 2020).

The turn in data analysis

Opportunities to collect and use data of previously unknown mass, granularity, and complexity, in new formats and based on non-scientific and unknown data-generating processes require analytical models that adequately address these data characteristics (e.g., Amaturo and Aragona, 2019; Edelmann et al., 2020). In recent years, impressive computer hardware innovations regarding storage capacities, computing power, interconnectedness, task division, and data transmission evoked the development of such computationally intensive statistical software solutions, creating an algorithmic culture of statistical modeling without assuming an underlying stochastic data model as in the traditional statistical modeling culture (Breiman, 2001). This algorithmic culture is strongly affected by machine learning, a field of sub-symbolic AI research dominated by informatics but with substantive roots in statistics (Friedrich et al., 2022).

Machine learning (ML) lacks a precise definition, being "as much a culture defined by a distinct set of values and tools as it is a set of algorithms" (Grimmer et al., 2021, p. 397). Besides processing numerical data, ML algorithms are also developed to process text data. This is demonstrated by some articles in this volume (Haensch et al.; Munnes et al.; Egger and Yu). For a comprehensive overview of the various ML algorithms, see the textbooks of Bishop (2006), Hastie et al. (2009), Goodfellow et al. (2016), Mohri et al. (2018), Sutton and Barto (2018), Jurafsky and Martin (2023), Murphy (2022).

The field is broadly classified into two domains: supervised and unsupervised learning. Although both share the automated extraction of information from data, they differ in their learning objectives. Supervised ML utilizes labeled output data Y and input data X to learn the input-output mapping for predictive and regression purposes. In contrast, the primary purpose of

unsupervised ML is to detect and describe systematic patterns (latent structures) in input data X without labeled output data Y. However, this binary classification of ML approaches is neither disjoint nor exhaustive (Molina and Garip, 2019). While some ML algorithms can be used in both domains, others belong to neither. The latter is—among others—the case for reinforcement learning and some speech and language processing algorithms. Furthermore, some algorithms can be principally assigned to one domain, but contain features from the other. An example is generative adversarial networks (GANs), classified as unsupervised ML models because no human labeling of the input data is required. However, GANs are trained on the principle of self-supervision; that is, the algorithm initiates a data labeling process to solve some classification problems. A typical field of application for GANs is manipulating audio or video material producing deepfakes (Eberl et al. in this volume). It is also worth noting that many algorithms subsumed under the ML paradigm already have a long social science research tradition but are not explicitly designated as an ML application. Prominent examples are linear modeling, hierarchical agglomerative and k-means clustering, k-nearest neighbor algorithms, principal component analysis, and neural network analysis.

As outlined, the primary goal of supervised ML applications is the prediction of $\hat{Y}$ from X. In contrast, the traditional stochastic statistical modeling approach, referred to as "generative modeling" (Donoho, 2017), focuses on parameter estimation. That is, on the generation of $\hat{\beta}$, which represent the estimated effect sizes of the effect of X on Y (Mullainathan and Spiess, 2017). It requires specifying the functional form of the joint distribution of X and Y (Athey and Imbens, 2019). This modeling perspective is in line with the epistemic focus on causal explanation, particularly with the tasks of causal inference and generative mechanism detection (e.g., Gangl, 2010; Hedström and Ylikoski, 2010; Imai et al., 2011; Winship and Morgan, 2015). It leads to "simple and interpretable models" (Molina and Garip, 2019, p. 29) that mimic the data-generating process. These models are based on strict theoretical assumptions, tied to a set of testable propositions (Grimmer et al., 2021). However, ML-based prediction models are much more data hungry (e.g., the simulation study of van der Ploeg et al., 2014) and complex, with up to millions of parameters and more opaque input-output-functions (Grimmer et al., 2021) that "produce black-box results that offer little insight on the mechanism linking the inputs to the output" (Molina and Garip, 2019, p. 29). The primary objective is predictive accuracy maximization in out-of-sample (training data) conditions, provoking data-driven *ad hoc* modeling decisions without substantial theoretical foundation (Radford and Joseph, 2020). This has relevant implications for the applicability of ML algorithms in sociology.

(i) For explanatory purposes, ML modeling strategies require conceptual and technical optimization to generate valid interpretable results that illuminate the generative social mechanisms based on massive amounts of DBD (e.g., the discussion in Radford and Joseph, 2020; Hofman et al., 2021; Breznau, 2022). This includes an adequate construct-measurement match and measurement modeling (Jacobs and Wallach, 2021).

(ii) The data deluge and the availability of data-driven ML algorithms for analytical processing evoked a *debate on the relevance of (social) theory*. The positions range from "the end of theory" and "correlation supersedes causation" proclamations (e.g., Anderson, 2008) to the call for a strong emphasis on theoretical reasoning to counteract technical limitations, problematic assumptions, limited interpretability, and false conclusions (e.g., Radford and Joseph, 2020; Wolbring, 2020). In any case, prominent examples such as the mispredictions of Google Flu Trends (e.g., Butler, 2013; Olson et al., 2013; Lazer et al., 2014) illustrated the demand for a flexible methodological framework with theory, traditional data sources and methods, as well as DBD and algorithmic approaches as complementary elements to be integrated to maximize knowledge gain (Lazer et al., 2014; Schnell, 2019). Also, unsupervised ML algorithms as exploratory tools could contribute to the inductive process of theory development.

(iii) ML algorithms optimized for prediction offer an opportunity to extend the key epistemological goals in sociology. While the prediction task has so far only played a minor role alongside the explanation task (e.g., Chen et al., 2021), its relevance has become particularly evident during the COVID-19 pandemic (e.g., Pavlović et al., 2022). The pandemic situation required predicting the consequences of strict policy measures (e.g., social distancing, the closing of schools, lockdowns) on various aspects of social life (e.g., student learning outcomes, mental health issues, domestic violence, social and economic inequalities, poverty) to support policy decision making (e.g., Jahn et al., 2022). In addition, Watts (2014) argued that the development of theory and causal explanations could also benefit from a stronger focus on prediction in sociology.

(iv) Assessing the quality of (out-of-sample) predictions requires respective performance metrics. Alongside the traditional technical measures (e.g., accuracy, precision, sensitivity, specificity, AUC, e.g., Steyerberg, 2010), increasing importance is attached to "social" metrics. These account for predictive fairness by quantifying the total amount of bias (for a typology of potential biases at the intersections between data, algorithms, and users, see Mehrabi et al., 2021) that causes a diverging predictive performance across and statistical discrimination against specific groups along ascriptive attributes, such as gender, age, and ethnicity. Although several fairness criteria have been developed based on different definitions of fairness (for an overview, see, e.g., Caton and Haas, 2020; Mitchell et al., 2021; Han et al., 2022; Pessach and Shmueli, 2022), additional concepts with respective evaluation criteria are needed to assess the overall social impact of algorithmic predictions on decision-making in detail. Sociology can play a decisive role in developing such a conceptual framework (e.g., Gerdon et al., 2022; Starke et al., 2022). An example is provided by Kuppler et al. (in this volume), advocating a conceptual differentiation between algorithmic fairness and distributive justice.

Outlook

This editorial highlights the digital revolution's impact on social sciences—particularly on empirical sociology—from an epistemological, data, and analytical perspective. In line with the thematic orientation of the Research Topic, it focuses on big data and machine learning, which are two core elements of the nascent

and interdisciplinary field of computational social science (CSS). Building on Lazer et al. (2020) and Leitgöb and Wolbring (2021), we finally share some thoughts on the institutional processes required to establish this computational turn as a sustainable success story.

(i) *Universities need to adopt their institutional structures and facilities to meet the demands*. This includes an organizational restructuring to facilitate interdisciplinary collaboration and the financing of the computational infrastructure mandatory for the storing, linking, and high-speed processing of massive amounts of data under the highest security standards.

(ii) *Social science education needs to be reformed*. In particular, the traditional training in methods and methodology, focusing on survey data and classical frequentist statistics must be supplemented by CSS elements based on mathematics, computational statistics, informatics, and data science to maximize the students' (digital) data literacy. Besides training in gathering, processing, analyzing, and visualizing big digital data with software packages such as R or Python, this also includes conducting simulation studies (e.g., Keuschnigg et al., 2018). The success of implementing these topics in the sociology curricula will determine the future viability of the discipline and the extent to which sociology will play a leading role in CSS.

(iii) *Big centralized data infrastructure needs to be established*. This infrastructure is intended to serve the systematic comprehensive collection, processing, and secure storage of any social science data in accordance with legal data protection standards. The main objective is to provide this data to the scientific community for secondary data analysis. In addition to the financial resources, technical innovations, and know-how, this requires a new culture of willing data and code sharing from the stakeholders such as researchers, universities, public authorities, and social media companies (Lazer et al., 2020).

(iv) *Detailed data protection regulations and ethical guidelines are necessary to establish handling security for researchers*. The progressive digital technologies enable researchers to explore, in principle, entirely new methodological pathways in studying social phenomena and generating empirical evidence for decision-making. However, relevant legal and ethical questions still need to be resolved to legitimize the use of these methodological innovations, especially because many data are sensitive or difficult to anonymize (e.g., Salganik, 2018). While legal data protection frameworks are set in principle in most countries (e.g., by the General Data Protection Regulation, applicable in all European Union member states since 2018), there has been uncertainty about how existing legislation will be handled in practice (for some brief examples, see Leitgöb and Wolbring, 2021). Likewise, a comprehensive set of tailored and broadly accepted ethical standards is still unavailable in this developing field of research (e.g., Hand, 2018; Piano, 2020).

(v) *The application of AI innovations in teaching and throughout the research process needs to be regulated*. Current developments, particularly the distribution of the chatbot software ChatGPT (Generative Pre-trained Transformer),[1] illustrate that AI systems can be used not only for data analysis in the social sciences. Instead, these systems allow a wide range of tasks to be solved throughout the research process and can also be used by lecturers and students in academic training courses. Initial reactions to these innovations range from banning of the use of ChatGPT for students (e.g., at Sciences Po Paris[2]) to active considerations of how to collaborate with generative AI to delegate tasks and maximize knowledge acquisition. In any case, standards should be developed on how to regulate the use of AI systems and how their contribution to scientific work should to be disclosed.

The above aspects outline the key efforts required to provide the CSS agenda with a solid foundation for long-term success. The Research Topic aims to serve as a platform for different contributions to the core elements of CSS: big data and machine learning. Ideally, the Research Topic and its articles encourage further research and contribute to the progress that the digital revolution has brought to social science research methodology.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

1 For details, see https://openai.com/blog/chatgpt (02/08/2023).

2 For details, see https://newsroom.sciencespo.fr/sciences-po-bans-the-use-of-chatgpt (02/08/2023).

References

Adams, J., and Brueckner, H. (2015). Wikipedia, sociology, and the promise and pitfalls of big data. *Big Data Soc.* 2, 1–5. doi: 10.1177/2053951715614332

Allen, J. A., Fisher, C., Chetouani, M., Chiu, M. M., Gunes, H., Mehu, M., et al. (2017). Comparing social science and computer science workflow processes for studying group interactions. *Small Group Res.* 48, 568–590. doi: 10.1177/1046496417721747

Amaturo, E., and Aragona, B. (2019). Methods for big data social sciences. *Math. Popul. Stud.* 26, 65–68. doi: 10.1080/08898480.2019.1597577

Amaya, A., Biemer, P. P., and Kinyon, D. (2020). Total error in a big data world: adapting the TSE framework to big data. *J. Surv. Stat. Methodol.* 8, 89–119. doi: 10.1093/jssam/smz056

Anderson, C. (2008). *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*. Wired. Available online at: https://www.wired.com/2008/06/pb-theory (accessed April 26, 2023).

Athey, S., and Imbens, G. (2019). Machine learning methods that economists should know about. *Annu. Rev. Econom.* 11, 685–725. doi: 10.1146/annurev-economics-080217-053433

Beyer, M. A., and Laney, D. (2012). *The Importance of "Big Data". A Definition*. Stamford: Gartner Research.

Biemer, P. P., and Amaya, A. (2021). "Total error frameworks for found data," in *Big Data Meets Survey Science. A Collection of Innovative Methods*, eds C. A Hill, O. P. Biemer, T. D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov, et al. (Wiley: Hoboken), 133–16.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York, NY: Springer.

Blei, D. M., and Smyth, P. (2017). Science and data science. *Proc. Nat. Acad. Sci. U. S. A.* 114, 8689–8692. doi: 10.1073/pnas.1702076114

Breiman, L. (2001). Statistical modeling: the two cultures. *Stat. Sci.* 16, 199–231. doi: 10.1214/ss/1009213726

Breznau, N. (2022). Integrating computer prediction methods in social science: a comment on Hofman et al. (2021). *Soc. Sci. Comp. Rev.* 40, 844–853. doi: 10.1177/08944393211049776

Butler, D. (2013). When Google got flu wrong: US outbreak foxes a leading web-based method for tracking seasonal flu. *Nature* 494, 155–156. doi: 10.1038/494155a

Caton, S., and Haas, C. (2020). Fairness in machine learning: a survey. *arXiv*. arXiv:2010.04053v1

Cesare, N., Lee, H., McCormick, T., Spiro, E., and Zagheni, E. (2018). Promises and pitfalls of using digital traces for demographic research. *Demography* 55, 1979–1999. doi: 10.1007/s13524-018-0715-2

Chen, X., Wu, X., Hu, A., He, G., and Ju, G. (2021). Social prediction: a new research paradigm based on machine learning. *J. Chin. Sociol.* 8, 1–21. doi: 10.1186/s40711-021-00152-z

Christen, P., Ranbaduge, T., and Schnell, R. (2020). *Linking Sensitive Data. Methods and Techniques for Practical Privacy-Preserving Information Sharing*. Springer: Cham.

Connelly, R., Playford, G. V., and Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Soc. Sci. Res.* 59, 1–12. doi: 10.1016/j.ssresearch.2016.04.015

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., et al. (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *J. Surv. Stat. Method.* 8, 4–36. doi: 10.1093/jssam/smz041

Couldry, N., and Mejias, U. A. (2021). The decolonial turn in data and technology research: what is at stake and where is it heading? *Inf. Commun. Soc.* 26, 1–17. doi: 10.1080/1369118X.2021.1986102

Diekmann, A. (2020). "Die Renaissance der "Unobstrusive Methods" im digitalen Zeitalter," in *Grundlagen – Methoden – Anwendungen in den Sozialwissenschaften. Festschrift für Steffen-M. Kühnel*, eds A. Mays, V. Dingelstedt, S. Hambauer, F. Schlosser, J. Berens, J. Leibold, et al. (Wiesbaden: Springer VS), 161–172.

Donoho, D. (2017). 50 years of data science. *J. Comput. Graph. Stat.* 26, 745–766. doi: 10.1080/10618600.2017.1384734

Edelmann, A., Wolff, T., Montagne, D., and Bail, C. A. (2020). Computational social science and sociology. *Annu. Rev. Sociol.* 46, 61–81. doi: 10.1146/annurev-soc-121919-054621

Engel, U., Quan-Haase, A., Liu, S. X., and Lyberg, L. (eds.). (2022a). *Handbook of Computational Social Science. Volume I: Theory, Case Studies, and Ethics*. Routledge: London.

Engel, U., Quan-Haase, A., Liu, S. X., and Lyberg, L. (eds.). (2022b). *Handbook of Computational Social Science. Volume II: Data Science, Statistical Modelling, and Machine Learning Methods*. Routledge: London.

Friedrich, S., Antes, G., Behr, S., Binder, H., Brannath, W., Dumpert, F., et al. (2022). Is there a role for statistics in artificial intelligence? *Adv. Data Anal. Classif.* 16, 823–846. doi: 10.1007/s11634-021-00455-6

Gangl, M. (2010). Causal inference in sociological research. *Annu. Rev. Sociol.* 36, 21–47. doi: 10.1146/annurev.soc.012809.102702

Gerdon, F., Bach, R. L., Kern, C., and Kreuter, F. (2022). Social impacts of algorithmic decision-making: a research agenda for the social sciences. *Big Data Soc.* 9. doi: 10.1177/20539517221089305 [Epub ahead of print].

Golder, S. A., and Macy, M. W. (2014). Digital footprints: Opportunities and challenges for online social research. *Annu. Rev. Sociol.* 40, 129–152. doi: 10.1146/annurev-soc-071913-043145

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.

Grimmer, J., Roberts, M. E., and Stewart, B. M. (2021). Machine learning for social science: an agnostic approach. *Ann. Rev. Polit. Sci.* 24, 395–419. doi: 10.1146/annurev-polisci-053119-015921

Han, X., Shen, A., Cohn, T., Baldwin, T., and Frermann, L. (2022). "Systematic evaluation of predictive fairness," in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, Vol 1* (Stroudsburg, PA), 68–81.

Hand, D. J. (2018). Aspects of data ethics in a changing world: where are we now? *Big Data* 6, 176–190. doi: 10.1089/big.2018.0083

Harari, G. M., Müller, S. R., Aung, M. S., and Rentfrow, P. J. (2017). Smartphone sensing methods for studying behavior in everyday life. *Curr. Opin. Behav. Sci.* 18, 83–90. doi: 10.1016/j.cobeha.2017.07.018

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction, 2nd Edn*. New York, NY: Springer.

Hedström, P., and Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annu. Rev. Sociol.* 36, 49–67. doi: 10.1146/annurev.soc.012809.102632

Hofman, J. M., Watts, D., J., Athey, S., Garip, F., Griffiths, T. L., et al. (2021). Integrating explanation and prediction in computational social science. *Nature* 595, 181–188. doi: 10.1038/s41586-021-03659-0

Hsieh, Y. P., and Murphy, J. (2017). "Total Twitter error: Decomposing public opinion measurement on Twitter from a total survey error perspective," in *Total Survey Error in Practice*, eds P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, et al. (Hoboken: Wiley), 23–46.

Imai, K., Keele, L., Tingley, D., and Yamamoto, T. (2011). Unpacking the black box of causality: learning about causal mechanisms from experimental and observational studies. *Am. Polit. Sci. Rev.* 105, 765–789. doi: 10.1017/S0003055411000414

Jäckle, A., Burton, J., Couper, M. P., and Lessof, C. (2019). Participation in a mobile app survey to collect expenditure data as part of a large-scale probability household panel: coverage and participation rates and biases. *Surv. Res. Methods* 13, 23–44. doi: 10.18148/srm/2019.v1i1.7297

Jacobs, A., and Wallach, H. (2021). "Measurement and fairness," in *Proceedings of the 2021 ACM Conference of Fairness, Accountability, and Transparency* (Virtual Event Canada), 375–385.

Jahn, B., Friedrich, S., Behnke, J., Engel, J., Garczarek, U., Münnich, R., et al. (2022). On the role of data, statistics, and decisions in a pandemic. *Adv. Stat. Anal.* 106, 349–382. doi: 10.1007/s10182-022-00439-7

Jarvis, B. F., Keuschnigg, M., and Hedström, P. (2022). "Analytical sociology amidst a computational social science revolution," in *Handbook of Computational Social Science. Volume I: Theory, Case Studies, and Ethics*, eds U. Engel, A. Quan-Haase, S. X. Liu, and L. Lyberg (Routledge: London), 33–52.

Jurafsky, D., and Martin, J. H. (2023). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 3rd Edn*. Available online at: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf (accessed April 26, 2023).

Keusch, F., Bach, R., and Cernat, A. (2022). Reactivity in measuring sensitive online behavior. *Int. Res.* 83, 210–235. doi: 10.1108/INTR-01-2021-0053

Keusch, F., Struminskaya, B., Antoun, C., Couper, M. P., and Kreuter, F. (2019).Willingness to participate in passive mobile data collection. *Public Opin. Q.* 83, 210–235. doi: 10.1093/poq/nfz007

Keuschnigg, M., Lovsjö, N., and Hedström, P. (2018). Analytical sociology and CSS. *J. Comp. Soc. Sci.* 1, 3–14. doi: 10.1007/s42001-017-0006-5

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data Soc.* 1, 1–12. doi: 10.1177/2053951714528481

Klumpe, B., Schröder, J., and Zwick, M. (eds.) (2020). *Qualität bei zusammengeführten Daten. Befragungsdaten, administrative Daten, neue digitale Daten: Miteinander besser?* Wiesbaden: Springer VS.

Kreuter, F., Haas, G.-C., Keusch, F., Bähr, S., and Trappmann, M. (2020). Collecting survey and smartphone sensor data with an app: Opportunities and challenges around privacy and informed consent. *Soc. Sci. Comput. Rev.* 38, 533–549. doi: 10.1177/0894439318816389

Laney, D. (2001). *3-D Data Management: Controlling Data Volume, Velocity, and Variety*. META Group Research Note. (Stamford). Available online at: https://www.gartner.com/en/blog

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of the Google flu: traps in big data analysis. *Science* 343, 1203–1205. doi: 10.1126/science.1248506

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A. L., Brewer, D., et al. (2009). Computational social science. *Science* 323, 721–723. doi: 10.1126/science.1167742

Lazer, D., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., et al. (2020). Computational social science: obstacles and opportunities. *Science* 369, 1060–1062. doi: 10.1126/science.aaz8170

Leitgöb, H., and Wolbring, T. (2021). "Die Methoden der sozialwissenschaftlichen Datenerhebung im digitalen Zeitalter. Entwicklungen, Möglichkeiten und

Herausforderungen," in *Sozialwissenschaftliche Datenerhebung im digitalen Zeitalter*, eds T. Wolbring, H. Leitgöb, and F. Faulbaum (Wiesbaden: Springer VS), 7–43.

Mayer-Schönberger, V., and Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. New York, NY: Houghton Mifflin Harcourt.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Comp. Surv.* 54, 1–35. doi: 10.1145/3457607

Mitchell, M. (2019). *Artificial Intelligence. A Guide for Thinking Humans*. London: Pelican Books.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. (2021). Algorithmic fairness: choices, assumptions, and definitions. *Ann. Rev. Stat. Appl.* 8, 141–163. doi: 10.1146/annurev-statistics-042720-125902

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of Machine Learning, 2nd Edn*. Cambridge, MA: MIT Press.

Molina, M., and Garip, F. (2019). Machine learning for sociology. *Annu. Rev. Sociol.* 45, 27–45. doi: 10.1146/annurev-soc-073117-041106

Mullainathan, S., and Spiess, J. (2017). Machine learning: an applied econometric approach. *J. Econ. Perspect.* 31, 87–106. doi: 10.1257/jep.31.2.87

Murphy, K. P. (2022). *Probabilistic Machine Learning: An Introduction*. Cambridge, MA: MIT Press.

Olson, D. R., Konty, K. J., Paladini, M., Viboud, C., and Simonsen, L. (2013). Reassessing Google flu trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLoS Comput. Biol.* 9, e1003256. doi: 10.1371/journal.pcbi.1003256

Pavlović, T., Azevedo, F., De, K., Riano-Moreno, J. C., Magli,ć, M., Gkinopoulos, T., et al. (2022). Predicting attitudinal and behavioral responses to COVID-19 pandemic using machine learning. *PNAS Nexus* 1, pgac093. doi: 10.1093/pnasnexus/pgac093

Pearl, J. (2010). The foundations of causal inference. *Sociol. Methodol.* 40, 75–149. doi: 10.1111/j.1467-9531.2010.01228.x

Pessach, D., and Shmueli, E. (2022). A review on fairness in machine learning. *ACM Comp. Surv.* 55, 1–44. doi: 10.1145/3494672

Piano, S. L. (2020). Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Human. Soc. Sci. Commun.* 7, 9. doi: 10.1057/s41599-020-0501-9

Radford, J., and Joseph, K. (2020). Theory in, theory out: the uses of social theory in machine learning for social science. *Front. Big Data* 3, 18. doi: 10.3389/fdata.2020.00018

Rubin, D. A. (2008). For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* 2, 808–840. doi: 10.1214/08-AOAS187

Salganik, M. J. (2018). *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton University Press.

Schnell, R. (2019). ""Big Data" aus sozialwissenschaftlicher Sicht: Warum es kaum sozialwissenschaftliche Studien ohne Befragungen gibt," in *Erklärende Soziologie und soziale Praxis*, eds D. Baron, O. Arránz Becker, and Lois, D. (Wiesbaden: Springer VS), 101–125.

Sen, I., Flöck, F., Weller, K., Wei,ß, B., and Wagner, C. (2021). A total error framework for digital traces of human behavior on online platforms. *Public Opin. Q.* 85, 399–422. doi: 10.1093/poq/nfab018

Silver, N. (2012). *The Signal and the Noise. Why So Many Predictions Fail – but Some Don't.* New York, NY: Penguin Press.

Starke, C., Baleis, J., Keller, B., and Marcinkowski, F. (2022). Fairness perceptions of algorithmic decision-making: a systematic review of the empirical literature. *Big Data Soc.* 9. doi: 10.1177/20539517221115189 [Epub ahead of print].

Steyerberg, E. W. (2010). *Clinical Prediction Models. A Practical Approach to Development, Validation, and Updating*. New York, NY: Springer.

Stier, S., Breuer, J., Siegers, P., and Thorson, K. (eds.) (2020). Integrating Survey data and digital trace data: Key issues in developing an emerging field. *Soc. Sci. Comp. Rev.* 38. doi: 10.1177/0894439319843669

Sutton, R. S., and Barto, A. G. (2018). *Reinforcement Learning. An Introduction, 2nd Edn*. Cambridge, MA: MIT Press.

Törnberg, P., and Törnberg, A. (2018). The limits of computation: a philosophical critique of contemporary big data research. *Big Data Soc.* 5. doi: 10.1177/2053951718811843 [Epub ahead of print].

Törnberg, P., and Uitermark, J. (2021). For a heterodox computational social science. *Big Data Soc.* 8. doi: 10.1177/20539517211047725 [Epub ahead of print].

van der Ploeg, T., Austin, P. C., and Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med. Res. Methodol.* 14, 137. doi: 10.1186/1471-2288-14-137

van Dijck, J. (2014). Datafication, dataism and dataveillance: big data between scientific paradigm and ideology. *Surveill. Soc.* 12, 197–208. doi: 10.24908/ss.v12i2.4776

Watts, D. J. (2014). Common sense and sociological explanations. *Am. J. Sociol.* 120, 313–351. doi: 10.1086/678271

Winship, C., and Morgan, S. L. (2015). *Counterfactuals and Causal Inference. Methods and Principles for Social Research, 2nd Edn*. New York, NY: Cambridge University Press.

Wolbring, T. (2020). "The digital revolution in the social sciences: five theses about big data and other recent methodological innovations from an analytical sociologist," in *Sociology of the Digital – Digital Sociology, Sonderband 23 der Zeitschrift Sozialen Welt*, eds S. Maasen, and J.-H. Passoth (Baden-Baden: Nomos), 60–72.

frontiers | Frontiers in Big Data

METHODS
published: 04 May 2022
doi: 10.3389/fdata.2022.886362

Examining Sentiment in Complex Texts. A Comparison of Different Computational Approaches

Stefan Munnes[1*], *Corinna Harsch*[1†], *Marcel Knobloch*[1†], *Johannes S. Vogel*[1,2†], *Lena Hipp*[1,2†] *and Erik Schilling*[3]

[1] WZB Berlin Social Science Center, Berlin, Germany, [2] Faculty of Economics and Social Sciences Chair of Inequality Research and Social Stratification Analysis, University of Potsdam, Potsdam, Germany, [3] Institute for German Philology, Ludwig Maximilian University of Munich, Munich, Germany

Can we rely on computational methods to accurately analyze complex texts? To answer this question, we compared different dictionary and scaling methods used in predicting the sentiment of German literature reviews to the "gold standard" of human-coded sentiments. Literature reviews constitute a challenging text corpus for computational analysis as they not only contain different text levels—for example, a summary of the work and the reviewer's appraisal—but are also characterized by subtle and ambiguous language elements. To take the nuanced sentiments of literature reviews into account, we worked with a metric rather than a dichotomous scale for sentiment analysis. The results of our analyses show that the predicted sentiments of prefabricated dictionaries, which are computationally efficient and require minimal adaption, have a low to medium correlation with the human-coded sentiments (r between 0.32 and 0.39). The accuracy of self-created dictionaries using word embeddings (both pre-trained and self-trained) was considerably lower (r between 0.10 and 0.28). Given the high coding intensity and contingency on seed selection as well as the degree of data pre-processing of word embeddings that we found with our data, we would not recommend them for complex texts without further adaptation. While fully automated approaches appear not to work in accurately predicting text sentiments with complex texts such as ours, we found relatively high correlations with a semiautomated approach (r of around 0.6)—which, however, requires intensive human coding efforts for the training dataset. In addition to illustrating the benefits and limits of computational approaches in analyzing complex text corpora and the potential of metric rather than binary scales of text sentiment, we also provide a practical guide for researchers to select an appropriate method and degree of pre-processing when working with complex texts.

Keywords: sentiment analysis, German literature, dictionary, word embeddings, automated text analysis, computer-assisted text analysis, scaling method

OPEN ACCESS

Edited by:
Tobias Wolbring, University of Erlangen Nuremberg, Germany

Reviewed by:
Dimitri Prandner, Johannes Kepler University of Linz, Austria
Ulf Liebe, University of Warwick, United Kingdom

***Correspondence:**
Stefan Munnes
munnes@wzb.eu

†These authors have contributed equally to this work and share senior authorship

Specialty section:
This article was submitted to Data Science, a section of the journal Frontiers in Big Data

Received: *28 February 2022*
Accepted: *14 April 2022*
Published: *04 May 2022*

Citation:
Munnes S, Harsch C, Knobloch M, Vogel JS, Hipp L and Schilling E (2022) Examining Sentiment in Complex Texts. A Comparison of Different Computational Approaches. Front. Big Data 5:886362. doi: 10.3389/fdata.2022.886362

1. INTRODUCTION

Quantitative text analysis has enabled researchers to process vast amounts of text in research designs of unprecedented size. Computational methods ranging from prefabricated, "off-the-shelf" dictionary approaches to fully automated machine learning approaches (Grimmer and Stewart, 2013) have been used to reliably analyze text corpora that are too large to read in a

lifetime, including social media data (e.g., Twitter, Reddit), parliamentary debates, and online product reviews.

These new possibilities raise questions, however, about the validity and accuracy of computational methods used with different types of texts. While a given method may produce outstanding results for one text corpus, it may perform poorly on another. In this study, we therefore sought to answer the following question: Can computational methods also be used to predict the sentiment in linguistically complex texts—and if so, which methods should researchers choose to maximize accuracy and minimize costs? To assess whether and how accurately automated approaches can predict the sentiment of complex texts, we applied different methods to a corpus of reviews of contemporary German books, including both novels and non-fiction publications.

Book reviews constitute a challenging text type for computer-assisted text analysis. First, they tend to include different latent dimensions. In addition to a summary of the book's content, they contain the reviewer's judgment of the book. Sometimes they refer to other books or to current or past events. Second, the language used in reviewing books—novels in particular—itself tends to exhibit literary characteristics. Ambiguity, irony, and metaphors are difficult to capture, however, with automated approaches. Third, and closely related to the first two points, in contrast to texts that clearly express positive or negative assessments (e.g., product reviews), book reviews tend to lean in a positive direction. Low-quality books are either not reviewed at all or are criticized in cautious and ambiguous terms.

Our text corpus consists of a combination of a random sample and a purposive sample of book review summaries (N = 6,041) published on the German online literary magazine Perlentaucher. Based on this corpus, we compared the correlations between the sentiment that human coders identified in a given review ("the gold standard") with the sentiment that different approaches predicted. Given the complexity and nuances of book reviews, we worked with a metric rather than a binary scale for sentiment analysis when applying different dictionary and scaling methods. In addition to prefabricated dictionaries (Remus et al., 2010; Rauh, 2018; Tymann et al., 2019), we also assessed the accuracy of self-created dictionaries based on word embeddings (GloVe: Pennington et al., 2014), and both supervised (wordscores: Laver et al., 2003) and unsupervised (wordfish: Slapin and Proksch, 2008) scaling methods. Given the importance of data pre-processing in computer-assisted text analysis, we also systematically varied the degree of text and dictionary manipulation when trying out the different methods to assess the influence on accuracy. With our analyses, we sought to provide guidance to other researchers in their decision-making processes for or against different methods.

The results of our comparison of the different approaches and different degrees of corpus pre-processing and dictionary modifications can be summarized as follows: First, prefabricated dictionaries, which are computationally efficient and require minimal, if any, adaption, such as the inclusion of negations, had a low to medium correlation with the human-coded sentiments (r between 0.32 and 0.39). Second, self-created dictionaries using word embeddings (both pre-trained and self-trained), which impose higher coding intensity on researchers, performed poorly with our corpus (r between 0.10 and 0.28). We would therefore not recommend them without further adaptations for complex text corpora similar to ours. Third, the fully automated approach we used in our analyses (wordfish) performed worst on our corpus, with correlations near 0. The semi-automated approach (wordscores), by contrast, which requires intensive human-coding of the training data, worked quite well. The correlations with the human-coded data ranged between 0.58 and 0.61 depending on the degree of pre-processing.

With these insights, our study makes the following contributions: First, we explore the potentials and limits of computational approaches for analyzing complex text corpora with regard to their validity and efficiency and provide researchers with a practical guide for selecting an appropriate method and the appropriate degree of pre-processing. Second, in contrast to most sentiment analyses, we work with a metric rather than a binary sentiment measure to take nuanced judgments into account, which may be beneficial for the analyses of many other complex text corpora as well. Third, we provide researchers, especially those working with non-English text corpora, with practical hints for creating context-specific dictionaries. Last but not least, by analyzing texts from outside the political arena, our analyses of a corpus of book reviews from contemporary German literature may inspire research projects outside established fields.

2. BACKGROUND

2.1. Content Analysis in Times of Mass Communication

The analysis of text has always been of interest to social scientists. Words—both spoken and written—are an integral part of social realities and exert an enormous influence on individual behaviors and attitudes (e.g., Martin, 1991; Glasze, 2008; Klüver, 2009; Fisher et al., 2013; Walton and Boon, 2014; Ng and Leung, 2015). The major technique used to systematically extract data from different forms texts and classification of documents is content analysis. It is "a scientific tool" (Krippendorff, 2018, p. 18) to examine patterns in communication in a replicable and valid manner. Qualitative approaches to content analysis primarily rely on an interpretive understanding of meaning and semantic contexts; quantitative approaches, by contrast, use word frequencies, distributions, and statistics to classify texts. One of the key advantages of using content analysis to analyze social phenomena is its noninvasive nature, which sets it apart from approaches that simulate social experiences or collect survey answers. A major challenge for quantitative text analysis, on the other hand, is the variability of word meanings in different contexts.

The first content analyses were conducted at the beginning of the last century, when mass media had become a major communication tool, as a form of newspaper analysis. It became more relevant over the course of multiple economic crises and the two world wars as propaganda analysis (for the historical overview, see Krippendorff, 2018). After Berelson's (1952) characterization of quantitative content analysis as "a

research technique for the systematic, objective, and quantitative description of the manifest content of communication" (p. 18), content analysis was applied to more and more research fields (for an overview, see Grimmer and Stewart, 2013; Benoit, 2020). In political science, quantitative content analysis has been used to study topics ranging from public discourse to individual policy positions and ideological networks. For instance, Glasze (2008) examined the discursive construction of Francophonie as a global community, international organization, and geocultural space. Stephens-Davidowitz (2014) analyzed how Google search terms can indicate racist animus and examined their impact on presidential elections in the United States. Similarly, Tumasjan et al. (2010) explored whether political sentiments on Twitter can predict election results (cf. critically Jungherr et al., 2012). Laver et al. (2003) and Diaz et al. (2016) assessed policy positions. Klüver (2009) and Sagarzazu and Klüver (2017) analyzed party manifestos, legislative speeches, interest groups in the EU, and political communication strategies of coalition parties. Fisher et al. (2013) analyzed discussions on climate change in the US Congress and mapped the resulting ideological relationships to measure coalitions and consensus among political actors.

In sociology, too, the benefits of using quantitative content analysis to study social phenomena has been recognized in recent years, and the method has been widely applied. Schwemmer and Wieczorek (2020), for instance, studied the methodological divide and paradigmatic preferences in sociology by analyzing publications in generalist sociology journals. Bohr and Dunlap (2018) applied topic modeling in their analyzes of sociological publications to identify the key topics in environmental sociology and changes in them over time. In their analysis of newspaper articles and Wikipedia entries, Nelson and King (2020) examined how distinct strategies emerge in different environmental organizations by linking their actions to their goals. In her analysis of US newspaper coverage on Muslim and non-Muslim women, Terman (2017) found more and different types of reporting on Muslim women than on non-Muslim women who had experienced human rights violations. Bail (2012) studied how civil society organizations shaped the news media discourse in the years after 9/11 through pro- and anti-Muslim messaging in their press releases.

Quantitative content analysis has also been used to investigate questions of social inequality in general and gender inequality in particular. In an analysis of Wikipedia profiles, Wagner et al. (2016) showed that women's profiles were more likely than men's to contain information on topics related to family, gender, and relationships and that the descriptions of men and women differed in the abstractness of positive and negative qualities. By analyzing men's and women's advertisements of their services in an online marketplace for contract labor, Ng and Leung (2015) showed that women were more likely to emphasize the relational aspects of their work, whereas men focused on the transactional aspects. Similarly, Hannák et al. (2017) analyzed worker evaluations from the online freelance marketplaces TaskRabbit and Fiverr and found considerable gender and racial biases in these evaluations. Brown (2021) analyzed descriptions of artworks to examine whether artworks produced by men and women differed in their observable characteristics and whether similarly described artwork by men and women varied in listing prices.

2.2. Sentiment Analysis in Digital Ages

According to Liu (2010), textual information can be "broadly categorized into two main types: facts and opinions" (p. 627). With sentiment analysis, which can be thought of a special form of content analysis and which has become one of the most important ways to quantitively analyze large amounts of textual data during the last 20 years, researchers seek to capture the nonfactual part of texts. Sentiment analysis, which is sometimes also referred to as "opinion mining" (Liu, 2012), captures the subjectivity, emotionality, or attitude of the author as expressed in the text; these are the aspects that are "not open to objective observation or verification" (Pang and Lee, 2008, p. 9). Sentiment analyses typically rely on dichotomous sentiment classifications (positive vs. negative) and sometimes also include a neutral category; there are, however, also studies that measured more nuanced emotional aspects, such as joy, anger, or sadness (Alm et al., 2005; Wiebe et al., 2005; Nielsen, 2011).

At the outset, sentiment analysis was mainly a subfield in computational linguistics and computer science. It's rise is mainly associated with the development of Web 2.0 in the early 2000s, which led to an incredible growth in the number of public available messages containing emotionally loaded opinions in form of product reviews, blog posts, forums contributions, or social media content. In addition, the big-tech-fueled commercialization of the internet has fostered a strong interest in the valorization of personal postings, as business models are built on the analysis of user behavior. Therefore, sentiment analysis has become widespread, especially in the financial and management sciences, but also in service, healthcare and the political and social sciences because of its importance to society as a whole; [(Liu, 2010; Puschmann and Powell, 2018); for an historic overview, also see Mäntylä et al. (2018)].

In contrast to classical quantitative content analysis methods, such as topic modeling or genre classification, in this method, the sentiments analyzed can be expressed in more subtle ways, including via the use of metaphors and irony. This makes sentiments much more difficult to detect (Pang et al., 2002). As a restricted natural language processing (NLP) problem, sentiment analysis does not need to understand the semantics of every sentence or the entire document but only some aspects of it. There are, however, two difficulties here: first, the task of determine the object to which the opinion is related and, second, the highly context-dependent nature of human language, which is especially true for evaluations (Liu, 2010). Ambiguity is also a problem in human coding, where coders do not always clearly come to the same conclusion about the subjective expression of opinion (van Atteveldt and Peng, 2018).

2.3. Various Computerized Methods

A key aspect of computerized sentiment analysis is that it is a tool to approximate human judgement. Obvious advantages of computerized methods include the reduced time and costs; researchers can thus deal with much larger corpora of texts (King, 2011). However, researchers have struggled with problems

TABLE 1 | Overview of various sentiment classification methods.

Type	Method	Validity and reliability	Time and costs
Gold standard	Human-coded	++	++
Dictionary	Prefabricated	−	−−
	Corpus-specific (e.g., word embeddings)	+	+
Maschine learning	Supervised (e.g., wordscores)	+	++
	Unsupervised (e.g., wordfish)	−	−−

concerning the validity and accuracy of computerized methods compared to human judgment. For this reason, computerized coding is compared with the gold standard of manual coding of sentiment by human coders on different text with different languages, as we do in this article (Nelson et al., 2018; Puschmann and Powell, 2018; van Atteveldt et al., 2021).

Broadly speaking, the available computerized methods can be classified as first, prefabricated dictionaries, second, constructed dictionaries for specific contexts, and third, machine learning (Rudkowsky et al., 2018). Each of these methods comes with different advantages and disadvantages and presumably varies in their performance in accurately classifying texts or predicting text sentiment. See **Table 1** for a general overview of the methods that will be discussed.

One of the most common, intuitive, and feasible methods of measuring text sentiment entails the use of dictionaries. Dictionary methods use the appearance rate of certain words (or combinations of words) to measure specific characteristics of the text (Grimmer and Stewart, 2013, 274). Dictionaries usually contain a list of words with a certain score (i.e., negative or positive) attached to them (DiMaggio, 2015, 274). The frequency with which words in either one of these categories appears in a text document is then used to measure the polarity of this document. Prefabricated dictionaries impose low costs on researchers and are ideal for replication purposes. There are a number of dictionaries, in different languages, that are easy to download, and some are already included in common software packages.

The advantages of dictionary approaches are that they are easy to use, computationally efficient, reliable, and require minimal working time if prefabricated dictionaries are used. Some potential shortcomings of dictionary methods are that they lack specificity, sensitivity, and validity (Benoit, 2020, 14f.). That is, instead of associating all relevant words—and only those—with positive or negative sentiments, dictionary methods may identify content that is not relevant for classifying a text (a lack of specificity), may not identify all relevant content (a lack of sensitivity), or may identify content inaccurately (a lack of validity), as words can have multiple meanings ("polysemes") and may be used differently in different contexts (e.g., in ironic discourse) (Grimmer and Stewart, 2013; Muddiman et al., 2019, 274). Dictionary accuracy may therefore vary depending on both the dictionary used and the characteristics of the text corpus. Recent advances in the development of multilingual (Proksch et al., 2019) and corpus-based dictionaries (Rice and Zorn, 2021) have sought to take these challenges into account.

Researchers can also modify prefabricated dictionaries according to their needs or engage in the tedious process of creating their own custom dictionaries (e.g., Muddiman et al., 2019) when the text under examination is very specific and uses unusual vocabulary and idioms (which may be the case with book reviews). Rice and Zorn (2021), for instance, have shown how to use certain machine learning methods to create a corpus-specific dictionary for specialized vocabularies in different contexts. The basic idea is to use what are known as word embeddings to find words that are similar to selected positive and negative words. Word embeddings are representations of words and their contextual meanings in a real-valued vector space. These specific methods of word embeddings are part of the broader field of natural language processing and refers to the distributional hypothesis proposed by Harris (1954). This hypothesis states that words appearing in the same context share the same meaning. Since this method creates word vectors using the global word-word co-occurrence statistics from a text corpus and neural networks, it is much more advanced and complex than dictionary approaches. However, it can be used for specific corpuses, and no human-coded training data is needed.

To overcome the challenges and shortcomings of dictionary approaches, researchers may also consider using either supervised or unsupervised machine learning methods, also known as classification and scaling methods (Grimmer and Stewart, 2013). Supervised machine learning methods require researchers to specify the relevant dimensions of interest in a set of pre-coded training texts, for example, the topic or the positive/negative text sentiment. Based on the dimensions specified in this training set, machine learning methods subsequently try to predict the characteristics of the unrated set of test texts (Benoit, 2020). Usually, such approaches entail attempting to classify the sentiment of a text into two or three categories. Classifiers like naive Bayes, maximum entropy or support-vector machines are used for this purpose. For our approach, which involves measuring sentiment in a more differentiated way on a metric scale, scaling methods are suitable.

A prominent supervised scaling method is wordscores (Laver et al., 2003). Wordscores assigns texts to a position on a continuous scale—the range of which is provided through the pre-coded training set. As is the case with dictionary methods, wordscores and other scaling methods have several advantages: replicability, reliability, speed, and low cost. Major disadvantages of supervised scaling methods are that the scaling of the texts in the training dataset requires considerable human coding for texts that are not yet classified. Moreover, the only words that are considered in the test dataset are those that were scaled in the training dataset, and only the relative importance of these words for determining the text sentiment is not contingent on the larger text content (similar to dictionary methods).

In terms of unsupervised scaling methods, wordfish (Slapin and Proksch, 2008) shares many of the advantages of wordscores but can be applied without reference texts and therefore requires less time and entails lower costs for researchers. However, the scale that unsupervised methods such as wordfish identifies may

be unclear and corpus specific. As a result, it is difficult to replicate and compare the accuracy of text sentiment predictions across different corpora.

Regarding the current status of the general quality of the different methods, as of today, the "best performance is still attained with trained human or crowd coding" (van Atteveldt et al., 2021, p. 1). van Atteveldt et al. (2021) further conclude that neither dictionaries nor machine learning approaches "come close to acceptable levels of validity" (p. 1). While deep learning approaches outperform dictionary-based methods, they nonetheless fall short in comparison to human classification.

3. DATA

3.1. Book Reviews as an Example of Complex Text

To investigate how accurately these different computational methods predict the sentiment in complex texts, we draw on a corpus of reviews of contemporary German books, including novels and non-fiction publications. Book reviews pose numerous challenges for automated analysis. First, book reviews commonly consist of various latent dimensions and linguistic elements. They usually comprise an overview of the plot that is formulated in relatively neutral terms, a contextualization of the work within the contemporary literary landscape, and an evaluation of the book by the reviewer. However, these dimensions are neither easily separated from each other, nor is the reviewer's assessment necessarily confined to the evaluation part. If, for example, reviewers see deficits in a book's structure, they will typically not summarize it in a neutral way. Reviewers may also judge a book differently depending on whether they approve of current literary trends. Second, book reviews are often characterized by linguistic ambiguities—ironic passages, metaphors, or sentences that praise a key idea but critique its realization. Third, book reviews often aim at surprising readers by creating certain expectations, only to subvert them and arrive at the opposite conclusion. In addition, reviewers may have various intentions, each with different implications: They may want to highlight a book's deficits or demonstrate their own broad knowledge. Hence, a neutral review that arrives at a matter-of-fact evaluation is more the exception than the rule.

In order to separate the different textual dimensions from each other and to reduce the text corpus to those passages in which reviewers provide their evaluation of the book, we decided not to work with the full-length reviews published in newspapers. Instead, we assembled our text corpus by collecting short versions of book reviews that focused on reviewer judgments from the German online literary magazine Perlentaucher, which has been in existence since 1999. Perlentaucher provides its readers with a daily overview of reviews published in the most important German newspapers and broadcast over the German public radio station Deutschlandfunk.

3.2. Data Collection and Sampling

The textual data of the summarized Perlentaucher reviews were collected along with additional information about the authors and books through web-scraping in May 2021. [1] In total, 88,248 unique reviews of 54,744 books by 33,168 authors were collected. The mean number of reviews for the total of 51,126 books with at least one summary review on Perlentaucher is 2.44 (SD of 1.6). The median number of tokens (i.e., the building blocks of the text, which in our case are words) per review is 113, with 20 for the shortest and 932 for the longest review. For our analyses, we sought to reduce reviews of translations and non-fiction books in our sample.[2]

From this corpus, we first drew a random sample of more than 6,000 book reviews and supplemented these with a purposive sample of 612 additional reviews. The purposive sample consisted of books that were either very well or very poorly received, controversial, or widely debated in German feuilletons. This step of selection was supported by the literary experts we interviewed prior to data collection. The sample of randomly and purposively selected reviews was then used to establish the "true" sentiment of the short reviews—the "gold standard," which we used to evaluate the accuracy of the different types of computational methods. In addition, we used a corpus containing all reviews with two different pre-processing strategies to train the word embeddings with the GloVe model.

3.3. Human-Coded Sentiment Analysis of Book Reviews

A total of seven paid, trained raters—most of them students with a background in literary studies—hand-coded the sentiment of the texts on a scale from 1 to 7 (very poor to very good)[3] for 1,000 randomly drawn reviews[4] per rater from the sample described above. After the completion of the coding process, we excluded reviews with missing scores and reviews that did not contain an evaluation. The final dataset of the human-coded reviews contained 6,041 valid sentiment scores. As expected, the reviews in our sample tended toward positive evaluations (median sentiment of 6, mean 5.09, and SD 1.66). Of these reviews, 656 were double-coded. We used these double-coded reviews to assess inter-coder reliability.[5] The intraclass correlation coefficient (ICC) was 0.86 (95% CI: 0.84; 0.87). **Figure 1** provides a scatter plot of inter-coder ratings. Based on the high consistency in the ratings (Liljequist et al., 2019), we assumed that all other reviews were also thoroughly and accurately coded. For reviews that were validly double-coded, we randomly chose one of two sentiment judgments for our analyses in order to have the same uncertainty measure in the evaluation.

[1] All R-scripts and important data for replication can be found at the GitHub repository.

[2] To exclude translations, we relied on the standard phrase in Perlentaucher book descriptions stating the language in which the book was originally published ("Aus dem LANGUAGE von …") as well as books that Perlentaucher labeled as non-fiction ("Sachbuch") in the tag or topic classification of the book. Moreover, we scraped additional Dewey Decimal Classification data from the German National Library in order to identify reviews of fiction books.

[3] Coders could also indicate if they were not able to find any evaluation of the book in the review at all; these reviews were coded as missing values and excluded.

[4] Book titles were randomly drawn so that all associated reviews would be rated by a single coder. There were an average of 2.44 reviews per book.

[5] Raters did not know how many and which reviews were also coded by another rater.

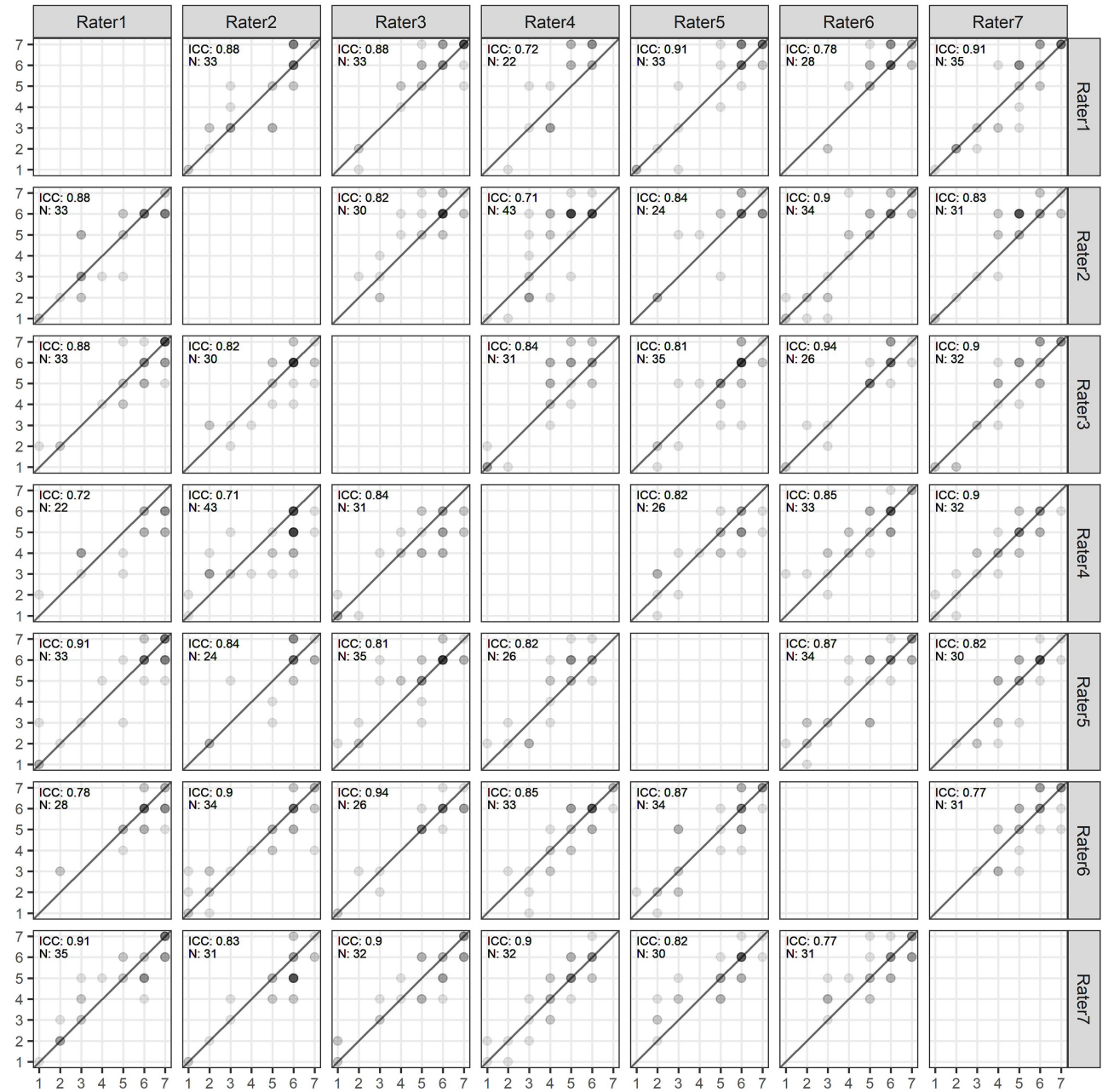


FIGURE 1 | Scatter plot and ICCs of ratings between pairs of coders.

3.4. Data Pre-processing

Data pre-processing is of vital importance for computational text analysis. Decisions about how to work with data should therefore always be made on the basis of pre-defined, methodological considerations (Denny and Spirling, 2018) as well as cost-benefit analyses associated with data cleaning and preparation. To enable researchers to make more informed decisions about the best degree of pre-processing for a given method, we examined how the accuracy of sentiment prediction of different methods varied between minimal and maximal levels of data pre-processing. The minimal pre-processing involved only the removal of punctuation, numbers, symbols, and separators from the reviews. The maximal pre-processing additionally involved the following alterations: We first stripped the reviews of the author names, the reviewer names, as well as the book titles and replaced all of them with empty tokens in order to maintain the original structure of the reviews. We then applied the same procedure to the tags and topics that had been assigned by Perlentaucher. These terms may affect how the different methods assess the sentiment of the reviews even if they are unrelated

TABLE 2 | Illustration of minimal vs. maximal pre-processing on an examplary review.

Original review	Tokens min. pre-processed	Tokens max. pre-processed
"Rezensentin Christiane Pöhlmann freut sich zu früh über Literatur aus Lettland. Inga Abeles Roman dämpft ihr Leseglück doch recht schnell mit der Geschichte einer jungen Lettin zwischen dem drängenden Wunsch nach Selbstverwirklichung als Drehbuchautorin und Depression, die Pöhlmann zufolge einfach zu viel zwischen die Buchdeckel klemmen will, Perspektivwechsel, Monologe, Briefe, alternative Milieus, abstrakte Passagen über Lektüre, Exil und Russland. Die persönliche Tragödie der Protagonistin kommt darüber zu kurz, bedauert Pöhlmann."	"Rezensentin" "Christiane" "Pöhlmann" "freut" "sich" "zu" "früh" "über" "Literatur" "aus" "Lettland" "Inga" "Abeles" "Roman" "dämpft" "ihr" "Leseglück" "doch" "recht" "schnell" "mit" "der" "Geschichte" "einer" "jungen" "Lettin" "zwischen" "dem" "drängenden" "Wunsch" "nach" "Selbstverwirklichung" "als" "Drehbuchautorin" "und" "Depression" "die" "Pöhlmann" "zufolge" "einfach" "zu" "viel" "zwischen" "die" "Buchdeckel" "klemmen" "will" "Perspektivwechsel" "Monologe" "Briefe" "alternative" "Milieus" "abstrakte" "Passagen" "über" "Lektüre" "Exil" "und" "Russland" "Die" "persönliche" "Tragödie" "der" "Protagonistin" "kommt" "darüber" "zu" "kurz" "bedauert" "Pöhlmann"	"" "" "" "freut" "" "" "frueh" "ueb" "literatur" "" "" "" "" "" "daempft" "" "leseglueck" "" "recht" "schnell" "" "" "" "" "" "jung" "lettin" "" "" "draengend" "wunsch" "" "selbstverwirklich" "" "drehbuchautorin" "" "depression" "" "" "zufolg" "einfach" "" "viel" "" "" "buchdeckel" "klemm" "" "perspektivwechsel" "monolog" "brief" "alternativ" "milieus" "abstrakt" "passag" "ueb" "lektu" "exil" "" "russland" "" "perso" "tragoedi" "" "protagonistin" "kommt" "darueb" "" "kurz" "bedauert" ""

to reviewers' evaluations of the book (for example, in the case of the book *Ein schlechter Verlierer* or the author Freya Stark, the word "schlechter" (bad) and the last name Stark (also the word for strong) may influence the review sentiment). Third, we stemmed and converted all words to lowercase, changed all special German characters such as umlauts to Latin characters, and stripped the corpus of common stopwords. For this, we used the standard German stopwords list from the quanteda R package (Benoit et al., 2018) with two modifications: We deleted negating and strengthening words[6] that may be important for sentiment detection and added review-specific words[7] to it.

For the minimally pre-processed corpus, the median number of tokens per review was 115 (range 45 and 932) in our sample, that is, human-coded reviews; the median number of unique tokens was 92 per review (range 37–488). The reviews in the corpus with maximal pre-processing were much shorter for both tokens (median 56, range 19–536) and unique tokens (median 53, range 19–365). The extensive pre-processing hence indeed shortened the corpus substantially (reduction in median number of all and unique tokens by half) and reduced the number of words that occurred frequently and were presumably unnecessary to determine the text sentiment (shown by the small difference in the medians of all vs. unique tokens). **Table 2** provides an illustration of how the original book review from Perlentaucher (column 1) changed with minimal data pre-processing (column 2) and maximal data pre-processing (columns 3).

4. METHODS

In our comparison of how accurately different computational methods can predict the nuanced sentiments and evaluations of book reviews, we drew on the following approaches: First, we applied three prefabricated, German dictionaries to our corpus, namely SentiWS (Remus et al., 2010), Rauh's German Political Sentiment Dictionary Rauh (2018), and GerVADER (Tymann et al., 2019). Second, we applied a self-created, corpus-based dictionary to our corpus that we constructed using the GloVe algorithm by Pennington et al. (2014). Third, we applied a supervised (wordscores by Laver et al., 2003) and an unsupervised method (wordfish by Slapin and Proksch, 2008).

In contrast to the majority of common sentiment analyses, which only differentiate between a positive and a negative and sometimes also a neutral category, we used a metric sentiment scale for our analyses. We did this for two reasons. First, we wanted to do justice to the specificity of our text corpus: Book reviews are generally not either entirely good or entirely bad, but instead contain subtle distinctions in a wide range of judgments. Second, we wanted to stress-test the various methods and assess how well computational methods map onto the fine-grained differences in the evaluations. To ensure comparability, we therefore worked with z-standardized scales.

4.1. Prefabricated Dictionary Methods

The first dictionary we used in our analyses was SentimentWortschatz (SentiWS), which was developed by the Department of Natural Language Processing at the University of Leipzig (Remus et al., 2010). SentiWS contains a list of 15,559 negative and 15,491 positive words—adjectives, verbs, and nouns, as well as their inflections. These features make SentiWS well-suited for our two pre-processing approaches, as we did not manipulate the capitalization and inflections of words (which in German can change their meaning) in the minimally pre-processing approach.[8]

In our analyses, we applied the SentiWS dictionary to both the minimum and maximum pre-processed corpus, once without and once with modifications to the dictionary. The modifications reduced the number of positive and negative words to 2,343 and 2,575, respectively. To include negations in the modified SentiWS dictionary and match them with negations in our corpus, we followed Rauh's recommendation Rauh (2018) and replaced six

[6]For example, "aber" (but), "kein" (no), "sehr" (very), "viel" (much).
[7]For example, "Autor" (author), "Rezensentin" (reviewer), "Buch" (book).
[8]The original SentiWS dictionary also contains weights for the strength of sentiment for each word, but we only used the binary sentiment version provided in the quanteda.sentiment R package (Benoit, 2021) for better comparability with the other approaches.

pre-determined German negating terms[9] with the English word "not" in our corpus. We connected the negating term with the following word as a bigram to form a single token that can be identified by the dictionary. To form the modified dictionary, we added a "not" negated version of each already existing token to the dictionary.

The second dictionary we used in our analyses was Rauh's German Political Sentiment Dictionary (Rauh, 2018), which is also available in the R package quanteda.sentiment (Benoit, 2021). The Rauh dictionary contains 74,160 entries, which are drawn from the SentiWS dictionary (Remus et al., 2010) and the GermanPolarityClues dictionary (Waltinger, 2010). In contrast to the two original underlying dictionaries, the Rauh dictionary also includes negated forms of each word. Accordingly, the entries are associated with four different keys: positive, negative, negated positive, and negated negative. To analyze the overall sentiment of a text, the negated positive words are meant to count as negative and the negated negative words as positive.

As with the other dictionary methods, we applied the Rauh dictionary to both the minimally and maximally pre-processed human-coded corpus. Similar to what we did in our analyses with the SentiWS dictionary, we replaced the negations in our text corpus with "not" and formed a bigram token. To compare the Rauh dictionary directly to the SentiWS dictionary, we also generated a minimally and maximally pre-processed version of the dictionary without the negated word forms. In the maximally pre-processed version, we performed the same steps as for the SentiWS dictionary: All words were stemmed, and German umlauts were transformed. This left us with a dictionary of 9,784 negative and 10,020 positive words in the dictionary containing negations. For the dictionary without negations, 6,161 negative and 4,028 positive entries were left.

The third dictionary we used in our analyses was GerVADER, a German adaption of the English language dictionary VADER (Hutto and Gilbert, 2014; Tymann et al., 2019). VADER consists of words taken from various other dictionaries such as the Linguistic Inquiry and Word Count dictionary (LIWC, Pennebaker et al., 2001) as well as special slang words and emoticons. The creators used crowd-coding to rate the polarity and intensity of each word. A strong feature of VADER are the heuristics implemented into the dictionary that allow a deeper understanding of text beyond bag-of-word analyses, in which the occurrence or frequency of words is used to classify texts, ignoring grammar or word order.[10] VADER, moreover, includes intensifying adverbs, such as "extremely," "very," or "marginally," and considers the mixed polarity of sentences starting with modifying conjunctions. VADER also examines trigrams preceding every word that carries sentiment and can therefore catch negations with a higher accuracy. VADER has been found to perform better in predicting text sentiment than other dictionary approaches and machine learning algorithms—and, in some instances, better than human coders (Hutto and Gilbert, 2014, 221).

The German VADER version, GerVADER, includes most of these features. The lexicon is based mainly on the SentiWS dictionary and was subsequently enlarged to include slang words. These words were then crowd-coded regarding polarity and intensity.[11] GerVADER, however, does not perform as well as the original VADER English language dictionary—most likely due to lexical and grammatical differences between German and English that are not captured by a simple translation (Tymann et al., 2019, 11). In German, moreover, negating words often appear after the verb at the end of the sentence. As VADER only considers negating words before the sentiment-laden word, negated words tend to be detected less frequently in German language corpora. Furthermore, GerVADER struggles to correctly classify longer sentences.

As with the other dictionaries, we processed the GerVADER dictionary according to our minimal and maximal criteria. Most notable in this case was the stemming, which greatly reduced the words contained in the dictionary. The original GerVADER dictionary used for the minimal approach contained 16,477 negative and 18,020 positive words. After preparing for the maximal approach, the dictionary contained 3,331 negative and 4,072 positive terms.

4.2. Word Embeddings: GloVe

In addition to these prefabricated dictionaries (and their modifications), we created a corpus-specific dictionary by drawing on a machine learning algorithm. We followed the example of Rice and Zorn (2021) and used the GloVe algorithm (Pennington et al., 2014) to generate word vectors from our corpus to build a corpus-specific dictionary.[12] We trained our own GloVe model, using the text2vev R Package (Selivanov et al., 2020), and created corpus-specific word embeddings. Here again, we varied the degree of pre-processing—this time for our total corpus of 88,248 reviews. For each pre-processed version, we also included a variant with additional bigrams in the word co-occurrence matrix to test whether negations and intensifications changed the results. For example, we wanted to see if word pairs like "not good" or "very good" would be part of the dictionary and would be attributed correctly.

There are various parameters in the modeling process that can be changed to identify the best model for a given dataset. For the purpose of our analyses, we followed the recommendations of Pennington et al. (2014) and Rodriguez and Spirling (2022).

[9]We added the word "ohne" (without) to Rauh's suggested list of "nicht" (not), "nichts" (nothing), "kein," "keine," and "keinen" (all inflections of the word no).

[10]For instance, VADER assigns higher scores to sentences ending with multiple exclamation marks or words that are written in all uppercase letters. This makes VADER especially useful for social media analyses, for which it was developed and for which it showed better results than other dictionaries. However, as our corpus was made up of reviews originally published in newspapers, the language is much more formal.

[11]It is important to note that, contrary to the original VADER, the raters did not receive financial compensation, which could have impacted their motivation and the data quality (Tymann et al., 2019, 6).

[12]We would like to point out that as of today, the word2vec algorithm (W2V), which was introduced by Google developers (Mikolov et al., 2013), is an additional, widely used and well documented algorithm that could be used for building a corpus-specific dictionary. W2V includes two different learning models: Continuous Bag of Words and Continuous Skip-Gram. While the first tries to predict every specific word based on a window of surrounding context words, the second tries exactly the opposite: It estimates the surrounding words from the specific word.

To have enough context for each token, we kept a minimum occurrence of five tokens. We also used a symmetric window size of 10, that is, five words before and five after the token. A larger window size (> 4) is recommended if the researcher is more interested in semantic than syntactic similarities. We also trained for the recommended 300 dimensions, the length of the resulting word vectors, with 10 iterations. This process resulted in four matrices of word vectors: The smallest is the maximum pre-processed variant with only onegrams (44,741 words and 105 MB of memory). The matrix with minimal pre-processing and onegrams contains 82,488 words and has a size of 194 MB. The matrix with maximal pre-processing and onegrams plus bigrams contains 95,674 words and has a size 226 MB. The matrix with minimal pre-processing contains 306,330 words and is 723 MB. On a computer with a CPU performance of 1.8 GHz and eight cores, the fitting of the models varied between 4 and 22 min.

As a next step, we used these four different matrices of word vectors to create our own dictionaries. This required positive and negative words as seeds to find similar words. To measure the similarity of the words represented as vectors, we used the cosine similarity. First, we used a list with 20 words, translated from (Rice and Zorn, 2021, henceforth RZ), which included generic and in principle interchangeable positive and negative terms, such as "brilliant" (brilliant), "wunderbar" (wonderful), and "schrecklich" (horrible). In a second step, we selected corpus-specific words from the hand-coded reviews that reflected the sentiment of the reviews, which we used as seeds (a total of 285 positive and 102 negative words, hence many more than in the first approach but including some very specific and rare words). These seeds were also pre-processed, so that they fitted the word vectors from the pre-processed corpus, which led to a seed corpus of 219 unique positive and 85 unique negative words for the maximally pre-processed corpus. In addition to typical words, these seeds also included words like "lustvoll" (lustful), "Poesie" (poetry), "Realismus" (realism), "Leichtigkeit" (easiness), or "kitschig" (cheesy), "billig" (cheap), "erwartbar" (expectable), and "Altherrenfantasie" (old men's fantasy).

We looped each list of seeds—both RZ's and the corpus-derived list—over the four word vector matrices. For each word in the dictionary, we collected the 400 words with the most similar vectors and kept words with a cosine similarity of at least 0.25. This relatively low similarity was a compromise between obtaining good similarity values and ensuring we had enough words to construct the final dictionary. In addition, only unique words that were not included in the other sentiment list were retained. Furthermore, only the same number of words per sentiment category was retained to avoid imbalance in the later matching process. Due to the exclusion of very rare words, the matrices of the word vectors no longer included all seeds. This resulted in a substantial variation of the dictionary length—from just 179 words per sentiment for maximum pre-processed and excluded bigrams with the RZ seeds to 1,017 for minimally pre-processed hand-coded seeds with bigrams included. See **Table 4** for an overview of the dictionaries along with the results.

Even if the first impression of this approach seemed to be promising, we also identified some conspicuous features of the resulting dictionaries that we consider worthwhile to briefly discuss. First, there were numerous words that, according to common understanding, do not express sentiments. The negative seed "Klischees" (clichés), for instance, yielded a list that included the non-evaluative word "Dimensionen" (dimensions) among others. Second, there were words with the exact opposite meaning from their seed. The word "Erstaunen" (astonishment), for example, was generated from the seed "Bedauern" (regret). Such mismatches were particularly likely to occur in the case of bigrams that involved negations. While bigrams such as "der_Stimulus" (the stimulus) or "gut_lesbar" (easy to read) yielded plausible lists of similar words, negations often fail to be assigned to the opposite negated sentiment.[13]

To further investigate the specific and relatively small corpus we used to train our GloVe models may mean that the results are not as good as a trained model on a larger corpus with much more contextual information for each word. We therefore also compared a pre-trained GloVe model with our model. The company deepset offers word vectors for free, trained with data from the German Wikipedia, which is a commonly used corpus for word embeddings due to its size. For pre-processing purposes, they only remove punctuation and lowercase, which is essentially the same as our minimally pre-processed corpus, and the minimal term frequency is also five. They also have a window size of 10,300 dimensions of vectors, and iterate 15 times. There are vectors for 1,309,281 words, much more than we achieve with our corpus. Because of the enormous number of words, we could let the minimum cosine similarity vary as a filter from 0.3 to 0.5 for both sources of seeds. Otherwise, we used the same procedure for selecting words. We obtained a dictionary size of 159 each for the RZ seeds and 322 for the human-coded ones for the most stringent selection of words with a cosine similarity of 0.5 to our seeds. With a cosine similarity of 0.3, the dictionaries contain 2,223 words each for the RZ seeds and 8,096 for the human-coded ones.

4.3. Scaling Methods: Wordscores and Wordfish

A third set of methods we used for our analyses were computational scaling methods, which have the advantage of being able to deal with very context-specific vocabulary. At the same time, they avoid much of the costly and labor-intensive preparation self-developed dictionaries require. Unlike methods using classification, the algorithms assign texts a position on a continuous scale (cf. Grimmer and Stewart, 2013, 292). Scaling methods are thus especially suitable for our approach, attempting to capture a more nuanced gradation of sentiment.

We used wordscores as an example of a supervised scaling method (Laver et al., 2003). We trained wordscores with the quanteda.textmodels R package (Benoit et al., 2021) with a

[13]Unfortunately, there is no simple way to pre-determine the quality of the choice of words in advance. We have deliberately chosen not to edit the dictionaries by hand, even though some ambiguities are clearly apparent. On the one hand, we assume that the meaning, which is partly not obvious to us, results methodically from the corpus. On the other hand, we would expect the wrong meanings to average out. Nevertheless, we assume that significant improvements could be made at this point in the procedure with some effort if the dictionaries were manually edited.

training dataset that included around 50% of the human-coded reviews in our corpus (N = 3,015) and captured the entire range of all seven sentiments. The minimally pre-processed training data contained a total of 12,517 unique words and the maximally pre-processed data a total of 8,610 unique words.

The unsupervised machine learning method we applied to our corpus was wordfish, also included in the quanteda.textmodels R package. The algorithm was developed by Slapin and Proksch (2008) and goes a step further than wordscores as it does not require any human input. As an unsupervised machine learning approach, this scaling method assigns texts to positions on a scale entirely determined by the computer. This happens based on similarity in word use. The model builds on an assumed Poisson distribution of words across the corpus, from which it derives its name. With known word or document parameters, it could be calculated as a Poisson regression. Since both are unknown, two regressions are calculated alternately until they converge. Compared to wordscores, it thus has significant advantages: It does not require any human-coding or a human selection of reference texts. This maximizes the potential for reducing costs and labor. The downside is that, due to the scaling dimension being corpus-specific, it does not allow for any comparisons between analyses. Since the range is not determined by the researcher beforehand, the model is only able to capture the main dimension differentiating the texts. Wordfish has been able to work well with political left-right scales (Slapin and Proksch, 2008). Whether the easily replicable, reliable, and exceptionally cost-efficient scaling method does equally well with the subtle sentiment of complex literature reviews is the object of our test.

5. FINDINGS

We now turn to the results of our analyses. In each of the sections below, we report the correlation between the human-coded sentiment of the reviews and the sentiment predicted by each method for the various levels of data pre-processing and degree of dictionary modification. In addition to reporting the substantive results in this section, we also develop recommendations for researchers interested in applying the different methods to complex text corpora.

5.1. Low to Medium Accuracy of Prefabricated Dictionary Methods

The accuracy of the different prefabricated dictionary approaches in predicting the sentiment of the book reviews is generally low, as can be seen from **Table 3**. First, the results of the SentiWS dictionary were not particularly good. Of the different pre-processing and dictionary variants, the lowest correlation was obtained with the maximally pre-processed approach that did not include negations (r = 0.29 with the human-coded sentiment). We were able to assign a sentiment for 6,033 out of the 6,041 human-coded reviews. On average, 8.55 words per review were matched with the dictionary content. To examine why SentiWS yielded a comparably low accuracy, we also counted the number of reviews whose predicted sentiment was completely off, that is, the deviation from the human-coded sentiment value was greater than two standard deviations. For the maximally pre-processed approach, this was the case for almost 552 reviews (10%). Under the condition of minimal processing, the correlation between the predicted and the human-coded sentiment value was slightly higher (r = 0.32) and results were further improved when negations were added (r = 0.38 with minimal pre-processing). After the inclusion of negations, however, only 6,012 reviews with an average of 6.34 matching words could be rated, and the number of ratings that were "completely off" also improved only slightly (427 reviews still had predicted sentiment values that were more than two standard deviations off; 7%). Based on these findings, we recommend adding additional negations to the SentiWS dictionary for the analysis of complex texts; other extensive pre-processing, however, may not be necessary.

Although the Rauh dictionary also performed rather poorly across all pre-processing variations in our corpus, it nonetheless yielded the second-best results of all the methods tested. With minimal pre-processing (both with and without negations), it achieved a correlation of 0.39 with the human-coded sentiment values. The original dictionary successfully determined the sentiment for 6,035 (6,038 without negations) reviews and matched a mean number of 8.23 (9.38) words per review on average. Moreover, the dictionary approaches with minimal pre-processing also performed better with regard to the number of predicted review sentiments that were more than two standard deviations away from the value that the human coders assigned (422 (7%) for the original dictionary with negations included and 429 (7%) for the dictionary with removed negations). We would therefore again recommend minimal pre-processing for the Rauh dictionary. Although including negations in the dictionary did not make sentiment determination considerably better, results did not deteriorate when the negated dictionary was combined with a minimal pre-processing approach. Since negations are already included in the Rauh dictionary, the extra step of excluding them was not worth the effort in our case.

Next, we turn to the results of the GerVADER dictionary. The results in **Table 3** show that although GerVADER successfully scales most texts (N = 6,029 for the minimally and N = 6,033 for the maximally pre-processed corpus), correlations were only slightly better than the original SentiWS. For the minimal corpus, the correlation with human-coded results was 0.34, the correlation of the maximum approach was even lower (r = 0.31). It is not surprising that the maximum pre-processing had no positive effect on the dictionary, as GerVADER is more context-dependent than the other dictionaries included in our analyses. Interestingly however, the GerVADER dictionary underperformed compared to the negated SentiWS dictionary—presumably due to the higher number of predicted review sentiments that can be considered "completely off" (633–660; 10–11%). Although VADER is a promising tool for sentiment analysis, its German version may lack proper language implementation. It also needs to be noted that both the original VADER as well as GerVADER were originally intended for sentence-level classifications (in contrast to longer texts such as a book review) and were originally based on a 3-point classification (positive, negative, and neutral) and not on the more nuanced scale that we imposed and assumed for our corpus.

TABLE 3 | Characteristics and results for prefabricated dictionaries.

Dictionary					Results			
Source	**Negation**	**Pre-processing**	**# Pos.**	**# Neg.**	**N**	**Cor.**	**Matches[a]**	**2 SD[b]**
SentiWS		Minimal	15,591	15,559	6,033	0.32	8.55 (0.07)	552
		Maximal	2,343	2,575	6,031	0.29	8.88 (0.15)	540
	Negation	Minimal	31,150	31,150	6,012	0.38	6.34 (0.05)	427
	Negation	Maximal	4,918	4,918	6,033	0.36	9.23 (0.16)	421
Rauh		Minimal	17,330	19,750	6,038	0.39	9.38 (0.08)	429
		Maximal	4,028	6,161	6,041	0.37	16.00 (0.27)	439
	Negation	Minimal	37,080	37,080	6,035	0.39	8.23 (0.07)	422
	Negation	Maximal	10,020	9,784	6,041	0.36	15.10 (0.26)	483
GerVADER		Minimal	18,020	16,477	6,029	0.34	-	633
		Maximal	4,072	3,331	6,033	0.32	-	660

[a]*Average number (and share of average number of tokens) of tokens matched by the dictionary.*
[b]*Number of reviews that deviate more than 2 standard deviations from the human-coded results.*

Both issues may be additional explanations for its comparably poor performance.

5.2. Low Accuracy of the Work-Intensive Self-Created Dictionary Using Word Embeddings

The results of the self-created GloVe dictionary are shown in **Table 4** and are neither good nor robust and vary greatly depending on seed selection and the degree of data pre-processing. Generally, the maximally pre-processed word vectors lead to better results than the minimally pre-processed vectors. The same applies to word vectors that do not contain bigrams. In terms of correlations, we observe slightly better and more consistent results with the human-coded seeds.

The best results were obtained with the maximally pre-processed word vectors that did not contain bigrams. For the human-coded seeds, we observed a correlation of 0.28, and a correlation of 0.26 for the RZ seeds. The worst results were from the minimally pre-processed corpus with bigrams included. While a correlation of 0.17 was still achieved with the human-coded seeds, the RZ seeds yielded a value of –0.01. We also observed only 3–4 matches with the human-coded seed dictionaries, in comparison to 15 at the top for the smaller RZ dictionaries. It seems that the smaller but more specialized dictionary of human-coded seeds matches fewer words in the texts, but that these lead to a more accurate sentiment score, especially when the dataset was maximally pre-processed. The major downside to the more specialized, human-coded seed dictionaries was that no sentiment could be assigned for around 200 to 500 reviews.

For the pre-trained word vectors, we found the same pattern. Here, again, the dictionary with the corpus-specific seeds performed significantly better. While the dictionary derived from the RZ seeds had a constant correlation of only 0.1, when cosine similarity was increased from 0.3 to 0.5., the correlation for the dictionary derived from the corpus-specific seeds increased from 0.15 to 0.26. On average, only 5 words were matched for the best score, and about 80 reviews could not be scored at all for both maximally pre-processed dictionaries. The number of reviews that were incorrectly rated (> 2SD) was not as high as with the self-trained word vectors.

All in all, our self-created dictionaries based on word embeddings underperformed compared to the easier-to-implement, prefabricated dictionaries that we used on our corpus. If word embeddings are used to create dictionaries, we recommend the following: Better results can be achieved with a maximally pre-processed corpus; the additional use of bigrams does not improve the dictionary's accuracy. Self-trained vectors perform better than pre-trained vectors. Corpus-specific seeds lead to more accurate results than generic seeds. Furthermore, at least for the hand-coded seeds, a higher similarity of the words improves the results. In short, the more specific the words in the dictionary, the better the results.

5.3. High Accuracy of Semi-supervised but Low Accuracy of Unsupervised Scaling Methods

The wordscores algorithm calculated sentiment positions for 98.4% of the minimally and 99.4% of the maximally pre-processed words. Since the training texts were coded relatively positively with only a few clearly negative reviews, wordscores also yielded many more positive than negative words. With a threshold of 4 on the original 7-point scale, 11,124 minimally and 7,760 maximally pre-processed words can be considered positive and 1,193 minimally and 797 maximally pre-processed words negative. However, as **Figure 2** illustrates, the actual attribution of sentiment is not binary but continuous: A word can also be only slightly more positive or negative than another. Words that occur frequently tend to be assigned a relatively neutral sentiment. This is not surprising, as a term that appears in both positive and negative reviews—for instance, pronouns or merely descriptive words—usually do not carry much clear sentiment. This is illustrated by the peak in **Figure 2**. Five frequent negative, neutral, and positive terms are highlighted

TABLE 4 | Characteristics and results for self- and pre-trained GloVe dictionaries.

Dictionary						Results			
Source	**Seeds**	**Sim.[a]**	**Ngram**	**Preproc.**	**# P./N.[b]**	**N**	**Cor.**	**Matches[c]**	**2 SD[d]**
self-trained	hc	0.25		Minimal	425	5,748	0.21	3.77 (0.03)	683
		0.25		Maximal	257	5,779	0.28	3.76 (0.06)	575
		0.25	Bigram	Minimal	1,017	5,585	0.17	2.97 (0.02)	704
		0.25	Bigram	Maximal	269	5,823	0.24	4.04 (0.07)	695
self-trained	RZ	0.25		Minimal	317	6,038	0.17	11.11 (0.09)	747
		0.25		Maximal	252	6,004	0.26	6.16 (0.10)	589
		0.25	Bigram	Minimal	452	6,041	-0.01	15.19 (0.13)	964
		0.25	Bigram	Maximal	179	5,919	0.15	4.78 (0.08)	720
pre-trained	hc	0.3		Case ins.	8,096	6,041	0.15	36.18 (0.30)	772
		0.4		Case ins.	1,916	6,041	0.23	16.39 (0.14)	681
		0.5		Case ins.	322	5,963	0.26	5.01 (0.04)	573
pre-trained	RZ	0.3		Case ins.	2,223	6,041	0.10	31.54 (0.26)	886
		0.4		Case ins.	811	6,041	0.10	20.14 (0.17)	828
		0.5		Case ins.	159	5,958	0.10	5.89 (0.05)	803

[a] *Minimum cosine similarity of word vectors to each seed.*
[b] *Number of positive and negative words each.*
[c] *Average number (and share of average number of tokens) of tokens matched by the dictionary.*
[d] *Number of reviews that deviate more than 2 standard deviations from the human-coded results.*

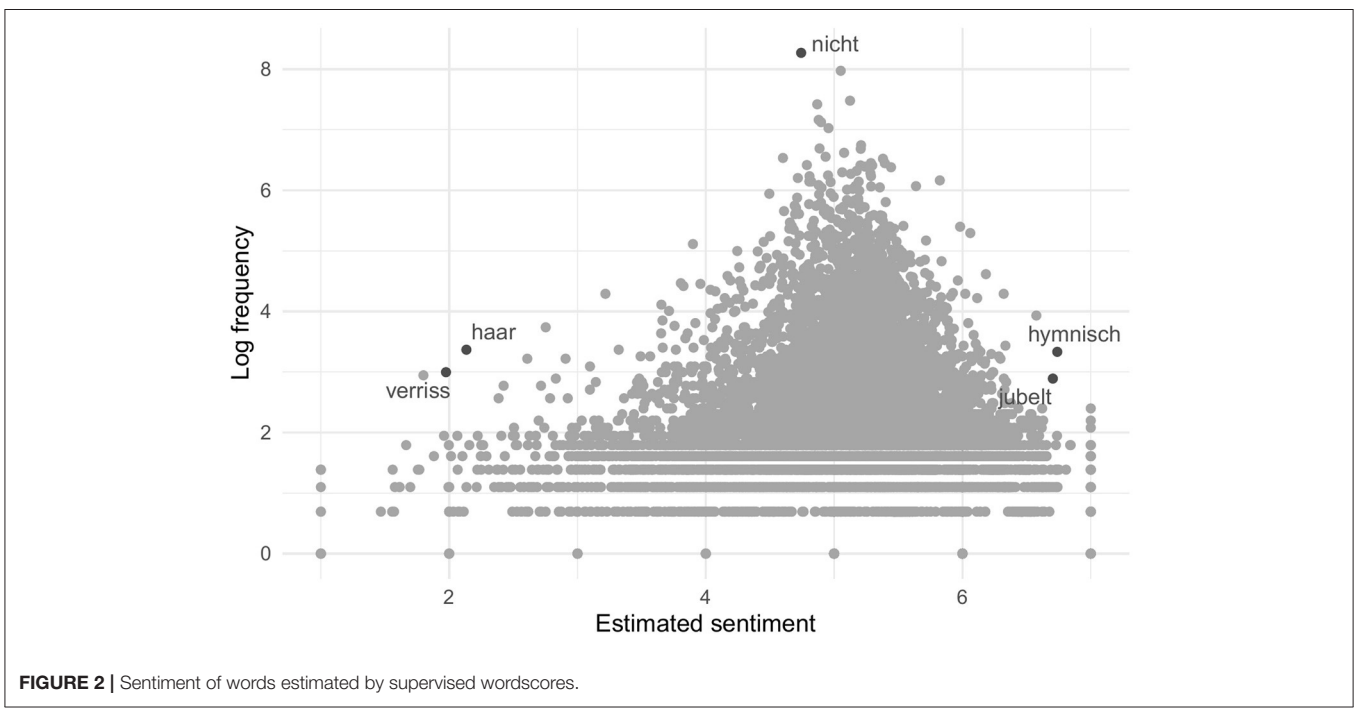


FIGURE 2 | Sentiment of words estimated by supervised wordscores.

as an example: "verriss" scorcher), "haar" (hair), "nicht" (not), "hymnisch" (anthemic), and "jubelt" (jubilates).

In the next step, the algorithm predicted the positions of the remaining 3,026 texts, based on the calculated ratings for the given words.[14] Since the wordscores "dictionary" is rather comprehensive, it matches, in clear contrast to the previous actual dictionaries, 99.9% (minimal corpus: 100%) of the 119.5 (minimal corpus: 58.9) words per review in the estimation set on average. This may explain the moderate to strong correlation of the estimated sentiment of the texts with our human-coded results of 0.58 for the minimally and 0.61 for the maximally pre-processed corpus. This is the best result we achieved and is 0.2 points higher than with the best dictionary approach. In

[14] The total runtime of the wordscores model was very moderate with 40–60 s per corpus.

TABLE 5 | Characteristics and results for supervised and unsupervised Methods.

Source	Pre-processing	# Pos.	# Neg.	N	Cor.	Matches[a]	2 SD[b]
Wordscores	Minimal	1,193	11,124	3,026	0.58	119.34	84
	Maximal	797	7760	3,026	0.61	58.91	76
Wordfish	Minimal	-	-	6,041	–0.05	119.50	1,095
	Maximal	-	-	6,041	–0.01	58.93	943

[a]*In contrast to dictionaries, almost all tokens (reported average) are used for scaling.*
[b]*Number of reviews that deviate more than 2 standard deviations from the human-coded results.*

addition, only 76 (3%)–84 (3%) (minimally to maximally) texts were rated more than two standard deviations off (see **Table 5**). This confirms our initial assumption that our corpus uses very specific language that is not adequately captured by prefabricated dictionaries. The method is also more accurate than the word embeddings approach, since it evaluates more words accurately. However, the cost for this good result is the amount of human coding required for the training texts (50% of the corpus).

Without relying on any human input, the wordfish algorithm calculated sentiment positions for all 12,517 minimally and 8,610 maximally pre-processed words in the corpus. Since the resulting scale is metric and exceeds the original seven points, however, a dichotomization into positive and negative appears difficult. While the median could serve as a threshold, this would obscure the expected unequal distribution of more positive than negative terms. We therefore refer to **Figure 3** to illustrate that the model has indeed converged and yields the expected Poisson distribution of words. The same five highlighted terms, however, already indicate that the estimation of sentiment was at most partially successful. While the words keep appearing in slightly different places, the opposing sentiment is no longer captured by the entirety of the scale.

Our doubts as to whether the wordfish estimation yields the sentiment of the reviews (rather than, for instance, the genre, the topic, or a mixture of these) grows when we compare the estimated sentiment positions of the texts with our gold standard, the human-coded results. While the unsupervised wordfish algorithm requires no human input for learning, estimates positions for all 6,041 texts, and matches 100% of the words in the estimation set, it yields a very weak correlation of –0.05 for the minimally and –0.01 for the maximally pre-processed corpus. In addition, 1,095 (18%) [minimal corpus: 943 (15%)] of texts were predicted more than two standard deviations off (see **Table 5**).[15]

Wordfish, despite its many advantages, is therefore not able to provide a useful sentiment estimation for our complex literature reviews. Since the algorithm only captures the least latent dimension, it appears that our text corpus is still too heterogeneous. For instance, some word positions point to a possibly involved dimension fiction–non-fiction, with a particular focus on music.[16] With further controls, such as in the enhanced Wordshoal algorithm of Lauderdale and Herzog (2016), which allows for control of intervening variables, the model might therefore yield better results. Yet as the necessary additional information (e.g., literary genre) is not available reliably for our source of literature reviews, for our corpus and with the information at hand, we recommend supervised scaling or dictionary methods instead.

[15]The total runtime of the wordfish model was 5–50 min per corpus.

[16]Features scaled as very negative, for example, were "bach" (river, but more likely the composer), "wohltemperiert" (well-tempered), "klavi[er]" (piano), "musikwissenschaft" (musicology), and "komponist" (composer).

6. CONCLUSION

In this paper, we applied different computational text analysis approaches to a corpus of short summaries of German book reviews to examine whether different computational methods accurately predict the sentiment in complex texts—and if so, under what conditions. Examining these questions is important for several reasons. First, social scientists are working increasingly with text-as-data to analyze topics of great political and societal interest, such as changes in political and social discourse and communication strategies or the representation of minorities in newspapers and Wikipedia entries. With increasing text complexity and potentially also increasingly complex questions, it is crucially important that researchers are aware of the potentials and limits of the different approaches and choose computational methods that work best on their corpus. Second, assessing text sentiments in complex texts and capturing gradual differences—for example, in the description of certain groups—tends to require more than a binary assessment of whether a text is positively or negatively loaded. Instead, researchers may be interested in assessing degrees of positivity and negativity. Third, although the introductory literature on approaches to quantitative text analysis is constantly growing, researchers lack sufficient guidance on what degree of data pre-processing and modifications to existing tools is beneficial when using different approaches.

With our analyses, we sought to contribute to each of these important points. In addition to comparing how well different computational methods—including three prefabricated German language dictionaries (SentiWS, Rauh, GerVADER), a self-created dictionary using pre- and self-trained word-embeddings (GloVe), and one supervised and one unsupervised scaling method (wordscores and wordfish)—predict the sentiment of complex texts, we used a metric instead of a binary scale to assess text sentiment, and systematically varied the degree of data pre-processing for each approach.

According to our analyses, predefined German-language dictionaries showed average performance on our corpus. Relying on predefined dictionaries is easy and inexpensive; however, the simple counting of predefined, labeled words does not account for the specific contexts in which words are used or correctly identify special linguistic features, such as metaphors, irony, and allusions. Additionally, dictionary approaches cannot solve another general problem of content analysis—the detection of a sentiment's object. With dictionary approaches, it is impossible for the researcher to differentiate between the content, the evaluation, and further contextual information that is included in unstructured texts. Based on our findings, we recommend that

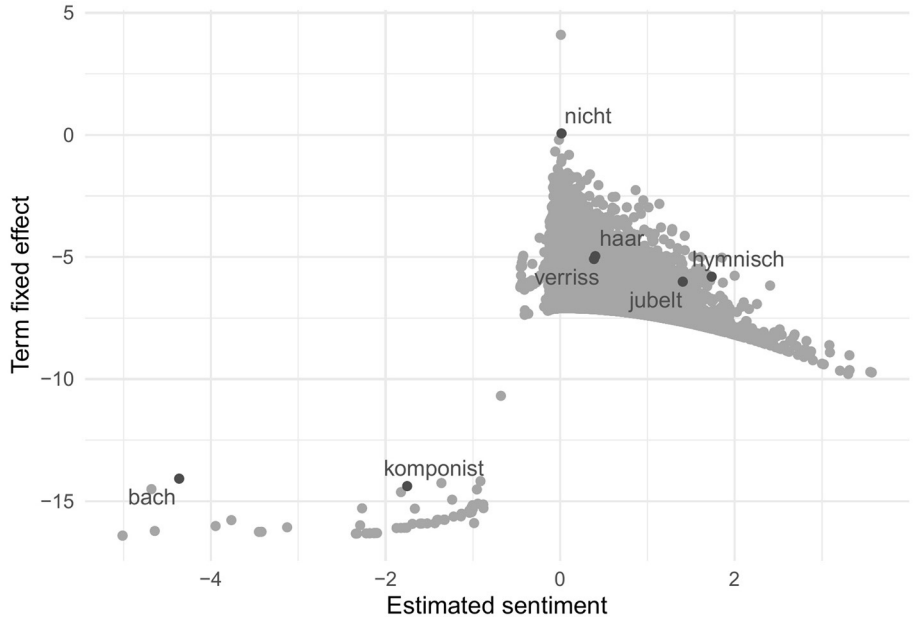


FIGURE 3 | Sentiment of words estimated by unsupervised wordfish.

researchers include negation terms when analyzing complex texts via these cost-efficient dictionary approaches.

Self-created dictionaries using word embeddings—both pretrained and self-trained—are a promising approach for analyzing texts for which predefined dictionaries were not designed. However, dictionary approaches using word embeddings impose high coding demands on researchers and actually performed poorly with our corpus. In theory, this approach intends to better capture the linguistic subtleties through the corpus-specific compilation of a list of words. When creating dictionaries based on word embeddings, researchers must deal with the trade-off between a small and highly specific dictionary and a large and unspecific dictionary by varying the cosine similarity to the chosen seeds. Although we sought to find a good compromise between a high similarity with the seeds and a sufficient number of words, with our corpus, the self-created dictionary was considerably less accurate in predicting the text sentiment than the prefabricated dictionaries. Furthermore, the results we obtained with word embeddings were not robust and varied considerably by seed selection and data pre-processing. Based on our experience, we suggest that researchers who apply the method manually should review the generated word lists and consider adding a small list of corpus-specific words to an existing dictionary.

There was considerable variation in the performance of the different machine learning approaches we applied. First, the accuracy in sentiment prediction based on wordfish, the unsupervised machine learning method, was even lower than the accuracy we obtained based on the prefabricated dictionaries. This low inaccuracy may be related to the many different latent dimensions that complex texts tend to have. In our text corpus, for instance, the content, genre, and evaluation of the book are all intermingled. The algorithm, however, only captures the least latent dimension. When using unsupervised scaling algorithms, researchers should try to reduce the number of text dimensions (which is a challenging task in unstructured texts, as was the case with ours). Second, the accuracy in sentiment prediction based on wordscores, the supervised machine learning method, was quite promising. The correlations between the predicted sentiment and the human-coded sentiments ranged between 0.58 (involving minimal data pre-processing) and 0.61 (with maximal data pre-processing). Given a sufficient number of classified texts, supervised learning methods fairly accurately identify patterns and predict the sentiment in even complex and specialized texts. The downside of the approach, however, is the high cost that the method entails in terms of the human coding necessary to train the model.

In conclusion, our results emphasize the importance of carefully choosing and evaluating different methods to ensure an optimal fit of the method to the data. Not only the methods used in the analyses but also the pre-processing influences the results, although not to a high and unambiguous degree. As a consequence, the research process should not be static, and the methods used should be constantly evaluated, adjusted, re-evaluated, and validated throughout the course of the project. In particular, by using word embeddings to create a corpus-specific dictionary, our results show both the potential and limits (as well as need for further advancements) of corpus-specific approaches. Overall, the analyses performed for this article provide researchers with some guidelines and ideas for how this can be done. In conclusion, we recommend scholars rely on supervised machine learning methods when

resources are available. When resources are unavailable, scholars can implement certain protocols to help other methods perform better.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/StefanMunnes/frontiers_literature/tree/master/data.

AUTHOR CONTRIBUTIONS

SM collected, organized, and cleaned the reviews database. MK and SM organized and supervised the sampling and hand coding process and wrote the section on data collection, manual coding, and pre-processing. CH, MK, and ES created a list of books for the purposive sample. CH and MK human-coded the seeds for the word embeddings. SM and JV pre-processed the data. Overall and final organization and cleaning of the code was done and coding and writing on embeddings and GloVe by SM. LH wrote the introduction and the conclusion and streamlined the article with support from SM. SM with help of the other authors contributed to the background section. Coding and writing related to the dictionaries was performed by SM, CH, and MK. Coding and writing on wordscores and wordfish by JV. Project management was the responsibility of LH, SM, and ES. LH and ES acquired the necessary funding. All authors proofread and approved the manuscript and participated in the conception and design of the study.

FUNDING

This research was partly funded by Junge Akademie. The publication of this article was funded by the Open Access Fund of the Leibniz Association.

ACKNOWLEDGMENTS

We thank our raters, Alexander, Antonia, Jan, Johanna, Norvin, Robin, and Sabine, for their thorough human-coding of the data and our literature experts for their help in identifying books that were controversial or received either bad or glowing reviews.

REFERENCES

Alm, C. O., Roth, D., and Sproat, R. (2005). "Emotions from text," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing-HLT'05*. (Vancouver: Association for Computational Linguistics).

Bail, C. A. (2012). The fringe effect: civil society organizations and the evolution of media discourse about islam since the september 11th attacks. *Am. Sociol. Rev.* 77, 855–879. doi: 10.1177/0003122412465743

Benoit, K. (2020). "Text as data: an overview," in *The SAGE Handbook of Research Methods in Political Science and International Relations* (SAGE Publications Ltd.), 461–497. Available online at: https://methods.sagepub.com/book/research-methods-in-political-science-and-international-relations/i4365.xml

Benoit, K. (2021). *quanteda.sentiment: Sentiment Analysis Using Quanteda*. R package version 0.22.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., et al. (2018). quanteda: An r package for the quantitative analysis of textual data. *J. Open Source Softw.* 3, 774. doi: 10.21105/joss.00774

Benoit, K., Watanabe, K., Wang, H., Perry, P. O., Lauderdale, B., Gruber, J., et al. (2021). *quanteda.textmodels: Scaling Models and Classifiers for Textual Data*. R package version 0.9.4.

Berelson, B. (1952). *Content Analysis in Communication Research. Foundations of Communications Research*. Free Press.

Bohr, J., and Dunlap, R. E. (2018). Key Topics in environmental sociology, 1990–2014: results from a computational text analysis. *Environ. Sociol.* 4, 181–195. doi: 10.1080/23251042.2017.1393863

Brown, T. (2021). *Qualities or Inequalities*?: *How Gender Shapes Value in the Market for Contemporary Art* (Dissertation). Duke University.

Denny, M. J., and Spirling, A. (2018). Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. *Polit. Anal.* 26, 168–189. doi: 10.1017/pan.2017.44

Diaz, F., Gamon, M., Hofman, J. M., Kiciman, E., and Rothschild, D. (2016). Online and social media data as an imperfect continuous panel survey. *PLoS ONE* 11, e0145406. doi: 10.1371/journal.pone.0145406

DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data Soc.* 2:2053951715602908. doi: 10.1177/2053951715602908

Fisher, D. R., Leifeld, P., and Iwaki, Y. (2013). Mapping the ideological networks of American climate politics. *Clim. Change* 116, 523–545. doi: 10.1007/s10584-012-0512-7

Glasze, G. (2008). Vorschläge zur operationalisierung der diskurstheorie von laclau und mouffe in einer triangulation von lexikometrischen und interpretativen methoden. *Histor. Soc. Res.* 33, 185–223. doi: 10.12759/hsr.33.2008.1.185-223

Grimmer, J., and Stewart, B. M. (2013). Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Polit. Anal.* 21, 267–297. doi: 10.1093/pan/mps028

Hannák, A., Wagner, C., Garcia, D., Mislove, A., Strohmaier, M., and Wilson, C. (2017). "Bias in online freelance marketplaces: evidence from taskrabbit and fiverr," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, OR: ACM), 1914–1933.

Harris, Z. S. (1954). Distributional structure. *Word* 10, 146–162. doi: 10.1080/00437956.1954.11659520

Hutto, C., and Gilbert, E. (2014). "Vader: a parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the International AAAI Conference on Web and Social Media, Vol.* 8, 216–225. Available online at: https://ojs.aaai.org/index.php/ICWSM/article/view/14550

Jungherr, A., Jürgens, P., and Schoen, H. (2012). Why the pirate party won the german election of 2009 or the trouble with predictions: a response to tumasjan, a., sprenger, t. o., sander, p. g., welpe, i. m. "predicting elections with twitter: What 140 characters reveal about political sentiment". *Soc. Sci. Comput. Rev.* 30, 229–234. doi: 10.1177/0894439311404119

King, G. (2011). Ensuring the data-rich future of the social sciences. *Science* 331, 719–721. doi: 10.1126/science.1197872

Klüver, H. (2009). Measuring interest group influence using quantitative text analysis. *Eur. Union Polit.* 10, 535–549. doi: 10.1177/1465116509346782

Krippendorff, K. (2018). *Content Analysis: An Introduction to Its Methodology*. SAGE PUBN.

Lauderdale, B. E., and Herzog, A. (2016). Measuring political positions from legislative speech. *Polit. Anal.* 24, 374–394. doi: 10.1093/pan/mpw017

Laver, M., Benoit, K., and Garry, J. (2003). Extracting policy positions from political texts using words as data. *Am. Polit. Sci. Rev.* 97, 311–331. doi: 10.1017/S0003055403000698

Liljequist, D., Elfving, B., and Skavberg Roaldsen, K. (2019). Intraclass correlation–a discussion and demonstration of basic features. *PLoS ONE* 14, e0219854. doi: 10.1371/journal.pone.0219854

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook Natural Lang. Process*. 2, 627–666.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lect. Hum. Lang. Technol.* 5, 1–167. doi: 10.2200/S00416ED1V01Y201204HLT016

Mäntylä, M. V., Graziotin, D., and Kuutila, M. (2018). The evolution of sentiment analysis-a review of research topics, venues, and top cited papers. *Comput. Sci. Rev*. 27, 16–32. doi: 10.1016/j.cosrev.2017.10.002

Martin, E. (1991). The egg and the sperm: How science has constructed a romance based on stereotypical male-female roles. *Signs* 16, 485–501. doi: 10.1086/494680

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv:1301.3781 [cs]*.

Muddiman, A., McGregor, S. C., and Stroud, N. J. (2019). (Re)Claiming our expertise: parsing large text corpora with manually validated and organic dictionaries. *Polit. Commun*. 36, 214–226. doi: 10.1080/10584609.2018.1517843

Nelson, L. K., Burk, D., Knudsen, M. L., and McCall, L. (2018). The future of coding: a comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociol. Methods Res*. 50, 202–237. doi: 10.1177/0049124118769114

Nelson, L. K., and King, B. G. (2020). The meaning of action: linking goal orientations, tactics, and strategies in the environmental movement. *Mobilization* 25, 315–338. doi: 10.17813/1086-671X-25-3-315

Ng, W., and Leung, M. D. (2015). For Love or money? gender differences in how one approaches getting a job. *SSRN Electron. J*. doi: 10.2139/ssrn.2583592

Nielsen, F. A. (2011). "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," in *Proceedings of the ESWC2011 Workshop on Making Sense of Microposts: Big Things Come in Small Packages* (CEUR-WS), 93–98.

Pang, B., and Lee, L. (2008). Opinion mining and sentiment analysis. *Found. Trends Inf. Retrieval* 2, 1–135. doi: 10.1561/9781601981516

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. *arXiv preprint* cs/0205070. doi: 10.3115/1118693.1118704

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). *Linguistic Inquiry and Word Count: Liwc 2001*. Mahway: Lawrence Erlbaum Associates.

Pennington, J., Socher, R., and Manning, C. (2014). "Glove: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha: Association for Computational Linguistics), 1532–1543.

Proksch, S.-O., Lowe, W., Wäckerle, J., and Soroka, S. (2019). Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches. *Legislative Stud. Q*. 44, 97–131. doi: 10.1111/lsq.12218

Puschmann, C., and Powell, A. (2018). Turning words into consumer preferences: how sentiment analysis is framed in research and the news media. *Soc. Media Soc*. 4:2056305118797724. doi: 10.1177/2056305118797724

Rauh, C. (2018). Validating a sentiment dictionary for german political language—a workbench note. *J. Inform. Technol. Polit*. 15, 319–343. doi: 10.1080/19331681.2018.1485608

Remus, R., Quasthoff, U., and Heyer, G. (2010). "Sentiws-a publicly available german-language resource for sentiment analysis," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (Leipzig: LREC'10), 1168–1171.

Rice, D. R., and Zorn, C. (2021). Corpus-based dictionaries for sentiment analysis of specialized vocabularies. *Polit. Sci. Res. Methods* 9, 20–35. doi: 10.1017/psrm.2019.10

Rodriguez, P. L., and Spirling, A. (2022). Word embeddings: what works, what doesn't, and how to tell the difference for applied research. *J. Polit*. 84, 101–115. doi: 10.1086/715162

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., and Sedlmair, M. (2018). More than bags of words: Sentiment analysis with word embeddings. *Commun. Methods Meas*. 12, 140–157. doi: 10.1080/19312458.2018.1455817

Sagarzazu, I., and Klüver, H. (2017). Coalition governments and party competition: political communication strategies of coalition parties. *Polit. Sci. Res. Methods* 5, 333–349. doi: 10.1017/psrm.2015.56

Schwemmer, C., and Wieczorek, O. (2020). The methodological divide of sociology: evidence from two decades of journal publications. *Sociology* 54, 3–21. doi: 10.1177/0038038519853146

Selivanov, D., Bickel, M., and Wang, Q. (2020). text2vec: Modern Text Mining Framework for R. *R package version* 0.6.

Slapin, J. B., and Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *Am. J. Pol. Sci*. 52, 705–722. doi: 10.1111/j.1540-5907.2008.00338.x

Stephens-Davidowitz, S. (2014). The cost of racial animus on a black candidate: Evidence using Google search data. *J. Public Econ*. 118, 26–40. doi: 10.1016/j.jpubeco.2014.04.010

Terman, R. (2017). Islamophobia and media portrayals of muslim women: a computational text analysis of US news coverage. *Int. Stud. Q*. 61, 489–502. doi: 10.1093/isq/sqx051

Tumasjan, A., Sprenger, T., Sandner, P., and Welpe, I. (2010). Predicting elections with twitter: what 140 characters reveal about political sentiment. *Proc. Int. AAAI Conf. Web Soc. Media*, Washington, DC, 4, 178–185.

Tymann, K., Lutz, M., Palsbröker, P., and Gips, C. (2019). "GerVADER-A german adaptation of the VADER sentiment analysis tool for social media texts," in *Proceedings of the Conference on "Lernen, Wissen, Daten, Analysen"*, eds R. Jäschke and M. Weidlich (Berlin), 178–189.

van Atteveldt, W., and Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Commun. Methods Meas*. 12, 81–92. doi: 10.1080/19312458.2018.1458084

van Atteveldt, W., van der Velden, M. A. C. G., and Boukes, M. (2021). The validity of sentiment analysis: comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Commun. Methods Meas*. 15, 121–140. doi: 10.1080/19312458.2020.1869198

Wagner, C., Graells-Garrido, E., Garcia, D., and Menczer, F. (2016). Women through the glass ceiling: gender asymmetries in Wikipedia. *EPJ Data Sci*. 5, 1–24. doi: 10.1140/epjds/s13688-016-0066-4

Waltinger, U. (2010). "GermanPolarityClues: a lexical resource for german sentiment analysis," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (Valletta: European Language Resources Association).

Walton, S., and Boon, B. (2014). Engaging with a Laclau Mouffe informed discourse analysis: a proposed framework. *Qual. Res. Organ. Manag*. 9, 351–370. doi: 10.1108/QROM-10-2012-1106

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Lang. Resour. Evaluat*. 39, 165–210. doi: 10.1007/s10579-005-7880-9

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Sociology

METHODS
published: 06 May 2022
doi: 10.3389/fsoc.2022.886498

A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts

Roman Egger[1] *and Joanne Yu*[2]*

[1] *Innovation and Management in Tourism, Salzburg University of Applied Sciences, Salzburg, Austria,* [2] *Department of Tourism and Service Management, Modul University Vienna, Vienna, Austria*

The richness of social media data has opened a new avenue for social science research to gain insights into human behaviors and experiences. In particular, emerging data-driven approaches relying on topic models provide entirely new perspectives on interpreting social phenomena. However, the short, text-heavy, and unstructured nature of social media content often leads to methodological challenges in both data collection and analysis. In order to bridge the developing field of computational science and empirical social research, this study aims to evaluate the performance of four topic modeling techniques; namely latent Dirichlet allocation (LDA), non-negative matrix factorization (NMF), Top2Vec, and BERTopic. In view of the interplay between human relations and digital media, this research takes Twitter posts as the reference point and assesses the performance of different algorithms concerning their strengths and weaknesses in a social science context. Based on certain details during the analytical procedures and on quality issues, this research sheds light on the efficacy of using BERTopic and NMF to analyze Twitter data.

Keywords: topic model, machine learning, LDA, Top2Vec, BERTopic, NMF, Twitter, covid travel

OPEN ACCESS

Edited by:
Dimitri Prandner, Johannes Kepler University of Linz, Austria

Reviewed by:
Tobias Wolbring, University of Erlangen Nuremberg, Germany
Ruben Bach, University of Mannheim, Germany

***Correspondence:**
Joanne Yu
joanne.yu@modul.ac.at

Specialty section:
This article was submitted to Sociological Theory, a section of the journal Frontiers in Sociology

Received: *28 February 2022*
Accepted: *19 April 2022*
Published: *06 May 2022*

Citation:
Egger R and Yu J (2022) A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. Front. Sociol. 7:886498. doi: 10.3389/fsoc.2022.886498

INTRODUCTION

With its limitless availability of constantly growing datasets and simultaneous increase in computing power, the era of digital transformation has brought about the potential to alter social science (Lazer and Radford, 2017). These massive volumes of data assemble digital footprints and capture cumulative human activities, both individually and collectively (Boccia Artieri et al., 2021). As such, the rise of big data in the twenty-first century has prompted a demand for advanced analytic techniques such as machine learning, natural language processing (NLP), and topic modeling in order to uncover patterns and relations embedded in the data, reduce the dimensionality of data, and forecast future outcomes more effectively (Elragal and Klischewski, 2017). In particular, the use of topic modeling in social science [e.g., conventional models such as Dirichlet allocation (LDA) and non-negative matrix factorization (NMF)] has soared in popularity across various domains in the past years (Maier et al., 2018; Chen et al., 2019). These techniques rely on statistical modeling to extract topical patterns within a collection of texts (Egger and Yu, 2021). For instance, since a semantic relationship exists between terms like "apple," "pear," and "mango," they could be formed under a topic called "fruit" in a text corpus (i.e., a collection of documents). Typically, documents contain mixed membership, which means that a mixture of topics exists in the corpus (Maier et al., 2018).

To unfold the complex nature of social phenomena, topic models act as a bridge between social science and (un)structured analysis, different methods of reasoning, and big data analytics (Hannigan et al., 2019) due to their explorative character (Albalawi et al., 2020). In social science, implications of big data can range from macro-level analyses (e.g., social structure and human behavior) to micro-level analyses (e.g., individual relationships and aspects of daily activities). Based on observed phenomena and experiences, examples can be noted from a growing amount of literature analyzing the news (Chen et al., 2019), online reviews (Bi et al., 2019), and social media content (Yu and Egger, 2021), amongst others. Yet, while the discussion of big data in social science mainly circles around the critical perspective of the subject, the application itself is hardly ever deliberated. Although big data seems exceptionally promising, data is always preconfigured through beliefs and values, and numerous challenges must be acknowledged as every step in big data analysis depends on various decisive criteria, such as the selection of parameters, the evaluation of partial results, and the actual interpretations thereof (Lupton, 2015). With recent advancement in the NLP field, emerging modeling techniques such as BERTopic (Grootendorst, 2022) and Top2Vec (Angelov, 2020) further complicate the process of big data analytics, pressing the need to evaluate the performance of different algorithms. Additionally, while social scientists are interested in theory-based assumptions and their implications, data scientists focus on discovering new patterns (Cai and Zhou, 2016) that appear to be irrational due to their limited explanatory power for social phenomena (McFarland et al., 2016).

Social media has opened an entirely new path for social science research, especially when it comes to the overlap between human relations and technology. In this respect, the value of user-generated content on social media platforms has been well-established and acknowledged since their rich and subjective information allows for favorable computational analysis (Hu, 2012). For instance, recent research explored the social dynamics of sporting events based on Facebook comments (Moreau et al., 2021), while another study disclosed the social semiotics of different attractions using Instagram content (Arefieva et al., 2021). Scholars have also used Twitter posts related to the COVID-19 pandemic to construct individual's reactions (Boccia Artieri et al., 2021). From an epistemological viewpoint, what is common among these data-driven approaches is that they provide brand-new perspectives on interpreting a phenomenon and have the possibility to revamp state-of-the-art knowledge (Simsek et al., 2019). After all, many aspects of social science and social media intertwine in one way or another; while the former concerns human interaction, the latter escalates its essence to a much broader and global scale.

Nevertheless, despite the prominence of social media in today's society, posts are often text-heavy and unstructured, thereby complicating the process of data analysis (Egger and Yu, 2021). Such methodological challenges are particularly salient for those lacking programming knowledge and skills (Kraska et al., 2013). Certainly, recent advancements in visual programming software have enabled researchers to analyze social media data in a coding-free manner using topic modeling (Yu and Egger, 2021), yet the validity and quality of the findings based on such intuition remain questionable. One common misconception that may skew results is the use of default hyperparameter settings. Although the importance of model tuning has been frequently acknowledged (Zhou et al., 2017), little guidance can be found when analyzing social media data in social science. In addition, another barrier that hinders knowledge generation in social science contexts is the application of more traditional and commonly-adopted algorithms (Blair et al., 2020). For example, despite the popularity of LDA, the reliability and validity of results have been criticized since model evaluation is left behind (Egger and Yu, 2021).

Consequently, some social scientists have initiated a call to conduct more interdisciplinary research and evaluate model performance based on other new and emerging techniques (Reisenbichler and Reutterer, 2019; Albalawi et al., 2020; Egger and Yu, 2021). Appertaining to the insufficient knowledge of newly developed algorithms that could better handle the nature of social media data in social science, this study thus aims to evaluate and compare the performance of four topic modeling techniques, namely, LDA, NMF, Top2Vec, and BERTopic. Specifically, LDA is a generative statistical model, NMF uses a linear algebra approach for topic extraction, and BERTopic and Top2Vec use an embedding approach. By bridging the discipline of data science with social science, reviews of the strengths, and weaknesses of different tools are valuable to support applied social scientists in choosing appropriate methods. This research sheds light on the capabilities of alternative solutions that can facilitate social science scholars in coping with any methodological issues when addressing big data.

LITERATURE REVIEW

Making Sense of Social Media Using Machine Learning Models

With the omnipresent use of technologies, human communication has transcended time and space, both locally and globally (Joubert and Costas, 2019). Among the various types of communication tools, social media stands out as a vital medium in mediating and facilitating interactions between social actors (Murthy, 2012). As social media portrays human behavior and interactions, social scientists have proceeded with data mining (Boccia Artieri et al., 2021) and using NLP and machine learning approaches. In order to understand the vast numbers of posts shared on social media, NLP can comprehend human languages, as programmed for machines, to make predictions based on the observed social phenomena (Hannigan et al., 2019). On the other side, machine learning, as a part of artificial intelligence, refers to computational methods using existing databases (i.e., the training data) to build and train a model for prediction and better decision making (Zhou et al., 2017). The advantages of opening new horizons for sociological consideration through advanced data analytics can be witnessed in manifold contexts, including business, healthcare, education, and, more generally, the role of social activities in developing scientific knowledge (Yang et al., 2020).

Previous research has underlined that the digital revolution presents dynamics in exchange networks (Joubert and Costas, 2019) and implies one's self-perception (Murthy, 2012). Examples can be seen from microblogging sites such as Twitter, accumulating over 200 million daily active users. As social media transforms interactions into relationships, and those interactions evolve into experiences (Witkemper et al., 2012), continuous status updates are seen and valued as self-production (Murthy, 2012) and, thus, allow scientists to assess perspectives from the public's point of view (Joubert and Costas, 2019). For instance, in infodemiology, Xue et al. (2020) applied machine learning models to monitor public responses in relation to the COVID-19 discussion and concerns on Twitter. Likewise, in the highly-dynamic tourism industry, Lu and Zheng (2021) were able to track public opinions toward cruise ships during the COVID-19 pandemic based on collected tweets. Furthermore, unlike most networking platforms built upon existing friendships, the retweet function can disseminate information much faster (Park et al., 2016), thereby making Twitter an ideal medium for social science research.

Yet, regardless of which social media platform, theorization remains an integral part (Müller et al., 2016) of the emerging subject of big data in social science. Although some scholars believe that big data can, and should, be free of theory altogether (Anderson, 2008; Kitchin, 2014), it seems improbable to interpret results without a sufficient understanding of the social sciences (Mazanec, 2020). Nevertheless, methodological challenges often present themselves in parallel with epistemological developments. For instance, because algorithms are unable to structure free text, data preprocessing steps that require complex decision-making skills, such as cleaning, transformation, feature extraction, and vectorization, lay the foundation for further analysis (Albalawi et al., 2020). Though social scientists have the ability to preprocess the datasets, issues may arise in the following steps involving model evaluation and hyperparameter tuning (Blair et al., 2020). For the most part, these challenges can be traced back to the nature of social media content itself, which primarily consists of short, concise, text-heavy, and unstructured formats (Albalawi et al., 2020).

Topic Modeling as a Solution to Cope With Unstructured Text Data

As human language is an adaptive multilevel system, text length, syntactic complexity, and semantic plausibility have long been considered focal points in both psychology and linguistics (Bradley and Meeds, 2002). Together with the interplay between technology and modernization, their impact has also extended to social media. For instance, scholars have pointed out that shorter posts typically lead to a higher engagement rate on Facebook (Sabate et al., 2014), potentially because concise messages reduce the amount of cognitive effort needed for information processing (She et al., 2022). Across the various available types of platforms, Twitter, in particular, restricts each post to a maximum of 280 characters (Queiroz, 2018), and although these short and unstructured posts conform with social media practice, they increase the complexity for algorithms to make sense of digital interaction. Common challenges arise from using compound words, acronyms, and ungrammatical sentences (Ariffin and Tiun, 2020). Despite the productive and unexpressed nature of compound words they often complicate computational analysis (Krishna et al., 2016). Other difficulties emerge when data are meaningless (i.e., noisy data) or when there are many gaps present in the data (i.e., sparse data; Kasperiuniene et al., 2020).

In order to effectively extract features from a large corpus of text data, numerous text mining approaches have been introduced (Li et al., 2019), among which topic modeling serves as the most frequently adopted technique (Hong and Davison, 2010). In a nutshell, a topic model is a form of statistical modeling used in machine learning and NLP, as discussed earlier, that identifies hidden topical patterns within a collection of texts (Guo et al., 2017). Those viewed as the most established, go-to techniques include LDA, latent semantic analysis (LSA), and probabilistic LSA (Albalawi et al., 2020). More recently, however, newly developed algorithms such as NMF, Corex, Top2Vec, and BERTopic have also received, and are continuing to attract, increasing attention from researchers (Obadimu et al., 2019; Sánchez-Franco and Rey-Moreno, 2022). In the social sciences, topic models have formerly been applied to, for example, discover consumers' implicit preferences (Vu et al., 2019; Egger et al., 2022), identify semantic structures on Instagram (Egger and Yu, 2021), and improve recommendation systems (Shafqat and Byun, 2020). Despite the robustness of topic modeling algorithms, existing literature relies primarily on one single model, with LDA being the dominant method (Gallagher et al., 2017) and is typically viewed as the standard approach.

Regardless of the popularity of LDA within the social science branch, its efficacy in analyzing social media data has been highly criticized (Egger and Yu, 2021; Sánchez-Franco and Rey-Moreno, 2022). In the case of Twitter data, Jaradat and Matskin (2019) argue that, while multiple topics can coexist in a document, LDA tends to neglect co-occurrence relations. Likewise, other researchers emphasize that noisy and sparse datasets are unsuitable for LDA (Chen et al., 2019) due to a lack of features for statistical learning (Cai et al., 2018). Consequently, researchers have reinforced the value of newly developed algorithms as alternatives since they often outperform LDA, especially when analyzing short text data on social media (Egger, 2022b). Albeit new approaches have emerged and have been adopted to reveal novel insights, their innovative advantages (unintentionally) lower the significance of model evaluation. Evidence can be taken from social media research, to which applying evaluation techniques is yet to become mainstream (Reisenbichler and Reutterer, 2019). Furthermore, because models would be optimized in extracting any slight variant of a topic, depending on the purpose of the algorithm, the results might be skewed in a specific direction. These issues further highlight the unreliability of concentrating solely on one single topic model and, thereby, also strengthening the value and need to compare differing algorithms (Reisenbichler and Reutterer, 2019; Albalawi et al., 2020; Egger and Yu, 2021).

MATERIALS AND METHODS

Intrigued by the complexity of short-text social media data, the goal of this research is to compare different types of topic modeling algorithms in order to offer new insights and solutions to social scientists interested in investigating human interactions. Compared to other platforms, Twitter features concise posts, with a maximum of 280 characters per tweet, that can be identified *via* specific hashtags (Queiroz, 2018). The use of hashtags thus streamlines the information search process based on users' interests. Seeing the potential of social media in enhancing crisis communication (Femenia-Serra et al., 2022), this study makes use of Twitter posts related to travel and the COVID-19 pandemic as reference points for the evaluation of the four above-mentioned topic models (i.e., LDA, NMF, Top2Vec, and BERTopic). The detailed implementation process of this study proceeded as below.

Data Collection and Preprocessing

Data collection was conducted in November 2021 by using the data extraction software tool Phantombuster and searching for the terms #covidtravel as well as the combination of #covid and #travel to fetch tweets. The initial datasets included a total of 50,000 tweets posted in English; however, after cleaning the data and removing duplicate posts, the final datasets consisted of 31,800 unique tweets. After that, the data underwent preprocessing in which all mentions (e.g., @users), hashtags, unknown signs, and emojis were removed. It is important to note that, up to this point, original sentences were used for BERTopic and Top2Vec since both algorithms rely on an embedding approach, and keeping the original structure of the text is vital for transformer models.

On the other hand, the data for LDA and NMF was preprocessed further using NLP modules in Python. More precisely, stopwords were excluded, irrelevant text (e.g., numbers, abbreviations, and unknown characters) was removed, and tokenization was performed. Following this step, stemming and lemmatization were then conducted. The former process used Porter Stemmer to remove suffixes from words (e.g., investigating to investigate), whereas the latter used WordNet Lemmatizer to remove inflectional endings and to return a word to its base form (e.g., investigating to investigate). Lastly, the text was converted into term frequency-inverse document frequency (TF-IDF) weight for information retrieval based on the importance of a keyword.

Implementation of Topic Models

Model 1: Latent Dirichlet Allocation

LDA, the most popular topic modeling technique, is a generative probabilistic model for discrete datasets such as text corpora (Blair et al., 2020). It is considered a three-level hierarchical Bayesian model, where each collection item is represented as a finite mixture over an underlying set of topics, and each topic is represented as an infinite mixture over a collection of topic probabilities. Hence, as the number of topics need not be pre-defined (Maier et al., 2018), applying LDA provides researchers with an efficient resource to obtain an explicit representation of a document.

In this research, to pinpoint optimal values for the three hyperparameters required for LDA, a grid search was performed for the number of topics (K) as well as for beta and alpha. The higher the beta, the more words the topics consist of; likewise, the higher the alpha, the more diverse the topics are. The search for an optimal number of topics started with a range from two to 15, with a step of one. In the first step of the learning process, K was pre-defined, and the search for beta and alpha was applied accordingly. During the process, only one hyperparameter varied, and the other remained unchanged until reaching the highest coherence score. The coherence score, referring to the quality of the extracted topics, presented itself for 14 topics with a value of 0.52. The grid search then yielded a symmetric distribution with a value of 0.91 for both alpha and beta. Finally, to facilitate a clear interpretation of the extracted information from a fitted LDA topic model, pyLDAvis was used to generate an intertropical distance map (Islam, 2019). A screenshot of the statistical proximity of the topics can be seen in **Figure 1**. An interactive visualization is available at https://tinyurl.com/frontiers-TM.

Model 2: Non-negative Matrix Factorization

In contrast to LDA, NMF is a decompositional, non-probabilistic algorithm using matrix factorization and belongs to the group of linear-algebraic algorithms (Egger, 2022b). NMF works on TF-IDF transformed data by breaking down a matrix into two lower-ranking matrices (Obadimu et al., 2019). Specifically, TF-IDF is a measure to evaluate the importance of a word in a collection of documents. As demonstrated in **Figure 2**, NMF decomposes its input, which is a term-document matrix (A), into a product of a terms-topics matrix (W) and a topics-documents matrix (H) (Chen et al., 2019). The values of W and H are modified iteratively, where the former contains the basis vectors, and the latter contains the corresponding weights (Chen et al., 2019). It is necessary that all entries of W and H are non-negative; otherwise, the interpretation of topics with negative values would be difficult (Lee and Seung, 1999).

Since NMF requires the data to be preprocessed, necessary steps to be performed beforehand include a classical NLP pipeline containing, amongst others, lowercasing, stopword removal, lemmatizing or stemming as well as punctuation and number removal (Egger, 2022b). For this study, an open-source Python library, Gensim, was used (Islam, 2019) to estimate the optimal number of topics. By computing the highest coherence score, 10 topics could be identified.

Model 3: Top2Vec

Top2Vec (Angelov, 2020) is a comparatively new algorithm that uses word embeddings. That is, the vectorization of text data makes it possible to locate semantically similar words, sentences, or documents within spatial proximity (Egger, 2022a). For example, words like "mom" and "dad" should be closer than words like "mom" and "apple." In this study, a pretrained embedding models, the Universal Sentence Encoder, was used to create word and document embeddings. Since word vectors

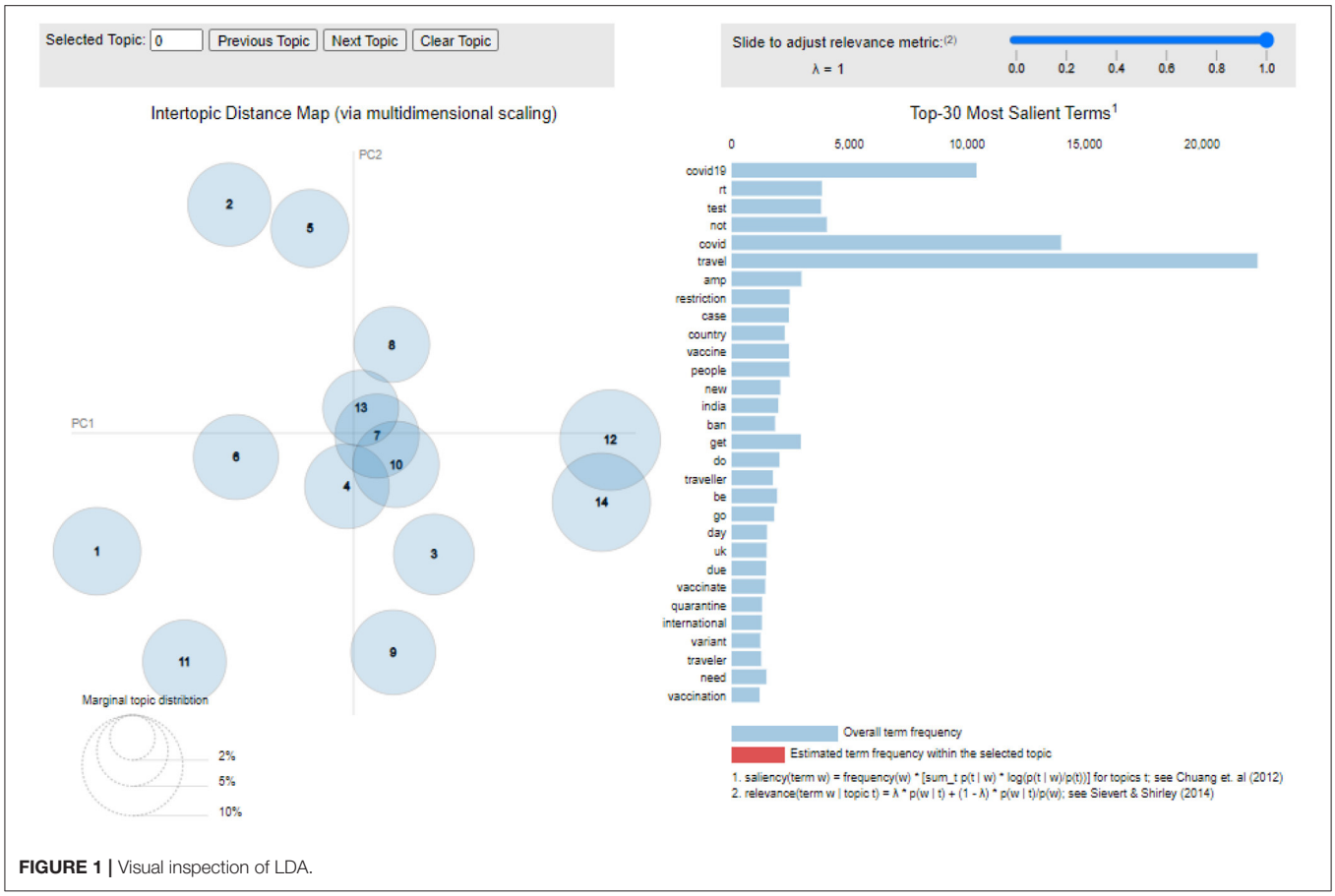


FIGURE 1 | Visual inspection of LDA.

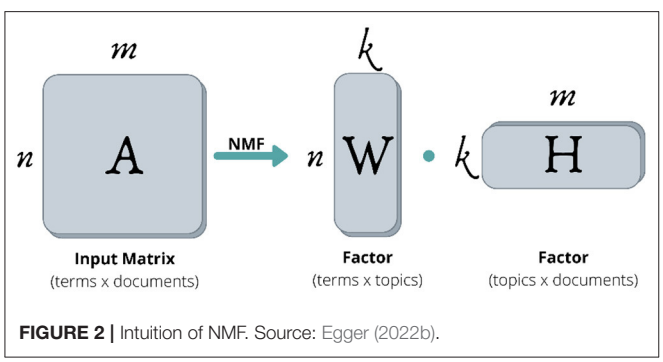


FIGURE 2 | Intuition of NMF. Source: Egger (2022b).

that emerge closest to the document vectors seem to best describe the topic of the document, the number of documents that can be grouped together represents the number of topics (Hendry et al., 2021).

However, since the vector space usually tends to be sparse (including mostly zero values), a dimension reduction was performed before density clustering. By using uniform manifold approximation and projection (UMAP), the dimensions were reduced to the extent that hierarchical density-based spatial clustering of applications with noise (HDBSCAN) could be used to identify dense regions in the documents (Angelov, 2020). Finally, the centroid of the document vectors in the original dimension was calculated for each dense area, corresponding to the topic vector.

Notably, because words that appear in multiple documents cannot be assigned to one single document, they were recognized by HDBSCAN as noise. Therefore, Top2Vec does not require any preprocessing (e.g., stopwords removal), or stemming and lemmatization (Ma et al., 2021; Thielmann et al., 2021). To conclude this model, Top2Vec automatically provided information on the number of topics, topic size, and words representing the topics.

Model 4: BERTopic

BERTopic (Grootendorst, 2020) builds upon the mechanisms of Top2Vec; hence, they are similar in terms of algorithmic structure. As the name suggests, BERT is used as an embedder, and BERTopic provides document embedding extraction, with a sentence-transformers model for more than 50 languages. Similarly, BERTopic also supports UMAP for dimension reduction and HDBSCAN for document clustering. The main difference between Top2Vec is the application of a class-based term frequency inverse document frequency (c-TF-IDF) algorithm, which compares the importance of terms within a cluster and creates term representation (Sánchez-Franco and

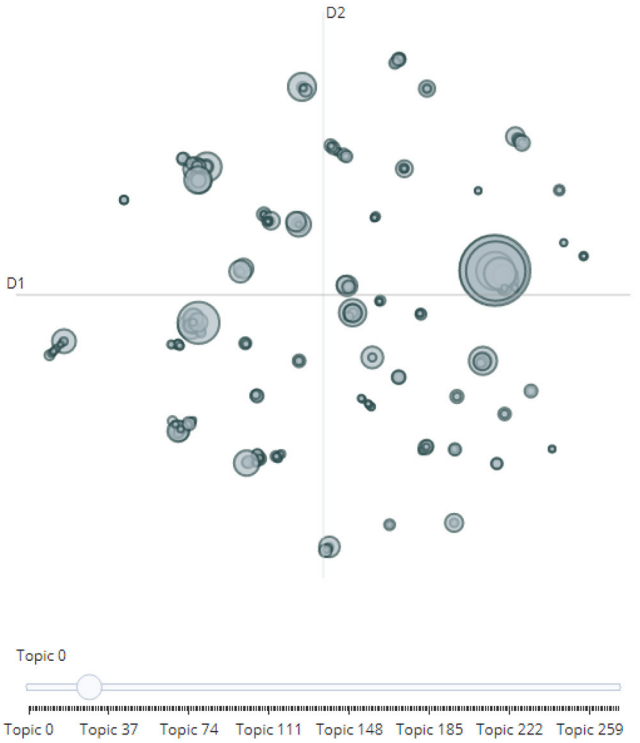


FIGURE 3 | BERTopic's interactive intertopic distance map.

Rey-Moreno, 2022). This means that the higher the value is for a term, the more representative it is of its topic.

BERTopic, similar to Top2Vec, differs from LDA because it provides continuous rather than discrete topic modeling (Alcoforado et al., 2022). The stochastic nature of the model thus leads to different results with repeated modeling. Once the model is computed, researchers can output the most important topics. Notably, Topic 0 with a count of−1 will always represent outliers and should not be considered any further. Researchers can also search for a keyword and receive the most important topics based on their similarity score along with the possibility to inspect individual topics based on their keywords. Ultimately, in order to better analyze the potentially large array of topics, BERTopic offers an interactive intertopic distance map for inspecting individual topics (Grootendorst, 2020). As illustrated in **Figure 3**, once an initial overview of the topics becomes available, an automated topic reduction can be performed again.

RESULTS

In essence, although topic models bring in statistical analysis and can advance social science research, each of the algorithms has its own uniqueness and relies on different assumptions. Quantitative methods are limited in their ability to provide in-depth contextual understanding, and the results cannot be compared with any single "value" (Egger and Yu, 2021). Thus, the interpretation of models still relies heavily on human judgment (Hannigan et al., 2019) and researchers' domain knowledge (Egger and Yu, 2022).

In the following section, a comparison of the obtained results will be divided into two parts, according to the nature of the algorithm: (1) LDA and NMF and (2) Top2Vec and BERTopic. The latter highlights the term search function as one of the pros of using a guided/seeding approach to delve deeper into a specific topic.

Comparison of LDA and NMF

Table 1 provides an overview of the 14 identified topics in the LDA model and the 10 topics from NMF. Names were given based on the terms that contributed the most to a topic in reference to their TF-IDF weights. Overall, several aspects point to common themes, such as expectations toward government response, discussion on $R_{(t)}$ values, and travel restrictions in different countries. Taking "government response" as an example, tweets seem to focus on people's expectations toward the White House (e.g., #whcovidresponse) and the US president (#potus, #vp). Although both models refer to the chance to reunite with their loved ones (e.g., #loveisnottourism), LDA, in particular, points out how the COVID-19 pandemic has influenced the Diversity Visa Program (e.g., #dv2021) application. Likewise, while both models disclose Twitter users' opinions on travel ban restrictions and quarantine, the LDA results appear to be more geographically oriented. For instance, when discussing the reproduction number, European countries, India, and the UK are more frequently mentioned. On the other hand, England and Scotland appear to be the main focal point concerning travel restrictions, and as for tweets related to quarantine, LDA reveals issues surrounding the Australian border.

Still, in spite of LDA performing seemingly better up to this point, the model produces more universal and irrelevant topics that, at the same time, barely offer any meaningful implications. This can be evidenced from the final four LDA topics listed in **Table 1**, which, based on the keywords, center on travel and COVID-19 on a broader level. Therefore, despite the fact that only a few NMF topics contain country-specific terms (e.g., New Zealand, India, and the UK), its value should not be underestimated. Due to a clear distinction between all the identified topics in the NMF model, this research concludes that the results obtained from NMF are more in line with human judgment, thereby outperforming LDA in general. Yet, as mentioned above, since topic extraction with LDA and NMF relies primarily on hyperparameters, most of the results are within expectation. As both models, however, do not allow for an in-depth understanding of the phenomenon, the next section will focus on the topic models that use embedding representations.

Comparison of BERTopic and Top2Vec

By relying on an embedding model, BERTopic and Top2Vec require an interactive process for topic inspection. As such, both algorithms allow researchers to discover highly relevant topics revolving around a specific term for a more in-depth understanding. Using Top2Vec for demonstration purposes, this section begins with the intuition behind the search query. Presuming that there is an interest in topics related to the

TABLE 1 | Topics identified by LDA and NMF.

	LDA		NMF	
No.	**Topic/content**	**Keywords**	**Topic/content**	**Keywords**
1	Government response	ban, travelgov, potus, dv2021, loveisnottourism, whcovidresponse, end, visa, please, vp	Government response	whcovidresponse, potus, loveisnottourism, cdcdirector, presssec, vp, cdctravel, cdcgov, liftthetravelban, cdctravel cdcdirector
2	Association for Molecular Pathology (AMP) / mask and virus	amp, travel, come, spread, mask, place, follow, stay, keep, virus	Association for Molecular Pathology (AMP) / desire to travel	covid, travel, people, amp, want, covid travel, time, travel covid, like, year
3	R_t value / India, UK, Europe	rt, travel, country, India, uk, covid, government, list, eu, news	R_t value	rt, covid, travel, https, covid19, traveler, rt ollysmithtravel, traveler, httpstco, ollysmithtravel
4	Travel restriction / England and Scotland	travel, covid, restriction, city, team, England, despite, event, expect, Scotland	Travel restriction	restriction, travel restriction, covid travel, covid19 travel, ease, covid restriction, travel, lift, covid19 restriction, restriction lift
5	Vaccination / border between Canada and the USA	vaccinate, covid19, international, traveler, travel, vaccination, Canada, border, US, fully	Travel ban / India and UK	ban, India, travel ban, travel India, uk, list, country, ban travel, red, variant
6	Quarantine and lockdown / Australia	traveler, day, quarantine, variant, allow, return, lockdown, Australia, break, two	General about travel / Canada	covid19, travel, covid19 travel, international, travel covid19, country, pandemic, international travel, vaccination, Canada
7	COVID-19 cases / USA	case, new, travel, health, state, tourism, public, number, close, include	Vaccination and quarantine	vaccinate, fully, fully vaccinate, vaccinate covid19, traveler, vaccinate traveler, traveler, quarantine, cdc, require
8	Flight / COVID-19 test	test, travel, need, positive, covid, flight, negative, air, take, airport	COVID-19 cases / New Zealand	case, new, covid case, covid19 case, new case, rise, Zealand, New Zealand, report, case covid19
9	Death / Florida	covid, die, death, cause, florida, child, spike, shoot, traveler002, flu	COVID-19 test	test, covid test, negative, positive, test travel, test positive, PCR, covid19 test, day, result
10	China and USA	travel, covid, call, china, business, 2020, trump, usa, dr	Vaccination pass	vaccine, covid19 vaccine, covid vaccine, passport, vaccine passport, require, vaccine travel, dose, mandate, vaccination
11	Unspecific I	not, covid, vaccine, people, do, travel, get, make, still, would		
12	Unspecific II	travel, may, covid, 2, please, 1, help, show, 3, pass		
13	Unspecific III	covid19, travel, due, pandemic, world, today, first, update, coronavirus, safe		
14	Unspecific IV	covid, be, go, travel, time, get, want, one, year, see		

term "cancel" during COVID-19, the Top2Vec model produces relevant outputs (topics) based on the order of their cosine similarity (Ghasiya and Okamura, 2021). Specifically, cosine similarity, ranging from 0 to 1, measures the similarity between the search term and a topic. In the case of this research, out of 309 topics, the similarity of Topic 10 proved to be the highest [*0.50*], followed by Topic 20 [*0.37*], Topic 7 [*0.33*], Topic 123 [*0.32*], and Topic 57 [*0.30*].

Thereafter, the most important keywords for each individual topic can be retrieved. For example, the keywords for Topic 10 include the following:

["refund," "booked," "ticket," "cancelled," "tickets," "booking," "cancel," "flight," "my," "hi," "trip," "phone," "email," "myself," "hello," "couldn," "pls," "having," "guys," "am," "sir," "supposed," "hopefully," "me," "excited," "postpone," "so," "days," "dad," "paid," "option," "customers," "request," "bihar," "thanks," "amount," "due," "waiting," "to," "got," "back," "impossible," "service," "hours," "complete," "before," "wait," "nice," "valid," "book"].

In order to acquire an overview of the importance of each term, a word cloud can be produced for better visualization (see **Figure 4**); but, ultimately, an inspection of individual tweets is also highly recommended. For instance, the findings suggest that document 20189 (tweets: "*@PaytmTravel Flight - AI 380 dated 9th April, 2020 (Canceled due to COVID). No Refund since then [. . .]*") has a similarity score of *0.8518*. This information allows one to gain deeper insights directly from the raw data. Meanwhile,

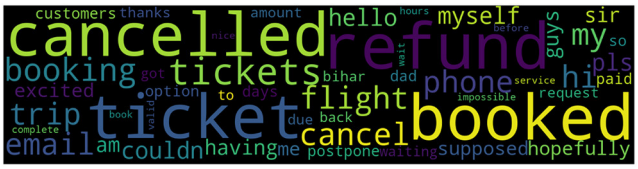


FIGURE 4 | Example of a word cloud based on the term "cancel."

in order to find more suitable keywords based on "cancel" for even further analysis, words that are most similar can be output with their similarity, such as "canceled [*0.60*]," "refund [*0.49*]," "booked [*0.47*]," "due [*0.46*]," and "ticket [*0.43*]."

Following the search process, a topic comparison between Top2Vec and BERTopic could be established. This time, "flight" and "travel bubble" were taken as other examples. Since cosine similarity has previously been introduced, the following section merely lists some of the keywords that facilitate topic naming. As mentioned above, this is because the results require human interpretation to make sense of the data (Hannigan et al., 2019).

Starting with "flight," **Table 2** provides an overview, out of the 343 identified topics, of the six most relevant ones taken from BERTopic and five, out of 253, from Top2Vec. Overall, Top2Vec topics appear to be more policy- and regulation-oriented, focusing on pre-departure testing requirements (e.g., negative PCR test and full vaccination) in countries such as Mexico, the Netherlands, and Canada. It also discusses the government's travel advice for public transport, such as in trains, buses, and flights. For a more qualitative inspection, relevant tweets can be reviewed; take, for example, "*Kind attention dear passengers traveling to [. . .] Please follow COVID-19 norms at the airport. Fly safe!*" and "*[My] flight [got] canceled by airlines due to covid. Also my travel insurance premium wasted*." On the other hand, topics identified by BERTopic are more related to the nature of air transport. Specifically, common issues shared on Twitter include the airline industry, flight routes, returning home, transmission through air, and air travel associations.

Turning to "travel bubble," both algorithms produced five relevant topics, as presented in **Table 3**. In this case, the BERTopic results seem to be more specific, with a clear distinction on travel between Australia and New Zealand, Singapore and Hong Kong, as well as Canada and Mexico. Other issues center on travel passes and business travel. With regards to Top2Vec, however, the results revealed a slight overlap. For example, the travel bubble between Australia and New Zealand is covered in four out of five topics; similarly, Singapore, Hong Kong, and Taiwan are also mentioned several times. In addition, Top2Vec produces topics with multiple aspects, which becomes especially apparent in the third and fourth topics. The third topic contains issues related to six different countries (i.e., Hong Kong, Singapore, Australia, New Zealand, the UK, and the Philippines), and the fourth includes quarantine regulations in eight countries (i.e., Singapore, Australia, New Zealand, Taiwan, Hong Kong, Korea, Hawaii, and Indonesia).

As a final note, when inspecting the keywords of BERTopic and Top2Vec, despite the redundancy of some terms (e.g., "travel bubble" and "travelbubble," as they are very close in the same vector-space), they can, in fact, provide valuable insights, especially for the process of topic naming. Mostly, the content of a topic can be understood based on frequently-repeated keywords. Moreover, regarding the logic of the algorithm, since BERTopic and Top2Vec should not be preprocessed, conjunction words (e.g., after, before to, from, at) are helpful for connecting the context. However, a major drawback without preprocessing is that (in)definite articles or be-verbs appearing in the keywords lists are often meaningless in comprehending a topic.

Hierarchical Topic Reduction of Top2Vec and BERTopic

Finally, it is worth noting that both Top2Vec and BERTopic allow for hierarchical reduction. Echoing this study's results, the number of extracted topics tends to be relatively large, thereby necessitating the need for intensive qualitative analysis. In order to streamline the analysis, the algorithms offer the possibility to reduce these topics further (Angelov, 2020). Starting with Top2Vec, a hierarchical reduction down to 10 topics is typically considered a good starting point to begin topic analysis. In the case of this research, the 10 remaining clusters deducted from the 253 original topics are presented in **Table 4**. Significantly, the original vectors remain after topic reduction, meaning that representative topics with keywords can still be sought after at any time.

Turning to BERTopic, since some of the topics are close in proximity, as could be observed in the intertopic distance map (**Figure 3**), visualization and topic reduction would provide a better understanding of how the topics truly relate to each other. To reduce the number of topics, hierarchical clustering was performed based on the cosine distance matrix between topic embeddings. This study thus took 100 topics as an example to provide an overview of how and to which extent topics can be reduced (**Figure 5**). Level 0 of the dendrogram demonstrates how similar topics (those with the same colors) have been clustered together. For example, Topic 4 (vaccine passports) and Topic 8 (the NHS COVID-19 app) were grouped together because of their adjacency. Correspondingly, Topic 6 (wearing face masks) and Topic 96 (mask mandate) were treated as part of the same cluster. In essence, a visualization as such can help researchers to better comprehend the algorithm's criteria by which topics are organized. After reviewing the proposed topic structure, researchers can then decide on a number of topics that also seem to be more realistic in an interactive manner.

However, for both algorithms, the underlying meanings of the topics are still subject to human interpretation. Nevertheless, although the intuition is to provide the best possible results, an optimal number of topics could not be established because most of the topics overlap with one another and cover a mixture of two to three different aspects. For instance, the results from Top2Vec (**Table 4**) present five topics associated with the US Diversity Visa program (e.g., dv, selectees fault, winners, an excuse, justice, interview, the petition, exam) and several terms related to

TABLE 2 | Topics identified by BERTopic and Top2Vec for "flight."

	BERTopic		Top2Vec	
No.	**Topic/content**	**Examples of keywords**	**Topic/content**	**Examples of keywords**
1	Airline industry	air travel, airline, air travel is, airlines, aviation, flights, the airline industry, the airline, airline industry, flight	Negative PCR / vaccination and quarantine	hours before, pre-departure, negative covid, all travelers, fully vaccinated, pcr, quarantine, days, requirement, mandatory
2	Flight routes	flights from, flights, direct flights, flights from india, canada eyes policy, canada eyes, india to canada, to canada, ban on direct, as india covid19	White House Secretary Tests Positive / travel guide from governmental institution	secretary, simon, house, white, tested positive, travel guidelines, cdc, mps, travelers, to follow
3	(Unable) to return home / Australian	australians, travel ban, fly home, fly home from, who fly home, who fly, to australia, australians who fly, covid travel ban, travel ban	Negative PCR / fully vaccinated before departure / foreign travelers / Mexico	negative covid, fully vaccinated, foreign travelers, pre departure, hours before, required to, before you, to enter, pcr, mexico
4	COVID transmission through air	the air, aerosols, droplets, air, airborne, covid travels, through the air, virus travels, how covid travels, covid travels through	Negative PCR / fully vaccinated before departure / foreign travelers / the Netherlands and Canada	negative covid, departure, hours before, international travelers, fully vaccinated, biden, the united, requirement, netherlands, canadians
5	Airports Authority of India (AAI) / India	aai, airports, aai airports, airport, the airport, flights, aai is, airports are, from aai, air traffic	Follow travel guidelines on public transport (train / bus / flight) / seek help and more info	train, bus, while traveling, covid appropriate, more information, to follow, covid guidelines, mandatory, by air, please help
6	Airport news	news airport airtravel, airtravel covid19 covid19india, airport airtravel, airport airtravel covid19, travelers news airport, airtravel covid19, travel covid19, flight travel covid19, air travel associations, airports air		

TABLE 3 | Topics identified by BERTopic and Top2Vec for "travel bubble."

	BERTopic		Top2Vec	
No.	**Topic/content**	**Examples of keywords**	**Topic/content**	**Examples of keywords**
1	Australia and New Zealand	travel bubble, travel bubble with, the travel bubble, australia travel bubble, zealandaustralia travel bubble, new zealandaustralia travel, zealand travel, zealand travel bubble, bubble with australia, after travel bubble	Australia and New Zealand / quarantine hotel	sydney, victoria, queensland, australia, hotel quarantine, nz, in hotel, quarantine free, lockdown, auckland
2	Singapore and Hong Kong	bubble, travel bubble, singapore, air travel bubble, travel bubble is, bubble is, singaporehong kong air, singaporehong kong, breaking singaporehong kong, as singapore battles	Australia and New Zealand / Singapore / Taiwan / vaccinated	zealand, quarantine free, singapore, hotel quarantine, 2 weeks, isolate, vaccinated travelers, lockdown, melbourne, Taiwan
3	Travel pass	travel pass, covid travel pass, eus covid travel, eus covid, the eus covid, covid travel, summer travel, travel passes, travel passes as, launch covid travel	Hong Kong and Singapore / Australia and New Zealand / green list / vaccinated / UK / Philippines	hong kong, singapore, zero covid, taiwan, green list, australia, vaccinated travelers, philippines, zealand, business travel
4	Nonessential travel / Canada and Mexico ferry / spread of COVID-19	canada and mexico, on non-essential travel, nonessential travel at, nonessential travel, ferry crossings, crossings with canada, ferry crossings with, land and ferry, and ferry crossings, spread of covid19	Quarantine free / Singapore / Australia and New Zealand / Taiwan / Hong Kong / Korea / Hawaii / Indonesia	quarantine free, singapore, hk, auckland, taiwan, korea, sydney, hawaii, indonesia, vaccinated travelers
5	Business travel	business travel, tourism, travel industry, the travel industry, tourism industry, and tourism, travel and tourism, and tourism industry, travel and, tourism industry the	Singapore / Hong Kong / Australia / Taiwan / fully vaccinated / green list	taiwan, singapore, hong kong, business travel, zealand, australia, fully vaccinated, portugal, green list, israel

TABLE 4 | Hierarchical topic reduction of Top2Vec.

No.	Topic/content	Examples of keywords
1	Diversity visa / Student life	byron, selectees fault, bay, mask, are increasing, student, the flu, exams, forever, first wave, take, traveling, covid positive, there, hands, rapidly, want, big, stop, death, interstate, fucking, haven, market, transmission, covid appropriate, bihar, to wear, short, exam, increasing
2	Diversity visa and visa petition / freedom / international travel / COVID-19 curfew	the petition, sign, tests for, pcr covid, selectees fault, boris, ford, ontario, want, curfew, premier, the airport, free, friend, trudeau, postpone, check out, rapidly, pakistan, shot, uk, enjoy, stay at, true, thread, toronto, travel insurance, international travel, normal, many countries, variants, overseas travel, freedom, mps, interstate, red list, folks, canadians, reasons, province, bihar
3	Diversity visa / unvaccinated people / vaccinate to prevent	selectees fault, centers for, di, disease, white, labor, fauci, economy, behavior, million, not being, market, shame, europeans, kerala, americans, control, here are, millions of, trump, unvaccinated, buy, weekend, make sure, oct, and tourism, dv, jobs, to protect, shop, this weekend, of vaccination, concerns, for your, air travel, next month, vaccines, open, to ease, political, millions, virus, prevention, cover, plans to, science, mexico, tourism
4	Politicians (Grant Shapps, Justin Trudeau, Biden, Trump, Anthony Fauci) / green list countries / international travel for vaccinated people / olympics / COVID-19 passport	on vaccination, eu, covid certificate, requirement for, ban, borders to, biden, grant, shapps, president, even worse, chinese, olympics, trudeau, european, required for, digital, vaccinated travelers, fauci, many countries, justice, vaccinated travelers, travel pass, visas, other countries, trump, the federal, countries, australians, green list, law, infected, joe, the border, for fully, interstate travel, europe, open, next month, covid passports
5	Pre-COVID and first wave / dreaming of travel	first wave, shelby, battle, solutions, simon, they find, the emergence, their journey, countless, lives, future, someone, human, money, an excuse, traveling, love, before covid, dose, happy, traveled, pfizer, from china, dream, together, selectees fault, died of
6	Complaints toward the US Diversity Visa Lottery program (COVID-19 as an excuse for the delay or cancellation thereof)	an excuse, toolset, selectees fault, even worse, on vaccination, uganda, death, justice, pcr tests, new cases, arabia, interview, the highest, united states, fun, winners, crazy, for fully, for foreign, nepal, imple, clear, african, nigeria, business travel, puerto rico, brexit, the airport, requiring, singapore
7	Yellow fever and COVID-19 vaccine / Saudi Arabia / COVID-19 cases	saudi, astrazeneca, journey, arabia, stay safe, new cases, covid numbers, dose of, nhs covid, wave, wear mask, got covid, yellow fever, pass, app, pre covid, doctors, eastern
8	Travel Destinations / Prevention / Travel Measures	dv, selectees fault, blaming, lanka, covid appropriate, rapidly, european, solutions, union, they find, the emergence, winners, travel advisory, increase, nepal, prevention, the delta, travel measures, covid cases, shelby, surge in, level, do not, new cases, travel related, eu, probably, hawaii, postpone, indian, to restrict, battle, florida, are increasing, rising covid, olympics, governor
9	Negative PCR test prior to departure / fully vaccinated for international travel	proof of, departure, hours before, covid appropriate, as long, will need, covid testing, negative covid, be fully, pre departure, to show, requirement for, you must, required to, by air, foreign travelers, test for, covid test, behavior, vaccinated against, test, pcr test, pcr tests, arrival, fully vaccinated, on vaccination, requirement, of vaccination, negative test, pcr, vaccination, negative, are fully, cdc, required, for international, requirements for, distancing, to require, guidance, on arrival, days of
10	Travel bubble / Australia (several cities included) and New Zealand / Hong Kong / Scotland / quarantine free / quarantine hotel	nsw, queensland, sydney, victoria, have tested, coast, shelby, melbourne, travel bubble, zealand, quarantine free, australia, in hotel, positive for, simon, wales, traveled from, kong, covid case, positive covid, battle, tested positive, first wave, vic, greater, auckland, woman, their journey, byron, the petition, hotel quarantine, scotland, south, army

politicians based in the USA and Canada (e.g., Grant Shapps, Justin Trudeau, Joe Biden, Donald Trump, Anthony Fauci). Similarly, making sense of the hierarchical clustering produced by BERTopic (**Figure 5**) also requires an enormous effort since the topic structure changes whenever researchers experiment with a different number of topics. Despite the possibility of using existing domain know-how to search for specific topics, a feature that is inexistent in other traditional algorithms, researchers should be well aware of the aforementioned issues. The overall process contains errors, and it may be quite labor-intensive to find a number that fits human judgment.

As shown in **Figure 5** below, the dendrogram produced by BERTopic shows the agglomeration levels of the individual topics. This visualization, in particular, aids in finding an appropriate number of k-topics. Furthermore, similar to Top2Vec, a table with keywords is obtained after fusing the topics; yet, it is also highly recommended to inspect individual raw documents for more appropriate interpretations.

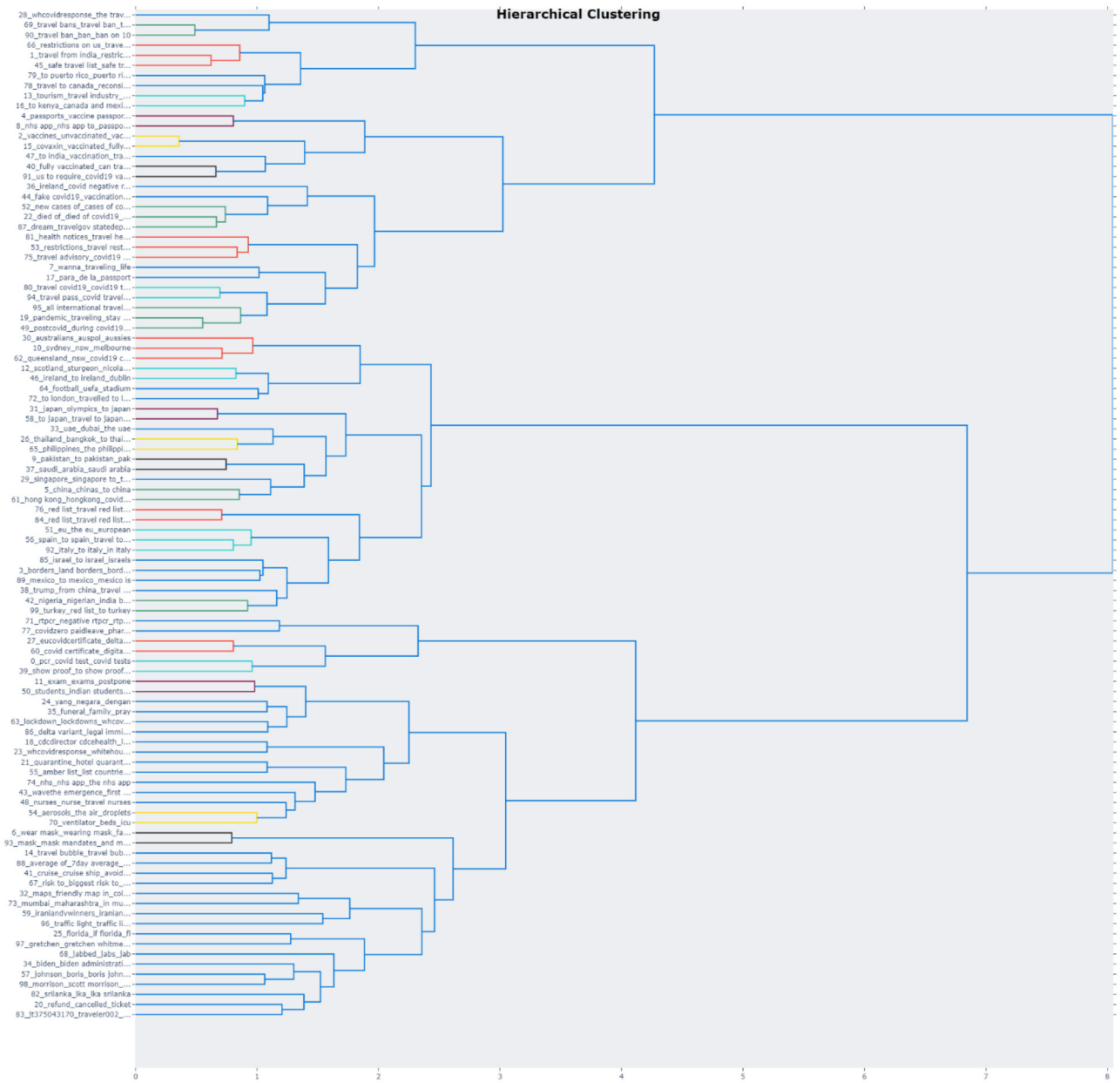


FIGURE 5 | Hierarchical reduction in BERTopic.

DISCUSSION AND CONCLUSION

Baring the difficulties of extracting useful information from short and unstructured texts in mind, this research intends to confront such challenges by comparing the results of four topic modeling algorithms. For an overall evaluation based on human interpretation, this study supports the potency of BERTopic and NMF, followed by Top2Vec and LDA, in analyzing Twitter data. While, in general, both BERTopic and NMF provide a clear cut between any identified topics, the results obtained from NMF can still be considered relatively "standard." Contrarily, in addition to the expected outcomes (i.e., topics), BERTopic was able to generate novel insights using its embedding approach. Although Top2Vec also uses pretrained embedding models, the results cover more topics that overlap and contain multiple concepts. On the other side of the spectrum, similar to NMF, the topics produced by LDA do not seem to be very intriguing, either. Thus, despite some Top2Vec topics appearing as irrelevant and difficult to understand, the model, even so, is capable of producing a few interesting findings rarely mentioned by other algorithms (e.g., politicians). As a result, in favor of extracting novel conclusions, this research recommends

Top2Vec over LDA. To provide a more solid foundation for these reasonings, a detailed evaluation for each algorithm will now be given.

First and foremost, compared to other techniques, BERTopic works exceptionally with pretrained embeddings (Sánchez-Franco and Rey-Moreno, 2022) due to a split between clustering the documents and using c-TF-IDF to extract topic representations. Especially owing to the c-TF-IDF procedure (Abuzayed and Al-Khalifa, 2021), BERTopic can support several topic modeling variations, such as guided topic modeling, dynamic topic modeling, or class-based topic modeling. Its main strength lies in the fact that the algorithm performs well on most aspects of the topic modeling domain, whereas others typically excel in one single aspect. Additionally, after having trained a BERTopic model, it is also possible to reduce the number of topics (Sánchez-Franco and Rey-Moreno, 2022), subsequently allowing researchers to settle on a number of (realistic) topics based on how many were actually produced.

Slightly different from BERTopic and the implementation of c-TF-IDF, Top2Vec creates jointly embedded word, document, and topic vectors to find topic descriptions (Angelov, 2020). The intuition behind this algorithm is that every input is considered a vector, and pivoting between them is trivial. Hence, Top2Vec can scale a large number of topics and vast quantities of data. Such strength is especially required when multiple languages emerge within a corpus (Hendry et al., 2021). The main disadvantage of Top2Vec, however, is that it is unqualified to work with a small amount of data (Abuzayed and Al-Khalifa, 2021; e.g., <1,000 documents). In fact, BERTopic and Top2Vec have a number of issues in common. For example, although outlier generation might be beneficial in some cases, the solutions might actually generate more outliers than expected. Meanwhile, another flaw involves topic distributions: they cannot be generated within a single document because each document is assigned to a single topic. Although probabilities can indeed be extracted, they are not equivalent to an actual topic distribution.

With regards to NMF and LDA, notwithstanding that both algorithms do not require social scientists to have prior domain knowledge, several topics identified by LDA in this study yielded either universal (Rizvi et al., 2019) or irrelevant (Alnusyan et al., 2020) pieces of information. Such an issue further reflects the study's findings of LDA being indeterministic (Egger and Yu, 2021). In order to achieve optimal results, LDA usually requires detailed assumptions concerning the hyperparameters; in particular, discovering the optimal number of topics typically proves to be a difficult task (Egger and Yu, 2021). Although NMF shares the same disadvantages, it can be assumed that NMF puts forward better results since the algorithm relies on TF-IDF weighting rather than raw word frequencies (Albalawi et al., 2020). Simultaneously, as a linear-algebraic model, scholars commonly agree that NMF works well with shorter texts (Chen et al., 2019), such as tweets. Since no prior knowledge is needed for topic extraction (Albalawi et al., 2020), this strength specifically benefits research based on social media data (Blair et al., 2020). Additionally, as LDA extracts independent topics from word distributions, topics that are deemed dissimilar in the document may not be identified separately (Campbell et al., 2015), thereby resulting in overlapping clusters (Passos et al., 2011). In opposition, other scholars believe that insufficient statistical information for feature extraction is the fundamental factor behind duplicate topics (Cai et al., 2018).

Lastly, when comparing BERTopic to NMF, a major shortcoming of NMF revolves around its low capability to identify embedded meanings within a corpus (Blair et al., 2020). Considering that the algorithm depends primarily on the Frobenius norm (Chen et al., 2019), which is typically useful for numerical linear algebra, this issue ultimately leads to difficulties in interpreting findings (Wang and Zhang, 2021). Though NMF can effectively analyze noisy data (Blair et al., 2020), others argue that accuracy cannot be guaranteed (Albalawi et al., 2020).

Based on the outcomes of this study, as discussed above, **Table 5** summarizes the pros and cons of applying LDA, NMF, BERTopic, and Top2Vec in order to help facilitate social scientists in the necessary preprocessing steps, proper hyperparameter tuning, and comprehensible evaluation of their results. However, researchers should take into account that, depending on the nature of the datasets, topic models may not always perform in the same fashion (Egger and Yu, 2021).

Theoretical and Practical Contributions

In light of the expansion of user-generated content, social media has broadened the horizons for human interaction and provoked new phenomena and social research for further investigation (Murthy, 2012; Rizvi et al., 2019; Boccia Artieri et al., 2021). Although several recent studies have vouched for the exploration of short-text social media data (Albalawi et al., 2020; Qiang et al., 2020), existing knowledge is rather restricted to conventional modeling techniques such as LDA and LSA (Albalawi et al., 2020). As the evolution of topic modeling has given rise to novel techniques, especially ones that have rarely been applied or evaluated in social science, this study is valuable in that it answers the call to assess topic modeling *via* a thorough comparison of four different algorithms (Reisenbichler and Reutterer, 2019). In addition, this research scrutinizes the bright and dark sides of applying embedded vs. standard topic models, but it also offers social science researchers insights into methodological challenges that may hinder knowledge generation.

Foreseeing that social scientists may indeed hesitate to choose an appropriate algorithm when analyzing social media data, this study presents possible methodological issues and promotes the efficacy of two different types of topic models. To be more precise, applying BERTopic to generate insights from short and unstructured text offers the most potential when it comes to embedding-based topic models. Thus, this study acknowledges the capability of BERTopic to encode contextual information (Chong and Chen, 2021), an aspect that may remain concealed by other models. Regarding traditional topic model algorithms, social science research is encouraged to consider NMF as an alternative approach to the commonly-adopted LDA (Gallagher et al., 2017). Certainly, however, it is essential to note that each model has its own strengths and shortcomings, and the findings require intensive qualitative interpretation. Finally, this study also strives to make another important contribution by outlining

TABLE 5 | Comparison of topic models.

	Advantages	Disadvantages
LDA	• Prior domain knowledge is not necessarily required • Finds coherent topics when correct hyperparameter tuning is applied • Can deal with sparse input • The number of topics is generally smaller than word-embedding based approaches; thus, it is easier to be interpreted • One document can contain several different topics (Mixed membership extraction) • Full generative models with multinominal distribution over topics are generated • Shows both adjectives and nouns within topics	• Detailed assumptions are required • Hyperparameters need to be tuned carefully • Results can easily produce overlapping topics as topics are soft clusters • Objective evaluation metrics are widely missing • The number of topics needs to be defined by the user(s) • Since the results are not deterministic, reliability and validity are not automatically ensured • Assumes that the topics are independent of each other; hence, only the frequency of the common occurrence of words is used • Word correlations are ignored, so no relationships between topics can be modeled
NMF	• Prior domain knowledge is not required • Supports mixed membership models; thus, one document can contain several topics • In contrast to LDA, which uses raw word frequencies, the term-document matrix can be weighted with TF-IDF • It proves to be computationally efficient and very scalable • Easy to implement	• Frequently delivers incoherent topics • The number of topics to be extracted must be defined by the user in advance • Implicit specification of probabilistic generative models
Top2Vec	• Supports hierarchical topic reduction • Allows for multilingual analysis • Automatically finds the number of topics • Creates jointly embedded word, document, and topic vectors • Contains built-in search functions (easy to go from topic to documents, search topics, etc.) • Can work on very large dataset sizes • It uses embeddings, so no preprocessing of the original data is needed	• The embedding approach might result in too many topics, requiring labor-intensive inspection of each topic • Generates many outliers • Not very suitable for small datasets (<1,000) • Each document is assigned to one topic • Objective evaluation metrics are missing
BERTopic	• High versatility and stability across domains • Allows for multilingual analysis • Supports topic modeling variations (guided topic modeling, dynamic topic modeling, or class-based topic modeling) • It uses embeddings, so no preprocessing of the original data is needed • Automatically finds the number of topics • Supports hierarchical topic reduction • Contains built-in search functions (easy to go from topic to documents, search topics, etc.) • Broader support of embedding models than Top2Vec	• The embedding approach might result in too many topics, requiring labor-intensive inspection of each topic • Generates many outliers • No topic distributions are generated within a single document; rather, each document is assigned to a single topic • Objective evaluation metrics are missing

guided modeling solutions that can be applied by social scientists to data analytics for knowledge extraction.

Limitations and Recommendations for Future Research

This research is certainly not without its limitations. While this study responds to a need to utilize Top2Vec and BERTopic for the analysis of short-text data (Egger and Yu, 2021; Sánchez-Franco and Rey-Moreno, 2022), novel language models, such as GPT3 and WuDao 2.0, have continued to emerge as time passes (Nagisetty, 2021), thereby acting as an excellent basis for even more powerful topic modeling approaches. To leverage the use of topic modeling methods, social scientists are encouraged to try and evaluate other newly developed algorithms and to keep their knowledge up to date. In the case of this study, Twitter was selected due to its strict regulations on the number of characters allowed per tweet, making it an ideal platform for exploratory research. Nonetheless, the methodological approach in this study should be applicable to other channels as well since social media posts, in general, are short and unstructured (Kasperiuniene et al., 2020). However, it is still critical to note that the nature of social media differs in terms of user demographics, text presentation, or rhetoric, amongst others. Thus, future research should continue to explore the effectiveness of topic modeling algorithms across other platforms. Lastly, acknowledging the epistemological challenges of big data is also of importance; regardless of the massive volumes of data that may appear tempting at face value, algorithms should be contextualized in a particular social framework (Egger and Yu, 2022). Although topic models have quantified short-text social media data, both

the interpretation and justification of the results come at the expense of data accuracy. Being equipped with extensive domain knowledge in data-driven science (Canali, 2016) would therefore allow social scientists to transform quantitative analytics into valuable insights for knowledge acquisition.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

RE collected and analyzed the data. JY wrote the manuscript in consultation with RE and interpreted the data. Both authors designed the study and were responsible for the overall management and planning. All authors contributed to the article and approved the submitted version.

REFERENCES

Abuzayed, A., and Al-Khalifa, H. (2021). BERT for Arabic topic modeling: an experimental study on BERTopic technique. *Proc. Comput. Sci.* 189, 191–194. doi: 10.1016/j.procs.2021.05.096

Albalawi, R., Yeap, T. H., and Benyoucef, M. (2020). Using topic modeling methods for short-text data: a comparative analysis. *Front. Artif. Intellig.* 3:42. doi: 10.3389/frai.2020.00042

Alcoforado, A., Ferraz, T. P., Gerber, R., Bustos, E., Oliveira, A. S., Veloso, B. M., et al. (2022). ZeroBERTo - leveraging zero-shot text classification by topic modeling. *arXiv [Preprint]. arXiv: 2201.01337*. Cham: Fortaleza, Portugal and Springer. Available online at: http://arxiv.org/pdf/2201.01337v1

Alnusyan, R., Almotairi, R., Almufadhi, S., Shargabi, A. A., and Alshobaili, J. (2020). "A semi-supervised approach for user reviews topic modeling and classification," in *2020 International Conference on Computing and Information Technology* (Piscataway, NJ: IEEE), 1–5. doi: 10.1109/ICCIT-144147971.2020.9213721

Anderson, C. (2008). *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*. Available online at: https://www.wired.com/2008/06/pb-theory/ (accessed February 1, 2022).

Angelov, D. (2020). *Top2Vec: Distributed Representations of Topics*. Available online at: http://arxiv.org/pdf/2008.09470v1 (accessed February 12, 2022).

Arefieva, V., Egger, R., and Yu, J. (2021). A machine learning approach to cluster destination image on Instagram. *Tour. Manag.* 85:104318. doi: 10.1016/j.tourman.2021.104318

Ariffin, S. N. A. N., and Tiun, S. (2020). Rule-based text normalization for Malay Social Media Texts. *Int. J. Adv. Comput. Sci. Appl.* 11:21. doi: 10.14569/IJACSA.2020.0111021

Bi, J.-W., Liu, Y., Fan, Z.-P., and Cambria, E. (2019). Modelling customer satisfaction from online reviews using ensemble neural network and effect-based Kano model. *Int. J. Prod. Res.* 57, 7068–7088. doi: 10.1080/00207543.2019.1574989

Blair, S. J., Bi, Y., and Mulvenna, M. D. (2020). Aggregated topic models for increasing social media topic coherence. *Appl. Intellig.* 50, 138–156. doi: 10.1007/s10489-019-01438-z

Boccia Artieri, G., Greco, F., and La Rocca, G. (2021). The construction of the meanings of #coronavirus on Twitter: an analysis of the initial reactions of the Italian people. *Int. Rev. Sociol.* 31, 287–309. doi: 10.1080/03906701.2021.1947950

Bradley, S. D., and Meeds, R. (2002). Surface-structure transformations and advertising slogans: the case for moderate syntactic complexity. *Psychol. Market.* 19, 595–619. doi: 10.1002/mar.10027

Cai, G., Sun, F., and Sha, Y. (2018). *Interactive Visualization for Topic Model Curation*. Tokyo: IUI Workshops.

Cai, T., and Zhou, Y. (2016). What should sociologists know about big data? *ISA eSymposium* 6, 1–9. Available online at: https://esymposium.isaportal.org/resources/resource/what-should-sociologists-know-about-big-data/

Campbell, J. C., Hindle, A., and Stroulia, E. (2015). Latent Dirichlet allocation: extracting topics from software engineering data. *Art Sci. Anal. Softw. Data* 9, 139–159. doi: 10.1016/B978-0-12-411519-4.00006-9

Canali, S. (2016). Big Data, epistemology and causality: knowledge in and knowledge out in EXPOsOMICS. *Big Data Soc.* 3:205395171666953. doi: 10.1177/2053951716669530

Chen, Y., Zhang, H., Liu, R., Ye, Z., and Lin, J. (2019). Experimental explorations on short text topic mining between LDA and NMF based Schemes. *Knowl. Based Syst.* 163, 1–13. doi: 10.1016/j.knosys.2018.08.011

Chong, M., and Chen, H. (2021). Racist framing through stigmatized naming: a topical and geo-locational analysis of #Chinavirus and #Chinesevirus on Twitter. *Proc. Assoc. Inform. Sci. Technol.* 58, 70–79. doi: 10.1002/pra2.437

Egger, R. (2022a). "Text representations and word embeddings. Vectorizing textual data," in *Applied Data Science in Tourism. Interdisciplinary Approaches, Methodologies and Applications*, ed R. Egger (Berlin: Springer), 16. doi: 10.1007/978-3-030-88389-8_16

Egger, R. (2022b). "Topic modelling. Modelling hidden semantic structures in textual data," in *Applied Data Science in Tourism. Interdisciplinary Approaches, Methodologies and Applications*, ed R. Egger (Berlin: Springer), 18. doi: 10.1007/978-3-030-88389-8_18

Egger, R., Pagiri, A., Prodinger, B., Liu, R., and Wettinger, F. (2022). "Topic modelling of tourist dining experiences based on the GLOBE model," in *ENTER22 e-Tourism Conference* (Berlin: Springer), 356–368. doi: 10.1007/978-3-030-94751-4_32

Egger, R., and Yu, J. (2021). Identifying hidden semantic structures in Instagram data: a topic modelling comparison. *Tour. Rev.* 2021:244. doi: 10.1108/TR-05-2021-0244

Egger, R., and Yu, J. (2022). "Epistemological challenges," in *Applied Data Science in Tourism. Interdisciplinary Approaches, Methodologies and Applications*, ed R. Egger (Berlin: Springer), 2. doi: 10.1007/978-3-030-88389-8_2

Elragal, A., and Klischewski, R. (2017). Theory-driven or process-driven prediction? Epistemological challenges of big data analytics. *J. Big Data* 4:2. doi: 10.1186/s40537-017-0079-2

Femenia-Serra, F., Gretzel, U., and Alzua-Sorzabal, A. (2022). Instagram travel influencers in #quarantine: communicative practices and roles during COVID-19. *Tour. Manag.* 89:104454. doi: 10.1016/j.tourman.2021.104454

Gallagher, R. J., Reing, K., Kale, D., and Ver Steeg, G. (2017). Anchored correlation explanation: topic modeling with minimal domain knowledge. *Trans. Assoc. Comput. Linguist.* 5, 529–542. doi: 10.1162/tacl_a_00078

Ghasiya, P., and Okamura, K. (2021). Investigating COVID-19 news across four nations: a topic modeling and sentiment analysis approach. *IEEE Access* 9, 36645–36656. doi: 10.1109/ACCESS.2021.3062875

Grootendorst, M. (2020). *BERTopic: Leveraging BERT and c-TF-IDF to Create Easily Interpretable Topics*. Zenodo. doi: 10.5281/zenodo.4430182

Grootendorst, M. (2022). *BERTopic: Neural Topic Modeling With a Class-Based TF-IDF Procedure. arXiv:2203.05794v0571*. Available online at: https://arxiv.org/pdf/2203.05794.pdf (accessed March 15, 2022).

Guo, Y., Barnes, S. J., and Jia, Q. (2017). Mining meaning from online ratings and reviews: tourist satisfaction analysis using latent dirichlet allocation. *Tour. Manag.* 59, 467–483. doi: 10.1016/j.tourman.2016.09.009

Hannigan, T. R., Haans, R. F. J., Vakili, K., Tchalian, H., Glaser, V. L., Wang, M. S., et al. (2019). Topic modeling in management research: rendering new theory from textual data. *Acad. Manag. Ann.* 13, 586–632. doi: 10.5465/annals.2017.0099

Hendry, D., Darari, F., Nurfadillah, R., Khanna, G., Sun, M., Condylis, P. C., et al. (2021). "Topic modeling for customer service chats," in *2021 International Conference on Advanced Computer Science and Information Systems* (Piscataway, NJ: IEEE), 1–6. doi: 10.1109/ICACSIS53237.2021.9631322

Hong, L., and Davison, B. D. (2010). Empirical study of topic modeling in Twitter. *Proc. First Workshop Soc. Media Analyt.* 2010, 80–88. doi: 10.1145/1964858.1964870

Hu, W. (2012). Real-time twitter sentiment toward midterm exams. *Sociol. Mind* 2, 177–184. doi: 10.4236/sm.2012.22023

Islam, T. (2019). *Yoga-Veganism: Correlation Mining of Twitter Health Data*. Anchorage, AK: Association for Computing Machinery.

Jaradat, S., and Matskin, M. (2019). "On dynamic topic models for mining social media," in *Lecture Notes in Social Networks. Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*, eds N. Agarwal, N. Dokoohaki, and S. Tokdemir (Berlin: Springer), 209–230. doi: 10.1007/978-3-319-94105-9_8

Joubert, M., and Costas, R. (2019). Getting to know science tweeters: a pilot analysis of South African twitter users tweeting about research articles. *J. Altmetr*. 2:2. doi: 10.29024/joa.8

Kasperiuniene, J., Briediene, M., and Zydziunaite, V. (2020). "Automatic content analysis of social media short texts: scoping review of methods and tools," in *Advances in Intelligent Systems and Computing. Computer Supported Qualitative Research*, eds A. P. Costa, L. P. Reis, and A. Moreira (Berlin: Springer), 89–101. doi: 10.1007/978-3-030-31787-4_7

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data Soc*. 1:205395171452848. doi: 10.1177/2053951714528481

Kraska, T., Talwalkar, A., Duchi, J. C., Griffith, R., Franklin, M. J., and Jordan, M. I. (2013). MLbase: a distributed machine-learning system. *CIDR* 1, 1–7. Available online at: http://www.cidrdb.org/cidr2013/Papers/CIDR13_Paper118.pdf

Krishna, A., Satuluri, P., Sharma, S., Kumar, A., and Goyal, P. (2016). "Compound type identification in sanskrit: what roles do the corpus and grammar play?," in *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing* (Osaka), 1–10.

Lazer, D., and Radford, J. (2017). Data ex machina: introduction to big data. *Ann. Rev. Sociol.* 43, 19–39. doi: 10.1146/annurev-soc-060116-053457

Lee, D. D., and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788–791. doi: 10.1038/44565

Li, Q., Li, S., Zhang, S., Hu, J., and Hu, J. (2019). A review of text corpus-based tourism big data mining. *Appl. Sci*. 9:3300. doi: 10.3390/app9163300

Lu, Y., and Zheng, Q. (2021). Twitter public sentiment dynamics on cruise tourism during the COVID-19 pandemic. *Curr. Iss. Tour*. 24, 892–898. doi: 10.1080/13683500.2020.1843607

Lupton, D. (2015). *The Thirteen Ps of Big Data. This Sociological Life*. Available online at: https://simplysociology.wordpress.com/2015/05/11/the-thirteen-ps-of-big-data/ (accessed February 14, 2022).

Ma, P., Zeng-Treitler, Q., and Nelson, S. J. (2021). Use of two topic modeling methods to investigate covid vaccine hesitancy. *Int. Conf. ICT Soc. Hum. Beings 2021* 384, 221–226. Available online at: https://www.ict-conf.org/wp-content/uploads/2021/07/04_202106C030_Ma.pdf

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., et al. (2018). Applying LDA topic modeling in communication research: toward a valid and reliable methodology. *Commun. Methods Measur.* 12, 93–118. doi: 10.1080/19312458.2018.1430754

Mazanec, J. A. (2020). Hidden theorizing in big data analytics: with a reference to tourism design research. *Ann. Tour. Res*. 83:102931. doi: 10.1016/j.annals.2020.102931

McFarland, D. A., Lewis, K., and Goldberg, A. (2016). Sociology in the era of big data: the ascent of forensic social science. *Am. Sociol.* 47, 12–35. doi: 10.1007/s12108-015-9291-8

Moreau, N., Roy, M., Wilson, A., and Atlani Duault, L. (2021). "Life is more important than football": comparative analysis of Tweets and Facebook comments regarding the cancellation of the 2015 African Cup of Nations in Morocco. *Int. Rev. Sociol. Sport* 56, 252–275. doi: 10.1177/1012690219899610

Müller, O., Junglas, I., vom Brocke, J., and Debortoli, S. (2016). Utilizing big data analytics for information systems research: challenges, promises and guidelines. *Eur. J. Inform. Syst*. 25, 289–302. doi: 10.1057/ejis.2016.2

Murthy, D. (2012). Towards a sociological understanding of social media: theorizing twitter. *Sociology* 46, 1059–1073 doi: 10.1177/0038038511422553

Nagisetty, V. (2021). *Domain Knowledge Guided Testing and Training of Neural Networks.* (Master's thesis), University of Waterloo, Waterloo, ON, Canada.

Obadimu, A., Mead, E., and Agarwal, N. (2019). "Identifying latent toxic features on YouTube using non-negative matrix factorization," in *The Ninth International Conference on Social Media Technologies, Communication, and Informatics* (Valencia), 1–6.

Park, S., Ok, C., and Chae, B. (2016). Using twitter data for cruise tourism marketing and research. *J. Travel Tour. Market*. 33, 885–898. doi: 10.1080/10548408.2015.1071688

Passos, A., Wallach, H. M., and McCallum, A. (2011). "Correlations and anti correlations in LDA inference," in *Proceedings of the 2011 Workshop on Challenges in Learning Hierarchical Models: Transfer Learning and Optimization (Granada)*, 1–5.

Qiang, J., Qian, Z., Li, Y., Yuan, Y., and Wu, X. (2020). Short text topic modeling techniques, applications, and performance: a survey. *IEEE Trans. Know. Data Eng*. 34, 1427–1445. doi: 10.1109/TKDE.2020.2992485

Queiroz, M. M. (2018). A framework based on Twitter and big data analytics to enhance sustainability performance. *Environ. Qual. Manag*. 28, 95–100. doi: 10.1002/tqem.21576

Reisenbichler, M., and Reutterer, T. (2019). Topic modeling in marketing: recent advances and research opportunities. *J. Bus. Econ*. 89, 327–356. doi: 10.1007/s11573-018-0915-7

Rizvi, R. F., Wang, Y., Nguyen, T., Vasilakes, J., Bian, J., He, Z., and Zhang, R. (2019). Analyzing social media data to understand consumers' information needs on dietary supplements. *Stud. Health Technol. Inform.* 264, 323–327. doi: 10.3233/SHTI190236

Sabate, F., Berbegal-Mirabent, J., Cañabate, A., and Lebherz, P. R. (2014). Factors influencing popularity of branded content in Facebook fan pages. *Eur. Manag. J.* 32, 1001–1011. doi: 10.1016/j.emj.2014.05.001

Sánchez-Franco, M. J., and Rey-Moreno, M. (2022). Do travelers' reviews depend on the destination? An analysis in coastal and urban peer-to-peer lodgings. *Psychol. Market*. 39, 441–459. doi: 10.1002/mar.21608

Shafqat, W., and Byun, Y.-C. (2020). A recommendation mechanism for under-emphasized tourist spots using topic modeling and sentiment analysis. *Sustainability* 12:320. doi: 10.3390/su12010320

She, J., Zhang, T., Chen, Q., Zhang, J., Fan, W., Wang, H., et al. (2022). Which social media posts generate the most buzz? Evidence from WeChat. *Internet Res*. 32, 273–291. doi: 10.1108/INTR-12-2019-0534

Simsek, Z., Vaara, E., Paruchuri, S., Nadkarni, S., and Shaw, J. D. (2019). New ways of seeing big data. *Acad. Manag. J*. 62, 971–978. doi: 10.5465/amj.2019.4004

Thielmann, A. F., Weisser, C., Kneib, T., and Saefken, B. (2021). "Coherence based document clustering," in *The International Conference on Learning Representations* (Online), 1–14.

Vu, H. Q., Li, G., and Law, R. (2019). Discovering implicit activity preferences in travel itineraries by topic modeling. *Tour. Manag*. 75, 435–446. doi: 10.1016/j.tourman.2019.06.011

Wang, J., and Zhang, X.-L. (2021). *Deep NMF Topic Modeling*. Available online at: http://arxiv.org/pdf/2102.12998v1 (accessed January 18, 2022).

Witkemper, C., Lim, C. H., and Waldburger, A. (2012). Social media and sports marketing: examining the motivations and constraints of Twitter users. *Sport Market. Quart*. 21, 170–183. Available online at: https://is.muni.cz/el/1423/podzim2013/ZUR589b/um/SM_W8_Twitter_Sports_Marketing.pdf

Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y., et al. (2020). Twitter discussions and emotions about the COVID-19 pandemic: machine learning approach. *J. Med. Internet Res*. 22:e20550. doi: 10.2196/20550

Yang, M., Nazir, S., Xu, Q., and Ali, S. (2020). Deep learning algorithms and multicriteria decision-making used in big data: a systematic literature review. *Complexity* 2020, 1–18. doi: 10.1155/2020/6618245

Yu, J., and Egger, R. (2021). Color and engagement in touristic Instagram pictures: a machine learning approach. *Ann. Tour. Res.* 2021:103204. doi: 10.1016/j.annals.2021.103204

Zhou, L., Pan, S., Wang, J., and Vasilakos, A. V. (2017). Machine learning on big data: opportunities and challenges. *Neurocomputing* 237, 350–361. doi: 10.1016/j.neucom.2017.01.026

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Sociology

METHODS
published: 09 June 2022
doi: 10.3389/fsoc.2022.910111

Efficient and Reliable Geocoding of German Twitter Data to Enable Spatial Data Linkage to Official Statistics and Other Data Sources

H. Long Nguyen, Dorian Tsolak, Anna Karmann, Stefan Knauff and Simon Kühne*

Faculty of Sociology, Bielefeld University, Bielefeld, Germany

More and more, social scientists are using (big) digital behavioral data for their research. In this context, the social network and microblogging platform Twitter is one of the most widely used data sources. In particular, geospatial analyses of Twitter data are proving to be fruitful for examining regional differences in user behavior and attitudes. However, ready-to-use spatial information in the form of GPS coordinates is only available for a tiny fraction of Twitter data, limiting research potential and making it difficult to link with data from other sources (e.g., official statistics and survey data) for regional analyses. We address this problem by using the free text locations provided by Twitter users in their profiles to determine the corresponding real-world locations. Since users can enter any text as a profile location, automated identification of geographic locations based on this information is highly complicated. With our method, we are able to assign over a quarter of the more than 866 million German tweets collected to real locations in Germany. This represents a vast improvement over the 0.18% of tweets in our corpus to which Twitter assigns geographic coordinates. Based on the geocoding results, we are not only able to determine a corresponding place for users with valid profile locations, but also the administrative level to which the place belongs. Enriching Twitter data with this information ensures that they can be directly linked to external data sources at different levels of aggregation. We show possible use cases for the fine-grained spatial data generated by our method and how it can be used to answer previously inaccessible research questions in the social sciences. We also provide a companion R package, `nutscoder`, to facilitate reuse of the geocoding method in this paper.

Keywords: Twitter, geocoding, spatial linkage, official statistics, regional analysis

OPEN ACCESS

Edited by:
Heinz Leitgöb, Catholic University of Eichstätt-Ingolstadt, Germany

Reviewed by:
Stefan Jünger, GESIS Leibniz Institute for the Social Sciences, Germany
Dimitri Prandner, Johannes Kepler University of Linz, Austria

****Correspondence:***
H. Long Nguyen
long.nguyen@uni-bielefeld.de

Specialty section:
This article was submitted to Sociological Theory, a section of the journal Frontiers in Sociology

Received: *31 March 2022*
Accepted: *10 May 2022*
Published: *09 June 2022*

Citation:
Nguyen HL, Tsolak D, Karmann A, Knauff S and Kühne S (2022) Efficient and Reliable Geocoding of German Twitter Data to Enable Spatial Data Linkage to Official Statistics and Other Data Sources. Front. Sociol. 7:910111. doi: 10.3389/fsoc.2022.910111

1. INTRODUCTION

Computational approaches that incorporate large volumes of online data and related methods into substantive research have become increasingly popular in the social sciences. There is now a rapidly growing literature which studies the use of digital trace data or big data for their use in social science projects (Jungherr, 2018; Stier et al., 2019; Choi, 2020). Within this literature, researchers have pointed to a number of issues that afflict many novel data types and online sources (Amaya et al., 2020; Sen et al., 2021).

Twitter is one of the most common sources for digital trace data and has been used extensively by social scientists as well as other researchers. Twitter is a microblogging platform launched in 2006 that allows users to publicly share short texts, images, or videos and to connect to and follow other users in professional or private networks. For researchers, Twitter is of particular interest, as its data is comparatively easy to access and collect (McCormick et al., 2015). Using Twitter data, researchers can study both the content of communication on Twitter—for example, by applying natural language processing techniques to large text corpora (e.g., Lwin et al., 2020)—as well as meta-information about the platform, usually to analyze networks of users (e.g., Ahmed et al., 2020). Applications of Twitter data analysis have been published in fields including political science, sociology, communication science, and public health studies (for an overview of research with Twitter data, see Karami et al., 2020).

One promising use of Twitter (meta) data is the analysis of geospatial information that accompanies tweets or user profiles (see Rieder and Kühne, 2018). Similar to research using regional properties to study survey respondents' living conditions (e.g., in urban sociology), research using Twitter data can examine the spatial distribution of tweets, compare the content of tweets across regions, or link Twitter data with external data sources by way of regional identifiers to study a variety of phenomena. Recent studies in the social sciences have used Twitter geoinformation to study the COVID-19 pandemic (Ntompras et al., 2022), influenza trends (Gao et al., 2018), crime (Hipp et al., 2018), language dialects (Huang et al., 2016), conspiracy theories (Stephens, 2020), polling (Beauchamp, 2017), travel and mobility (Blanford et al., 2015; Zhang et al., 2017; Wang et al., 2018; Levy et al., 2020), health behavior and outcomes (Wiedener and Li, 2014; Nguyen et al., 2017; Martinez et al., 2018), anti-immigrant attitudes (Menshikova and van Tubergen, 2022), happiness (Mitchell et al., 2013), and human behavior in environmental disasters (Murthy and Gross, 2017).

However, despite the vast amount of data, ready-to-use geospatial information—in the form of geographic coordinates—is only available for a small fraction of tweets. The majority of users choose not to provide the social network with GPS[1] access to their devices when sending tweets. Sloan and Morgan (2015) estimated the share of users who allow geotagging by Twitter to be 3.1%. At the tweet level, Sloan et al. (2013) estimated the share of geotagged tweets to be 0.85%. These results are supported by our analysis of over 866 million German tweets, in which the shares of tweets and users with Twitter geotags are 0.18 and 0.31% respectively. As a result, only a very small portion of Twitter data can be readily combined with external information about geographical areas, limiting the potential applications and increasing the threat of bias in estimates based on the data. For the latter, we already know from existing studies that in many countries, on average, Twitter users are more likely to be male, younger, more highly educated, wealthier, and to live in urban as opposed to rural areas (Blank, 2017; Yildiz et al., 2017; Beisch and Koch, 2021). Blank (2017) also points to systematic differences in online behaviors and attitudes that dramatically limit the potential for social science research when seeking to provide estimates for larger social groups (or even the general population). Further, Sloan and Morgan (2015) highlight additional biases in working with geotagged Twitter data by comparing users who allow geotagging of their tweets to those who do not: male and older users are more likely to share geotags and more likely to show a different set of languages in their tweets.

Clearly, adding missing but needed geographic information will increase the proportion of tweets or users that can be attributed to geographic regions, which will improve the usability of Twitter data for the study of regional context effects. In this paper, we propose a method to reliably and efficiently leverage the user-supplied free text in the location field of Twitter profiles to retrieve geographic locations as an alternative to the GPS geotags provided by Twitter. Since there are many more Twitter users who specify their profile locations than those who enable geotagging via GPS, this strategy can make a much larger portion of Twitter data usable for geospatial analysis, potentially decreasing the population bias in geotagged tweets for the analysis of regional relationships (Malik et al., 2015). Although profile locations are readily available along with tweet data, the challenge—due to the nature of the data as free text—is generally to identify as many real locations as possible while filtering out nonsensical or nonexistent locations (Hecht et al., 2011).

In addition to identifying real-world places that correspond to Twitter profile locations, we match them to (e.g., administrative) regions at different levels of spatial aggregation. Enriching Twitter data with this information ensures that it can be linked directly to regional data from other sources, such as official statistics. **Figure 1** shows the increase in the number of geolocated users achieved by our method, aggregated at the NUTS-3[2] level. While we focus on the specific case of German tweets and German administrative regions throughout this paper, our approach can easily be applied to other countries as well.

Building on our process of geocoding Twitter profile locations, we also provide `nutscoder`—a free, open-source software package in the R programming language—to help researchers implement our method in their analyses. To evaluate the results of our geocoding method, we a) assess the accuracy of the geocoded locations based on four common token-based and distance-based evaluation metrics and also compare, b) the spatial distribution of our geolocated tweets against the distribution of tweets geotagged by Twitter with respect to the distribution of real-world population as well as, and c) the

[1]To be precise, GPS is one specific satellite system for obtaining user locations. Modern technology also makes use of other systems such as GLONASS, Galileo, QZSS and BeiDou for this purpose. In this paper, however, we use GPS as an umbrella term to refer to this kind of technology.

[2]NUTS (*Nomenclature des unités territoriales statistiques*) is "a common classification of territorial units to enable the collection, compilation and dissemination of harmonized regional statistics in the EU and the UK." The NUTS system has a hierarchy of three levels. In Germany, NUTS-1 is federal states (*Bundesländer*), NUTS-2 is government regions (*Regierungsbezirke*), and NUTS-3 is districts (*Kreise*) or major, district-free cities (*kreisfreie Städte*) (European Commission, 2016).

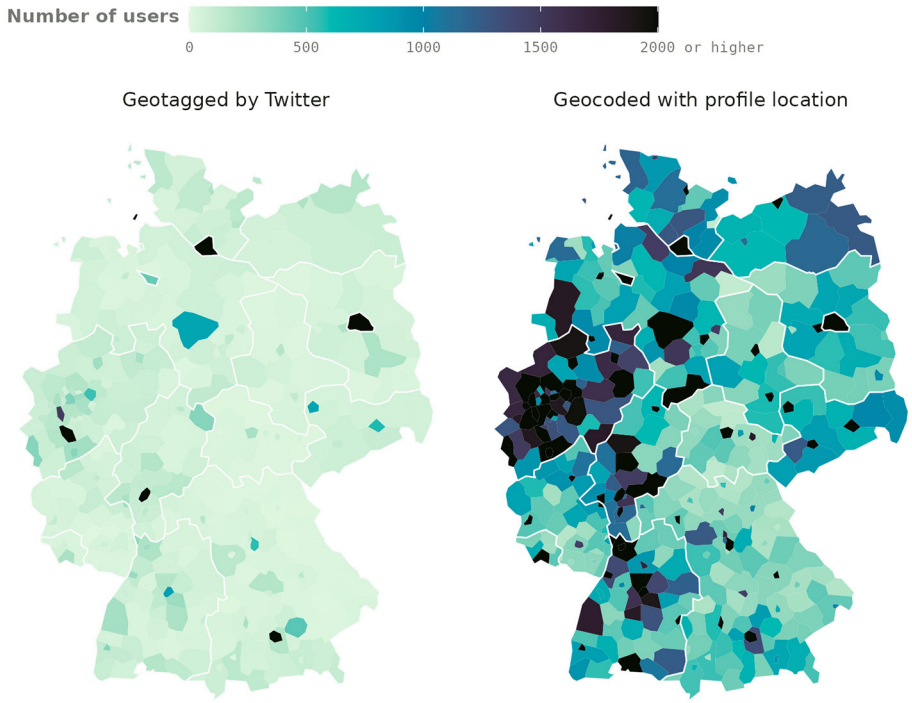


FIGURE 1 | Number of Twitter users who tweeted (retweets and tweets from verified accounts not included) between October 15, 2018, and October 14, 2021, per NUTS-3 region in Germany according to Twitter geotags and our geocoding results using user profile locations.

content of geolocated and non-geolocated tweets using a bag-of-word approach. Finally, we demonstrate the potential of our geocoded data for regional analyses in several use cases.

2. GEOLOCATION OF TWITTER DATA: BACKGROUND AND RELATED WORK

Regional analyses using Twitter data require data to be mapped to real locations of the world. Locations of tweets and users can be derived based on a variety of sources within Twitter data. The sources commonly used to locate Twitter users and tweets can be divided into three categories: a) Twitter metadata, b) Twitter user networks, and c) content of tweets (Miura et al., 2017; Zheng et al., 2018).

2.1. Twitter Metadata

Metadata is the data that accompany a tweet when a user posts it. A tweet's metadata includes information about the tweet, such as timestamp and information about the user, such as their display name and profile location text as well as GPS geotag (if available). Among these, GPS geotags are the most obvious source of location information, as they come in the form of geographic coordinates (longitude and latitude) and represent precise locations on the Earth's surface without any further processing. Thanks to their ease of use, tweet geotags are utilized by many researchers to locate tweets and users in their analysis (Mitchell et al., 2013; Hawelka et al., 2014; Wiedener and Li, 2014; Blanford et al., 2015; Shelton et al., 2015; Huang et al., 2016; Murthy and Gross, 2017; Nguyen et al., 2017; Zhang et al., 2017; Hipp et al., 2018; Martinez et al., 2018; Wang et al., 2018; Levy et al., 2020). However, this information is available for not even 1% of all tweets (Sloan and Morgan, 2015). Consequently, studies using exclusively tweets that are geotagged by Twitter limit themselves to a tiny subsample of the available data. Furthermore, the potential for more more granular regional analysis is severely restricted due to the small number of tweets available per spatial unit of analysis.

Twitter metadata provides another source for geographic locations in the user profile location field. This information is available for about two thirds of all users (Alex et al., 2016)[3], indicating the potential for much better coverage. Similar to tweet geotags, user profile locations are also intended to provide specific geographic information. Many studies to date have used location information derived from profile locations to supplement the information given by GPS geotags and provide a better sample size for analysis (Beauchamp, 2017; Stephens, 2020; Ntompras et al., 2022).

However, since user profile locations are simply free text fields for which Twitter has no constraints with regard to their correctness, many users misuse this field to state information that has nothing to do with their locations (Hecht et al., 2011). On the other hand, valid location names can take many forms due to, for

[3]This percentage refers to the number of users who sent English tweets collected in Alex et al. (2016). Analysis of our own dataset (Section 3.2) shows a similar proportion of Twitter users who provide a profile location.

example, abbreviation, capitalization, punctuation, and the order of the components of a place name. A method for geolocation based on user profile locations must therefore be able to recognize as many valid locations as possible among all available profile location text strings.

An obvious strategy for studies that use Twitter profile locations for geolocation is to employ pattern matching, for example, using regular expressions (regex), to assign profile location text to real-world location (e.g., Beauchamp, 2017). The challenge with this approach is twofold. First, the list of real location names must be large enough to cover all the regions in which the researcher is interested. This means not only having all the desired target regions, but also as many places as possible within those regions. For example, a researcher who wants to locate users in the state of Bavaria and only has the state's name in their reference list of places to match to Twitter profile locations will miss users who do not explicitly have "Bavaria" in their profile, but only the names of cities within the state such as "Munich" or "Nuremberg." Second, creating regex patterns that can reliably accommodate all possible variations in the spelling of place names is an almost impossible task. Thus, studies using *ad hoc* regex searches on user profile locations for real-world location detection are at risk of missing a significant proportion of valid location strings.

Alex et al. (2016) demonstrates a more complex approach for geolocation based on user profile locations. In this method, the Edinburgh Geoparser (Grover et al., 2010), which uses lexicon-based and rule-based named entity recognition and was originally developed to find real-world locations in regular running English text, is adapted to geolocate Twitter profile location strings and shows promising results (Alex et al., 2016). Also using specialized software—in this case, Yahoo's PlaceFinder API—to extract real-world locations from profile location text, Dredze et al. (2013) constructed a pipeline that is fast enough to return users' geographic locations in real time, proving useful for disease surveillance systems. Other applications of dedicated geolocation services and databases in the literature include the use of the Google Geocoding API and GeoNames[4] (Stephens, 2020; Ntompras et al., 2022). However, all these services are subjected to usage fees and/or restrictions regarding the size of the target name list as well as the speed of queries.

2.2. Twitter User Networks

GPS geotags and user profile locations cover the scope of Twitter data intended for the purpose of geolocation. In cases where these two pieces of information are not available, researchers must rely on other parts of Twitter data that do not explicitly refer to geographic locations but may still help to predict this information. The first of the two major approaches of this kind involves exploiting user networks—formed by interactions between Twitter users, such as following or mentioning one another—as a basis for inferring user locations. Simply put, network-based geolocation methods use available geographic information about users in a network and their relationships to predict geographic information for users for whom geographic information is not available in their metadata. This strategy relies on the assumption that users residing within the same area are more likely to communicate frequently (Ajao et al., 2015). While this is generally true (McGee et al., 2011, 2013; Jurgens, 2013), the likelihood of interactions between users also depends on a multitude of other factors, for example, users' popularity and topics of interest (Chandra et al., 2011; Li et al., 2012). A great number of methods have been developed to draw predictions about users' locations from their interaction networks (and the geographic information available from the aforementioned metadata for users in their networks), which typically involve probabilistic and machine learning models that incorporate the available spatial and network data (Backstrom et al., 2010; Davis Jr. et al., 2011; Jurgens, 2013; Rout et al., 2013; Cheng et al., 2014; Compton et al., 2014; Kong et al., 2014; Ghoorchian and Girdzijauskas, 2018). However, such methods cannot easily be scaled to real-world applications and their performance varies greatly depending on the geographic information available to be used as ground truth (Jurgens et al., 2015).

2.3. Content of Tweets

The final frequently used source of geographic information about Twitter data is the content of a tweet itself. This approach applies natural language processing methods on the text of a tweet to predict user location by leveraging words indicative of locality, for example, by being more commonly used in certain regions. Due to the unstructured nature of the data and the general complexity of the problem, geolocation methods using tweet content employ a wide range of techniques, ranging from maximum likelihood approaches to machine learning/deep learning models, both supervised and unsupervised (Cheng et al., 2010; Chandra et al., 2011; Wing and Baldridge, 2011; Roller et al., 2012; Han et al., 2013, 2014; Graham et al., 2014; Onan, 2017; Hoang and Mothe, 2018). Obviously, geolocation methods can also combine tweet content, including photos (Matsuo et al., 2017), with network data and metadata to achieve better results (Ren et al., 2012; Elmongui et al., 2015; Miura et al., 2017; Bakerman et al., 2018; Ribeiro and Pappa, 2018; Tian et al., 2020).

Compared to geolocation methods based on Twitter metadata, methods based on user networks and tweet content are more complicated because these data are not exclusively related to geographic locations, and thus geographic information in these data is sparse. Consequently, the results of network-based and content-based geolocation methods are highly uncertain in nature and generally less accurate. These methods therefore also require much more effort to validate and evaluate. Since the goal of our paper is to develop a method to geolocate data in a very large corpus of tweets in a reliable and efficient manner, Twitter metadata is the more suitable source of geographic information on which to base our method.

3. DATA

3.1. Data Collection

Data collection in from the official Twitter API started on October 5, 2018, and is still ongoing. In our queries to the Twitter

[4]www.geonames.org

API[5], we request real-time tweets that are tagged as German by Twitter's language detection and contain one of the 100 most common words—excluding punctuations and separators—in the German language[6]. The Twitter API requests return on average about 15 tweets per second (with some day-night cycle fluctuation), which amounts to 35–40 million tweets per month. While the Twitter API has a rate limit of 1% of all Twitter traffic globally, we believe this does not affect our data collection. Tromble et al. (2017) estimated the global rate to be 6,000 tweets per second in 2016, and based on Twitter's growth from 2016 to the present, we expect the amount of data that we collect to be well below the possible rate limit of about 60 tweets per second (1% of 6,000).

3.2. Dataset

Until March 2022, we have collected over 1.1 billion tweets (including retweets[7]). For the analysis in this paper, we use a subset over the 3-year period from October 15, 2018, to October 14, 2021. This subset does not include retweets. It also does not include tweets from so-called verified accounts, as these are mostly run by representatives of media and other organizations whose tweets tend to be neutral reporting of news and thus less interesting for our substantive applications in researching public attitudes and behaviors on the platform. With this restriction, our analysis sample consists of over 866 million tweets from 16.6 million users. Alongside the text of each tweet, the Twitter API provides additional information about the tweet, including a unique ID, the time of posting, the location of the device as a geographic coordinates, if available, and whether it was a retweet, as well as information about the user who posted the tweet, including a unique user ID, their username, follower count, profile description, and profile location, if available.

In order to link the data in the tweets with external data about geographical regions for use in regional analysis, we need an attribute that identifies the regions to which a tweet or its user can be assigned. When users give permission, Twitter collects their location in the form of geographic latitude and longitude. Researchers can easily pinpoint the location to which the specific latitude and longitude refer and choose the appropriate level of spatial and/or political aggregation—municipality, county, district, or state—to link the Twitter data with data from other sources.

In our dataset, however, only about 1.53 million or 0.18% of the tweets collected were tagged with geographic coordinates by Twitter. These geotagged tweets came from 51,180 Twitter users, or 0.31% of all the users in our analysis sample. This represents an even smaller amount of geographic information collected and shared by Twitter than what was reported in Sloan and Morgan (2015). This difference could be attributed to the fact that we only analyze German-language tweets, since users in Germany tend to be less willing than users in other countries to share geolocation information with their tweets (Scheffler, 2014).

TABLE 1 | NUTS-3 regions with the fewest users based on Twitter geotags.

NUTS-3	Name	Users
DEB3G	Kusel	6
DEG0D	Sömmerda	9
DE255	Schwabach	9
DE272	Kaufbeuren	10
DE22C	Dingolfing-Landau	11
DE247	Coburg	11
DE926	Holzminden	11
DE267	Haßberge	11
DEG0N	Eisenach, Stadt	11
DE234	Amberg-Sulzbach	12
DE23A	Tirschenreuth	12
DEB37	Pirmasens, kreisfreie Stadt	12
DEG06	Eichsfeld	12
DEG0A	Kyffhäuserkreis	12

If we use only those tweets in our dataset that were already geotagged by Twitter, we cannot perform meaningful regional analysis at the level of (and below) major cities (*kreisfreie Städte*) or counties (*Landkreise* or *Kreise*) in Germany. For many regions, the number of users who have at least one tweet with GPS coordinates falls in the low double-digit range or even below, with the lowest number being six (**Table 1**).

An alternative source of geographic information in Twitter data that is also easily accessible and can be exploited to increase the number of geolocated tweets is the profile's location field, in which Twitter users can enter an arbitrary text that will be displayed publicly. Assuming that the text in the profile location corresponds to a user's actual location, this information has the potential to make a much larger portion of Twitter data usable for regional analysis. In contrast to the low percentages of tweets and users with Twitter geotags, 569 million (65.66%) of our 866 million tweets (excluding retweets and tweets from verified accounts) collected during the 3-year period were posted by users who had entered something in the location field of their profiles. These users (9.2 million) make up 59.15% of the total number of users in our analysis sample (16.6 million users).

However, it has to be noted that not every Twitter user who uses the profile location field uses it for its designated purpose, as users can enter any text string 30 characters or shorter in this field. For example, many write indecipherable sequences of letters and emojis. Many others misuse this space to make their age and/or gender pronouns known. Other examples of non-location strings that users give as their location are "mind your own business," "dying of hunger," and "goat cheese radish tartine[8]."

[5]developer.twitter.com/en/docs/twitter-api

[6]Our list of the most common German words was compiled from the word list `DeReKo-2014-II-MainArchive-STT.100000` (www.ids-mannheim.de/digspra/kl/projekte/methoden/derewo/), from the Institute for German Language (www.ids-mannheim.de). Note that Twitter does not allow queries that filter tweets based solely on Twitter's language recognition. Therefore, it is necessary to provide a list of additional keywords or parameters—in our case, the 100 most common German words. A list of these words can be found in the (**Supplementary Section 1**).

[7]Retweeting is the act of sharing another user's tweet publicly on Twitter.

[8]These profile location strings are obfuscated to protect the privacy of users.

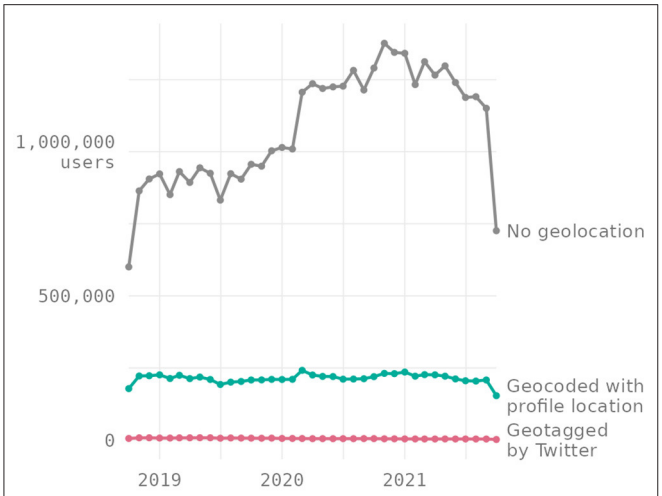


FIGURE 2 | Number of monthy active Twitter users in our dataset.

Figure 2 shows the number of monthly active users over the 3-year period in our analysis dataset, grouped by the source of geospatial information about their tweets: from Twitter's geotags, geocoded based on their profile location, or none at all. For every month, we count as active users those who posted at least once tweet during the month[9]. Across each month, the number of active users in our German Twitter dataset who could be geocoded via their profile locations is much higher than the number of users whose tweets were geotagged by Twitter, while most users, however, could not be assigned a location in Germany. Note that users are grouped by whether they could be geolocated, so the number of Twitter users with geocoded profile locations is lower than the number of users with a location text in their profiles presented above. Also, while the number of users without geographic locations shows a general upward trend with a significant jump in early 2020, this trend is not observed for the number of users with geographic locations[10].

4. GEOCODING TWITTER PROFILE LOCATIONS

4.1. Objectives

As mentioned earlier, metadata in the form of GPS coordinates needs virtually no processing, but is only available for a tiny fraction of all available tweets. Therefore, the purpose of our approach is to supplement this information with geographic information obtained from the profile location text, which is available for a large proportion of the data. Unlike geographic coordinates, locations as text strings need to be preprocessed in order to compile unambiguous geographic information, because a given place on Earth may be referred to in many ways. The process of extracting geographic information from text is called *geocoding*. Since geographic locations are unique and can often be identified as such, for example, in official statistics, free text locations in Twitter profiles need to be geocoded to enable a linkage of regional data with other data sources, which may then be leveraged for regional analysis.

A primary goal of our geocoding procedure is to discern—whenever possible—a corresponding spatial reference for a given location name in a Twitter user's profile. This means, on the one hand, that geocoding should allow for a variety of names that each location may be associated with. For example, we should be able to identify a Twitter user from the German city of Hamburg if they have a profile location that reads "Hamburg" or "HH" (its ISO code), or "Free and Hanseatic City of Hamburg" (its full official name). The language used for a place's name should also not influence where the place actually is: "Freie und Hansestadt Hamburg" (German), "Friee un Hansestadt Hamborg" (Low Saxon), "Hampuri" (Finnish), "Amburgo" (Italian), and "ハンブルク" (Japanese) should all be recognized as the same city. Furthermore, the geocoding results should not be dependent on the use of capitalization, punctuation, spaces, or the order of components in the location strings: "münchen," "MÜNCHEN BY," "München, Deutschland," and "Germany / Bavaria / Munich" should all be assigned the same spatial reference. Likewise, geocoding should also be insensitive to additional non-text elements in the location string, such as emojis and other special Unicode characters. On the other hand, the geocoding rules must also be strict enough so as not to mistake non-locations that users enter in their profile, such as those listed in Section 3, for real locations.

A second important objective of our geocoding procedure is to make it easy to determine whether an observation can be included in aggregated statistics at a certain level of spatial aggregation. In contrast to Twitter's geographic tagging with the use of GPS, the name of a region can only reveal its shape as a polygon on the surface of the Earth, but not an exact point, since a region spans a larger area. For an exact point on the Earth's surface, the associated data can be aggregated to any higher or lower regional level that encompasses that point. For polygons, however, the lowest possible level of spatial aggregation is their own boundary. Knowing the lowest possible level of aggregation for each region as well as the encompassing regions at higher levels of aggregation is important to identify the appropriate spatial reference that can be used to link Twitter data with data from other (e.g., administrative) sources, as data about regions at a lower level can be aggregated to a higher level, but data about a region at a higher level cannot be easily disaggregated to regions at a lower level. For example, if a user's profile location says "Munich," it is also non-problematic to use this observation as a part of the federal state of Bavaria, Germany in an analysis at the state level; however, the reverse is not true, since not every part of Bavaria is within the city of Munich, and a user with a profile location that says Bavaria cannot be part of an analysis of cities or other types of spatial units that are at a lower level of aggregation than federal states.

The sheer amount of data available (see Section 3) leads to an additional objective for our geocoding procedure: In order to make use of the geocoding results in our substantive research, we

[9]This definition is more conservative than the more common definition of active users, which also counts registered users who visited the platform but did not post anything.

[10]Due to the first and last date of the analysis subset being in the middle of the month, the first and last month has a substantially lower number of active users.

need to achieve the aforementioned goals for all of our collected tweets in a reasonable time. Additionally, as the data collection is ongoing, our geocoding tool chain should also be able to continuously process the new profile locations associated with the incoming tweets while avoiding repetitive geocoding of already processed locations to save time and computing resources, enabling us to establish a real-time pipeline for geocoding the collected Twitter data.

4.2. Implementation

Geocoding—the identification of geographic information based on the name of a place—is a common practice in spatial analysis that emerged and has continued to be refined over the last several decades (Goldberg et al., 2007). There are now a wide range of vendors and services available to facilitate the geocoding process, including free, open-source software solutions as well as enterprise-level products at global conglomerates like Google (Google Maps, 2022).

For our application, we opted for the open-source geocoder Nominatim, which allows users to search all of OpenStreetMap data (Nominatim, 2022b). OpenStreetMap is an initiative whose diverse contributors create and provide free geographic data about places all over the world (Map Foundation, 2021). By virtue of being free, open-source, and actively developed by a large community, both OpenStreetMap data and its search engine Nominatim offer themselves as a viable long-term solution for our purpose. Another advantage of Nominatim is the ability to geocode place names not only in English or the language of the country where a place is located, but also in many different other languages, especially for widely known place names.

Nominatim's search engine takes a text string as input and returns geographic information as well as other data from OpenStreetMap about the place in real life that corresponds to the input string. Thanks to sensible tokenization and normalization of OpenStreetMap place names as well as search input, Nominatim's text search engine can handle users' queries flexibly, also being tolerant of fuzzy matches and abbreviations (Hoffmann, 2021a,b; Nominatim, 2022d). Nominatim also provides a public instance at nominatim.openstreetmap.org, accompanied by an API that allows users to programmatically search for places in the OpenStreetMap database (Nominatim, 2022a).

It is important to note that Nominatim can return multiple places based on a given text string. This often occurs when there are multiple places with the same name, such as the US city of New York and the Munich hair salon named New York. In such cases, the places in the results are assigned a ranking based on Nominatim's internal search rank (e.g., a state has a higher search rank than a city, which has a higher rank than a suburb) or—when available—the Wikipedia importance ranking (Nominatim, 2022c). The latter is a function of the number of Wikipedia articles that are linked to a place's Wikipedia article (Nominatim, 2021). For our application, we limit the geocoding results to the first-ranked place that Nominatim returns for each location string.

By taking a list of all unique profile location strings that appear in our database, we reduce the number of cases for geocoding from 569 million tweets sent by users with a location in their profile to over 6 million location strings. After geocoding, the results can be joined back to user profiles via the location strings. However, despite the substantial reduction in the number of cases, the rate limit of 1 query per second of the public Nominatim server means that it would take us over 2 months to geocode the 6 million text strings that we have.

To overcome this problem, we host our own instances of Nominatim's database on our on-premise high-performance computing server (on which the relational database that contains all collected Twitter data is also hosted). More specifically, we deploy two Nominatim instances[11]: the first contains data for German places only and acts as a quick filter; the second covers the whole world and is used to perform the final geocoding step on the filtered profile location strings[12]. Not only does self-hosting free us from the query rate limit of the public Nominatim, it also enables complete access to Nominatim's database backend. The benefits of this are two-fold. First, we can exclude irrelevant places on the globe from the database, thus reducing the size of the database and making queries faster. Second, since this allows us to perform geocomputational operations such as spatial joins directly on objects in the database, we have flexible control over the geographic information that Nominatim queries return and are able to streamline it to our needs.

To preserve user privacy, we exclude the geocoding results in which the location text is matched with a place at the street address level, with the exception of train stations. This also greatly reduces the number of mishits, which are particularly prevalent for places at this level, as location strings containing common nouns are often matched with businesses such as shops and restaurants. For example, a user can specify their profile location as "Saturn" (presumably the planet), which is also the name of a chain of electronics stores in Germany and Luxembourg. Since there is no other place in Germany with a higher ranking that is also named Saturn, Nominatim will return the address of the Saturn store in Senden, Bavaria, which is the first-ranked result when searching for "Saturn."

In addition to geocoding the profile locations and retrieving the geographic information about the place that corresponds to each location, we create a dataset that contains the official names and codes of administrative regions at different levels in Germany as well as the geographic geometries (also commonly known as "shapes") of these regions. By performing spatial joins of the geocoded places' shapes on the shapes of the administrative

[11]The Nominatim database instances are containerized with Docker (image from github.com/mediagis/nominatim-docker) and deployed via Kubernetes, each with a maximum of 64 CPU threads and 16 GB of shared memory. Wikipedia data is imported into both instances to leverage the result ranking mechanism described above. PostGIS (postgis.net) is enabled in both Nominatim databases as well as the Twitter database to facilitate geocomputational operations.

[12]A more obvious setup would be to simply geocode all user profile locations in one pass with the global Nominatim instance. However, by first running all profile location strings through the German Nominatim instance, we can filter out a large number of irrelevant strings (i.e., non-locations or locations not in Germany) in much less time, since the German database is significantly smaller than the worldwide database (100 GB vs. 1.4 TB). Obviously, after this step, the profile location strings—now substantially fewer—still have to be geocoded with the global instance of Nominatim, since the Germany-only instance—due to the lack of data on places outside Germany—mistakes place names like "New York" for locations in Germany.

TABLE 2 | Random sample of geocoding results where the input is the Twitter profile location and the output is the corresponding administrative regions in Germany.

Profile location	NUTS-1	NUTS-2	NUTS-3
fRaNkFuRt	DE7	DE71	DE712
Aicha vorm Wald	DE2	DE22	DE228
Schwei	DE9	DE94	DE94G
Brochenzell	DE1	DE14	DE147
hh	DE6	DE60	DE600
nrw	DEA	–	–
Jena, Germany	DEG	DEG0	DEG03
Aub, Deutschland	DE2	DE26	DE26C
Germany-Mülheim an der Ruhr	DEA	DEA1	DEA16
Kuhbach im Schwarzwald	DE1	DE13	DE134

regions, we can determine all administrative regions at different levels to which a geocoded place can be assigned, as well as the lowest administrative level at which analysis can be done with the geocoded data. More precisely, a Twitter profile location is matched to an administrative region if the place that corresponds to this location lies completely within the boundaries of that region. For example, in addition to being assigned to the city of Munich, a user whose profile location reads "Munich, Germany" is also matched with the state of Bavaria as well as any administrative region that completely encompasses Munich.

Since our analysis only deals with Twitter users in Germany, only the geometries of German regions are included in the target dataset for the spatial joins. This means that profile locations referring to actual places outside of Germany such as "Vienna, Austria" are excluded from the final results, as no administrative region in Germany covers Vienna on the map. **Table 2** shows a sample of location strings and the NUTS codes of the regions that we could match with these strings using the described procedure.

To facilitate automation of the geocoding process and make it reusable in other research, we create the R package `nutscoder`, which makes it straightforward to perform the described geocoding steps to generate corresponding administrative region codes from location names as free text. `nutscoder` also generalizes our geocoding practice so that it is applicable not only to Twitter profile locations, but to any text strings that refer to real-world locations. With the ability to customize the target dataset of administrative regions, the same procedure can also be used to geocode locations outside of Germany. Without access to our private server, however, `nutscoder` can only use the public Nominatim server (or an instance of the Nominatim database and API self-hosted by the package users). The package is publicly available and can be installed from github.com/long39ng/nutscoder.

4.3. Results

In total, we are able to match German administrative regions to over 74,000 of the unique location strings available in our sample. Merging these geocoding results over the location text to the data on profiles and tweets, we obtain the geographic locations for a total of 229 million tweets—26.4% of our analysis subset.

TABLE 3 | Number of tweets per user from October 15, 2018, to October 14, 2021. Retweets and tweets from verified accounts are excluded.

	Mean	Median	SD	Max
Geocoded with profile location	230.0	9	1,939	792,298
Geotagged by Twitter	29.8	2	1,108	226,900
No geolocation	42.9	1	669	447,564

This represents a 150-fold increase over the number of tweets geotagged with GPS coordinates by Twitter (see Section 3)[13].

Perhaps surprisingly, the geocoded tweets were posted by only 6.23% (997,602 users) of all Twitter users in our dataset. A closer look at the data reveals the reason for this disproportion: **Table 3** shows that Twitter users whose profile location could be matched with administrative regions in Germany were apparently much more active according to our data. However, the underlying reason for this discrepancy may not be the inactivity of users whose profile location could not be assigned to a region in Germany, which the data seem to suggest, but rather that this group may tweet less in German and therefore appear far less frequently in our dataset.

5. EVALUATION

To evaluate the performance of our geocoding, we compare geocoding results with GPS geotags for the users for whom both these pieces of information are available, using common evaluation metrics (Section 5.1). Further, as studies have shown that the distribution of locations provided by Twitter via GPS tagging are biased in several dimensions (Malik et al., 2015; Arthur and Williams, 2019; Karami et al., 2021), we suspect similar issues with the geographic locations obtained via geocoding of user profile locations. To investigate this, we first look at whether geocoding via profile locations increases the potential bias in geolocated tweets by comparing the spatial distribution of users geolocated by Twitter and with our method (Section 5.2). Second, in Section 5.3, to assess whether geolocated tweets might differ in terms of content from non-geolocated tweets, we compare their respective bag-of-words distributions.

5.1. Geocoding Performance

Based on the assumption that GPS geotags from Twitter are the most reliable source of information about geographic locations, we use them as the basis for creating a gold standard to evaluate our geocoding results. Since GPS geotags are reported at the tweet level, the GPS–place-of-residence relation can be noisy. We apply several constraints when selecting the gold standard sample to ensure that locations provided by Twitter geotags and extracted from user profiles reflect the same underlying information (i.e.,

[13]The numbers presented in this section refer to the results of geocoding Twitter profile locations using administrative regions in Germany. This means that valid profile locations, that is, those that contain actual place names, but do not refer to locations in Germany, do not yield any results. In applications where locations outside Germany are also taken into account, the coverage provided by the geocoded user profile locations is likely to be much higher.

TABLE 4 | Performance of our geocoding method.

				Error distance (km)	
NUTS level	**N**	**Accuracy**	**Accuracy@161**	**Median**	**Mean**
NUTS-1	13,423	92.74	-	-	-
NUTS-2	12,919	90.92	-	-	-
NUTS-3	12,793	86.07	-	-	-
All levels	13,423	85.70	95.87	0	18.35

presumably the place of residence). Specifically, we select users for whom at least two geotags (which may refer to the same pair of coordinates) are covered by the same NUTS-3 region, and the geotags covered by said region account for more than half of all available geotags for the respective user. There are 13,423 users in our dataset whose geotags satisfy this condition and whose profile location could also be geocoded by our method[14]. The location to be used as the gold standard for a user is then calculated as the centroid of the geometry formed by all unique pairs of coordinates in the NUTS-3 region that covers the majority of that user's geotags.

We evaluate our geocoding results using four common metrics (Zheng et al., 2018): The first metric is accuracy, which treats location as discrete tokens and represents the percentage of cases in which the geocoded NUTS region matches the NUTS region containing the gold standard coordinates. The remaining three metrics are distance-based[15], including accuracy@161, a relaxed accuracy metric that accepts results within a distance of 161 km (100 miles) from the gold standard as correct, as well as median and mean error distance of the geocoded regions to the gold standard.

Table 4 shows the evaluation results. Our geocoding procedure achieved over 90% accuracy at the NUTS-1 and NUTS-2 levels, and over 85 at the NUTS-3 level as well as when considering geocoding results at all levels combined. Over 95% of the geocoded NUTS regions are less than 161 km from the gold standard, with the median and mean error distances at 0 and 18.35 km, respectively[16].

5.2. Spatial Distribution of Geocoded Users

As suggested above, GPS coordinates are expected to show more variability at the user level. Our data support this assumption, as users with geotags provided by Twitter have more unique locations on average (mean: 2.54, standard deviation: 5.55) than users with locations geocoded by our method (mean: 1.04, standard deviation: 0.022)[17]. Nevertheless, since the median is 1 in both cases, we can assume that most users can be assigned to one NUTS-3 region, even in the case of the geographic locations provided by Twitter.

Following the general idea that most users can be assigned to one location, that is, their primary residence, we assign each user the statistical mode of their available locations—either geocoded with profile location or geotagged by Twitter. This allows us to unambiguously link Twitter user data to data from other sources (i.e., a user can only be attributed to one region when linking with official regional statistics). For example, if a user is assigned to Berlin ten times and to Munich three times (due to changes in their profile location over time), this user will be assigned to Berlin in our analysis. If a user has multiple modes of locations (i.e., multiple locations with the highest number of tweets associated with each of those locations), we draw a random location from those.

Figure 3 shows the distribution of locations provided by Twitter and by our method compared to the general population[18]. The share of Twitter users within a NUTS-3 region shows a rank similarity to the actual share of the real population in that region. However, after including newly geolocated users based on profile locations, we find the same biases as in the Twitter geotagged sample—that is, most smaller regions are slightly underrepresented, while a few larger regions (mostly cities) are overrepresented. On the other hand, the differences in percentage point between the two samples and the actual population are small. The average absolute error[19]—which corresponds to the average vertical distance of the points to the diagonal in **Figure 3**—is 0.00173 percentage points for Twitter geographic locations and 0.00111 for geographic locations obtained via the profile locations. This is possible evidence that the observable bias compared to the general population distribution is not from the GPS-based geographic locations, but instead represents a bias inherent to the platform, i.e., general self-selection into Twitter. Nevertheless, as our user sample is 20 times larger and our tweet sample is 150 times larger, it enables a wide variety of regional analyses at finer levels of granularity. Examples of regionalized content analyses can be found in the following sections.

5.3. Content of Non-geolocated and Geolocated Tweets

As previous research has shown, geolocated tweets may be susceptible to sampling bias (Malik et al., 2015), but it is not entirely clear whether this also applies to their content. To assess potential differences between the content of non-geolocated and geolocated tweets, we compare these two samples with two

[14] An evaluation based on all users for whom Twitter geotags are available and whose profile locations could be geocoded by us (i.e., without the restrictions to filter for users in the gold standard sample in this section) is reported in the **Supplementary Section 2**.

[15] For these metrics, we calculate the distance between the gold standard point and the polygon of the geocoded NUTS region for each case.

[16] While these numbers appear to show much better overall performance compared to other methods of geocoding using Twitter profile locations, such as in Dredze et al. (2013), meaningful comparison is not possible, since they performed the geocoding on a much smaller sample of tweets that were posted by users from another country.

[17] We count the number of locations per user at the NUTS-3 level. This means that for a user, unique pairs of geocoordinates that fall within a NUTS-3 region are counted as a single location.

[18] Source: Destatis (2021).

[19] The absolute error for a NUTS-3 region is calculated as the absolute difference between the region's actual share of population and the share of Twitter users in the region. For example, if a NUTS-3 region has 0.01% of the actual population, but only 0.009% of Twitter users, the absolute error for this region is $|0.01 - 0.009| = 0.001$.

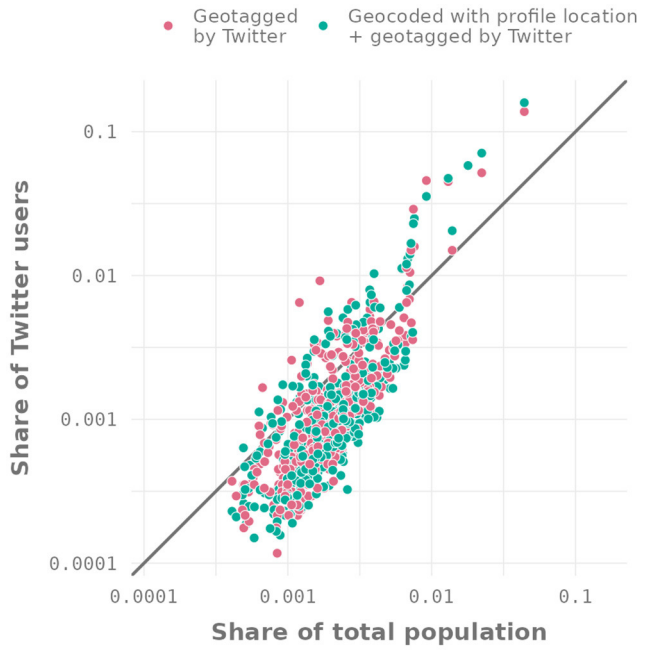


FIGURE 3 | Share of Twitter users geotagged by Twitter and geocoded with profile locations vs. share of the German population by NUTS-3 region.

common metrics using a bag-of-words approach. A bag-of-words for a document, or in our case for a collection of tweets, contains the count of each word ("token") after the data has been preprocessed and split into tokens.

We construct such a bag-of-words model, which we call "vocabulary," in the form of a table containing the number of occurrences of each word in our data, using all tweets (whether geolocated or not). We decompose the tweets into individual tokens according to the following scheme: First, we use a regular expression to filter out all URLs in the data. Then, we employ a tokenizer that lowercases all words and excludes all characters that are not in the *letter, lowercase* subcategory of the Unicode 6.0 standard[20]—except for the octothorpe (#), since its use as a "hashtag" on Twitter signifies a special meaning if prefixing a token. During vocabulary building, words that occur fewer than 25 times in the whole dataset are excluded as they are mostly misspelled, made-up words or more or less randomly occurring strings. What remains is a vocabulary containing 2.2 million unique tokens.

For the comparison of non-geolocated and geolocated tweets, we create two sub-vocabularies containing the word counts for tweets without geolocation and the word counts for tweets geolocated either by our method or by Twitter. In creating these vocabularies, we restrict ourselves to the token pool of the full vocabulary and again remove words that occur less than 25 times in the full dataset. Sub-vocabularies may, however, contain words that occur fewer than 25 times if the word has a low frequency in our data and is spread across the two sub-vocabularies.

We compute two common metrics to compare our sub-vocabularies of non-geolocated and geolocated tweets: the Jaccard S_J coefficient and the cosine similarity S_C. Since the Jaccard coefficient is the ratio between the size of the intersection of two sets and the size of their union, it measures the extent to which the sub-vocabularies contain the same words. It does not, however, take into account the distribution of words within the sets, that is, how many times a word occurs in each set. The cosine similarity is effectively calculated on the intersection of the two sets and is therefore agnostic to the set differences analyzed by the Jaccard coefficient, but can account for the word count differences within the intersection[21]. In our case, the Jaccard coefficient is $S_J(Vocabulary_{non-geo}, Vocabulary_{geo}) = 0.935$, while the cosine similarity is $S_C(Vocabulary_{non-geo}, Vocabulary_{geo}) = 0.996$. For both metrics, 1 represents the greatest possible similarity, and 0 the greatest possible dissimilarity. Although such summary statistics do not tell the whole story, they do show that the distribution of words in both data sets is extremely similar. The high Jaccard coefficient shows that both non-geolocated and geolocated tweets share more than 93% of words between them, with a large proportion of the words that are not shared across the vocabulary being odd words with rather low frequency (results not shown). The high cosine similarity supports this even more strongly. If the distribution of words among the common words were different in terms of their frequency, e.g., if some words were very prevalent in one corpus, but less common in the other (in relation to other words in the respective corpus), the cosine similarity would be low, which might ultimately indicate that some topics are less discussed or covered in one of the corpora. However, the very high cosine similarity is a strong indication that most words and (and possibly topics) are present to a similar extent in both non-geolocated and geolocated tweets.

6. APPLICATION EXAMPLES

In this section, we provide examples that demonstrate how regional variance observed in Twitter data can be used to approximate real-world behavior in the case of elections and regional party support, and how regional variance in dialects and gender-inclusive language can be captured in tweets. Furthermore, these simplified examples show that different types of analyses are possible at both the user and tweet level, and that digital behavior and communication correspond to known regional differences in the real world. In this respect, the forthcoming use cases display the potential of the geocoded data in sociological and political science analyses to reveal spatial variations in public discourse and behavior.

[20]The Unicode 6.0 standard includes 1,759 lowercase letters from multiple languages in its specification (www.unicode.org/versions/Unicode6.0.0/UnicodeStandard-6.0.pdf).

[21]The cosine similarity, interpreted for the case at hand, corresponds to the angle between the vocabularies, that is, the vectors of term frequencies. Hence, despite the difference in size in the absolute values between the partial vocabularies, no further normalization is necessary.

6.1. Voting Behavior and Party Support in Tweets

One advantage of our geocoding technique is that it significantly enhances the possibility for regionalized content analysis using Twitter data. Although analyses of regional differences in party support, political attitudes, and voting behavior have already been conducted with Twitter data (Beauchamp, 2017; Lopez et al., 2017), our data offer large gains in the number of cases available at the lower regional levels. Compared to survey data, analysis using Twitter data is comparatively inexpensive and can enable real-time tracking of regional public opinion (nowcasting)—a major challenge for survey projects (see Lopez et al., 2017).

To demonstrate the potential of this approach, we analyze hashtags in support for the German Green Party shortly before the September 2021 federal election and use party support on Twitter as a predictor of Green Party vote shares at the NUTS-2 level. For this purpose, we analyze data from the 30-day period (August 28, 2021, to September 26, 2021) leading up to the election on September 26, 2021, as this is the period when there is the most support and publicity for the party. First, we take data containing hashtags that indicated support for the Green Party[22] and collect the count of users who tweeted using one of these hashtags at least once across the 38 NUTS-2 regions that we previously geocoded using the method presented above.

We compare the regional distribution of this quantity with the distribution of Green party votes in the 2021 federal election[23]. As we would expect a greater number of Twitter users who support the Green party as well as pro-Green votes in more populous regions, we divide both of our counts—the number of users tweeting in support for the Greens and the number of Green votes—by the total population at the NUTS-2 level. By doing this, both quantities are normalized by the same regional constant and, therefore, more comparable.

The Pearson correlation coefficient for party support on Twitter and actual voting behavior at the NUTS-2 level shows a significant positive relationship between the two quantities [$r_{(35)} = 0.528$ at $p < 0.001$]. However, it is evident from **Figure 4** that this correlation is in part driven by the two major cities of Berlin and Hamburg, which are overrepresented on Twitter and at the same time have comparatively strong levels of support for the Green party in the election. These results suggest that Twitter data geolocated by our method can—to some extent—provide an approximation for a known regional quantity, namely the level of electoral support for the Green Party in a given region in this example.

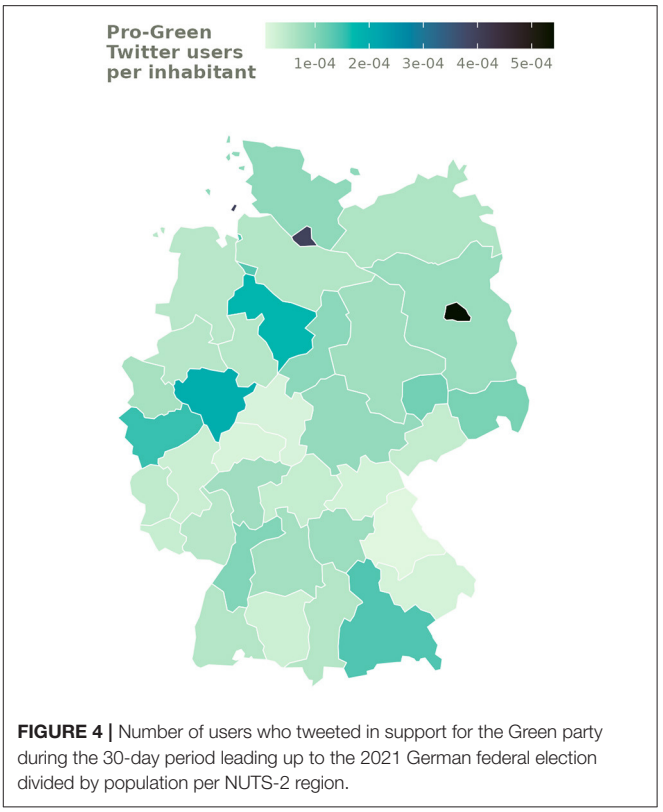


FIGURE 4 | Number of users who tweeted in support for the Green party during the 30-day period leading up to the 2021 German federal election divided by population per NUTS-2 region.

6.2. Regional Dialects

Like many other languages, German is characterized by different regional dialects. We perform a tweet-level analysis to capture linguistic differences in social media communication and investigate whether known regional dialects are represented in a similar pattern in digital communication.

An example of different dialects in Germany is the use of words for bread rolls, which are most commonly called *Brötchen*, but are usually called *Semmel* in southeastern Germany[24]. We test our data against this rather fuzzy concept of regional dialects, this time using data from the entire 3-year period covered by our dataset.

We search for tweets that mention bread rolls by performing a pattern match on a list of German names for bread rolls against our database (see **Supplementary Section 3.2** for the list of patterns used). In this analysis, we do not normalize by the number of users and simply count the number of tweets that match one of the corresponding words describing a bread roll, as we are interested in the most frequently used expression by region. For each NUTS-3 region, we calculate the total number of occurrences of the above two terms for bread rolls in tweets that can be attributed to that region based on Twitter geotags or our geocoding results. **Figure 5** shows the spatial distribution

[22]The hashtags we use are: #diesmalgrün (#thistimegreen), #grünwählen (#electgreen), #bereitweilihresseid, (#readybecauseyouare), #grün (#green), #grüne (#greens) and the respective version with mutated vowels replaced (The full query to the database can be found in the **Supplementary Section 3.1**). It should also be noted that hashtags like #green or #greens are often used in news reports and may not represent actual support for the party. However, because we excluded retweets and verified Twitter accounts, which are mostly a superset of professional accounts such as news outlets, we assume that these hashtags much more accurately represent party support.

[23]We use the second vote (*Zweitstimme*), which voters cast for a party at the national level, not for a regional candidate. Since party votes are only available at the district level (Der Bundeswahlleiter, 2021), we aggregate these election results at the NUTS-2 level. We also exclude Saarland, where it was not possible to cast a second vote for the Greens in the 2021 federal election (tagesschau.de, 2021).

[24]Other variations also exist across Germany, but they occur much less frequently compared to these two.

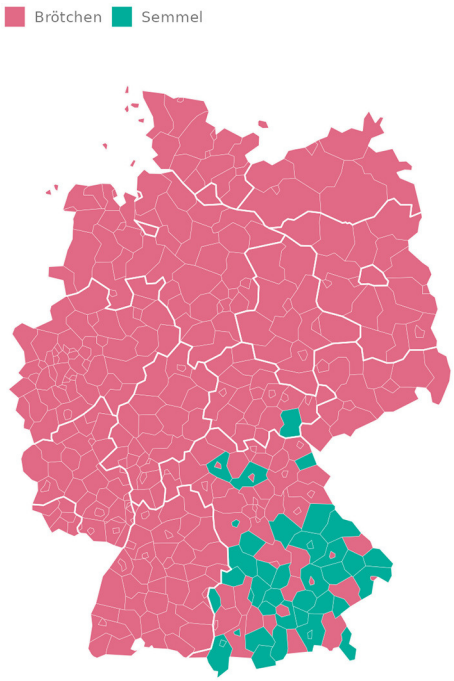


FIGURE 5 | Most common name for German bread rolls by NUTS-3 region.

of the words *Brötchen* and *Semmel* across NUTS-3 regions. For each region, the word most frequently used in tweets by users from that region is shown. In 361 regions, *Brötchen* is the most frequently used word for bread rolls, while in 40 regions, *Semmel* is most often used. As expected, all regions that favor *Semmel* are located in southeastern Germany. Yet, even in a large part of southeastern Germany, *Brötchen* is still predominant, being a very common word that is widely known throughout Germany.

This example shows that, first, our data is able to capture regional variation in dialects, a concept rather difficult to quantify, especially when dealing with a word that is a common description known throughout Germany. Second, and more interestingly, in our example, regional variation cannot be captured as precisely if we aggregate tweets at the NUTS-2 level. In the NUTS-2 aggregate, *Brötchen* is more common than *Semmel* in all but two regions. This is due to the fact that even in southeastern Germany, there are many NUTS-3 regions where *Brötchen* is either more common, or less common but not significantly so. When aggregating at the NUTS-2 level, the total number of occurrences of *Brötchen* outweighs *Semmel*, despite the presence of subregions where *Semmel* is used more frequently. This exemplifies a case where finer-grained spatial analysis—enabled by the data geocoded with our method—allows for the uncovering of regional patterns that would otherwise go undetected.

6.3. Regional Variation in the Use of Gender-Inclusive Language

The German language uses gendered nouns, distinguishing three genders: masculine, feminine, and neuter. While there is an ongoing effort to make German more gender-neutral, both spoken and written German still tend to be biased toward masculine forms. Efforts to include all genders extend to the development of more gender-inclusive language. For example, the common noun *Mitarbeiter* (employees), a masculine plural noun, can be written in a more gender-inclusive way as *MitarbeiterInnen*, *Mitarbeiter_innen*, *Mitarbeiter*innen*, or *Mitarbeiter:innen*[25]. We show that our data can also be used to capture regional differences in the usage of gender-inclusive language. Here, we again aggregate users in our data who have used gender-inclusive plural nouns in at least one original tweet[26], this time at the NUTS-3 level (401 regions). We divide this count by the number of unique users in each respective region to get an estimate of the share of users who use gender-inclusive language when tweeting.

Figure 6 shows the distribution of the share of users who use gender-inclusive language across the 401 NUTS-3 regions. It is apparent that major cities tend to have higher shares of users tweeting with gender-inclusive forms of plural nouns. A possible hypothesis could be that Twitter users from cities are more gender-aware than users from rural areas. To assess this hypothesis, we calculate the Pearson correlation between the share of users using gender-inclusive language and the population density of the respective region. The resulting correlation coefficient $r_{(399)} = 0.482$ at $p < 0.001$ suggests that living in a less populous area may indeed be linked to less frequent use of gender-inclusive language.

A possible explanation for this correlation could be a larger share of academics or a larger young female population in urban areas. Combining data from INKAR (*Indikatoren und Karten zur Raum- und Stadtentwicklung*, English: indicators and maps of spatial and urban development) (Bundesinstitut für Bau-, Stadt- und Raumforschung, 2022) with our regional aggregates of Twitter data, we compute three linear regression models (**Table 5**) where the response variable in each case is the proportion of gender-inclusive language users in a region. Explanatory variables include the logarithm of population density (since the distribution of the population density is right-skewed), the proportion of employees with an academic degree, and the proportion of women aged 20–40 in the total population.

Our results show a positive effect of population density on the share of gender-inclusive language users (Model 1). However, the inclusion of the share of employees with an academic degree (Model 2) leads to a positive and significant effect of this predictor as well as a substantial increase in explanatory power, while the effect of population density diminishes. Finally, when the proportion of women aged 20–40 is added as a covariate (Model 3), which also has a significant positive effect, the effect of population density becomes no longer significant. This suggests that the correlation between population density and gender-inclusive language is indeed an effect of the demographic structure of the NUTS-3 regions.

[25]This list of possible variants is exhaustive.

[26]The regex pattern to query usage of gender-inclusive language is reported in the **Supplementary Section 3.3**.

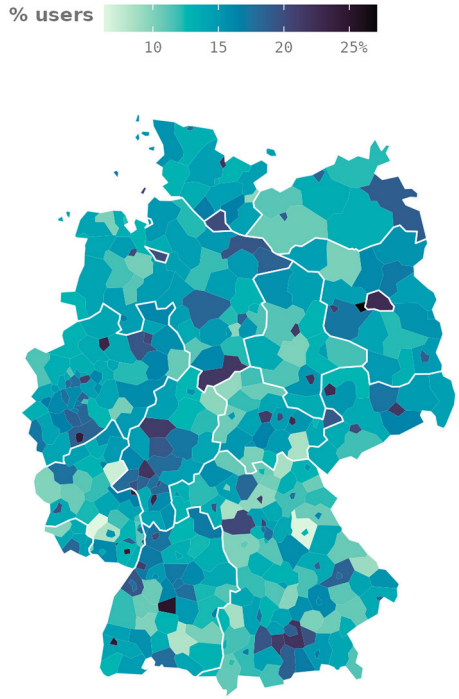


FIGURE 6 | Percentage of Twitter users who used gender-inclusive language at least once by NUTS-3 region.

TABLE 5 | Regression models of the proportion of gender-inclusive language users in NUTS-3 regions.

	Model 1	Model 2	Model 3	Model 4
(Intercept)	0.058***	0.082***	−0.110*	−0.106*
	(0.007)	(0.007)	(0.047)	(0.048)
Population density (log)	0.016***	0.003*	0.002	0.003
	(0.001)	(0.002)	(0.002)	(0.002)
Share academic employees		0.004***	0.004***	0.003***
		(0.000)	(0.000)	(0.000)
Share female population (20-40y)			0.004***	0.004***
			(0.001)	(0.001)
λ				0.185**
				(0.069)
R^2	0.273	0.470	0.492	0.503
Num. obs.	401	401	401	401
Log likelihood	856.052	919.463	927.781	930.947

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$.

Examining the residuals of the OLS models reveals the presence of spatial autocorrelation, with Moran's I significant at $p < 0.05$ in all three models. This suggests potential biases in the estimation of parameters in the presented linear models. To account for spatial dependence in the unobservables, we add a spatial autoregressive error term (Model 4)[27]. While the λ parameter is positive and significant, indicating spatial clustering among the unobserved characteristics, the coefficients of the spatial error model for the independent variables remain very similar to those of the OLS model, further supporting the results reported in the previous paragraph on the effects of the proportion of academics and young female population on the use of gender-inclusive language.

[27]Spatial error model: $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$, where $\mathbf{u} = \lambda\mathbf{W}\mathbf{u} + \varepsilon$ and $\mathbf{W}$ is the spatial weights matrix (Rüttenauer, 2022).

7. DISCUSSION

Digital behavioral data and big data are becoming an increasingly important resource for social science research. In this respect, Twitter is one of the most widely used data sources, not least because of the ease of access to the data for research purposes.

In this paper, we implemented a method for geocoding Twitter users and tweets using the user profile locations to substantially increase the amount of Twitter data usable for regional analyses. By using a self-hosted, customized database of the OpenStreetMap search engine Nominatim to geocode profile locations in our dataset of German tweets, we achieved an 150-fold increase in the number of tweets that can be geolocated in Germany, from 0.18 to 26.4%. With the new, larger sample, we were able to confirm the biases in the spatial distribution of Twitter users highlighted in previous research, with larger cities overrepresented, and smaller cities and rural areas underrepresented compared to the actual population. We developed and maintain a companion free open-source R package, `nutscoder` (github.com/long39ng/nutscoder), which facilitates straightforward reuse of our geocoding procedure and extends the applicability of our method to administrative regions outside Germany.

We evaluated our geocoding results based on a number of parameters. First, the assessment of the geocoding performance based on comparisons of geocoded profile locations and geotags provided by Twitter showed a high level of accuracy of our results. Second, the geolocated and non-geolocated tweets do not appear to differ systematically in terms of word occurrences. Consequently, tweets geolocated using our method could represent an almost random subsample of all tweets for many applications. However, further analysis is needed to assess the potential bias in the content of geolocated tweets compared to non-geolocated tweets.

Moreover, we have demonstrated through a number of use cases that our geolocated data are able to capture a) known regional differences (predicting party votes on the regional level), b) fuzzy regional differences (reproducing the spatial distribution of known regional dialects), and c) previously unknown regional differences, for example in the use of gender-inclusive language between urban and rural areas.

Many other applications of analyzing regionalized Twitter data are potentially possible, including monitoring regional changes in attitudes and behavior over time, deriving proxy information about regions that can be used as

explanatory variables. In particular, when research aims to compare small regions or small time periods, survey data are usually not suitable, and indicators derived from Twitter data may be able to fill certain data gaps. Thus, although Twitter does not allow for deriving population parameter estimates in almost all cases, it can be useful for a number of research applications and should be further studied and evaluated by social science methodology research.

By standardizing the geocoding results to official codes of administrative regions, our procedure makes it simple to combine the geocoded data with regional data from other sources, such as official statistics. This approach also has the additional benefit of being less privacy-sensitive compared to exact point coding. Of course, the geocoding output is not limited to administrative regions. By customizing the target geographic data on which we perform spatial joins of the geocoding results, we can modify the output to any desired set of regional identifiers.

Compared to approaches that model Twitter user networks and tweet content to infer users' real-world locations, our method of geocoding the profile location text should be able to provide more reliable results at much higher speed. Since we only geocode the information that explicitly relates to the users' locations, our geocoding results have a much lower degree of uncertainty and require much less effort to validate compared to the above alternatives. This makes our geocoding method particularly suitable for applications that work with very large amounts of data and/or in real time. Moreover, using our method to obtain more geographic information based on user profile locations provides more data that can be used for both training and evaluation of more sophisticated methods, thereby improving the efficacy of these methods. Given that many users do not provide profile locations— and many of those who do, do not provide actual locations—more sophisticated, specialized geolocation methods are the likely next step that will allow us to achieve better spatial coverage of Twitter data in future studies.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: Redistribution of the collected Twitter data is restricted by the Twitter Terms of Service, Privacy Policy, Developer Agreement, and Developer Policy. The IDs of the geocoded tweets and the geocoding results associated with those IDs are available at gitlab.ub.uni-bielefeld.de/geocoding-german-twitter/geocoded-tweets. The geocoding procedure can be reproduced with the code in the paper's GitLab repository (gitlab.ub.uni-bielefeld.de/geocoding-german-twitter/geocoding-german-twitter) and/or with the use of the companion R package nutscoder (github.com/long39ng/nutscoder). Requests to access these datasets should be directed to HLN, long.nguyen@uni-bielefeld.de.

ETHICS STATEMENT

Ethical review and approval was not required for the current study in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

SKü, HLN, and DT conducted the literature review. HLN designed and implemented the geocoding. HLN and DT performed the evaluation of the geocoding results. DT and AK performed the analyses in the application examples. SKn managed the infrastructure for data collection, management, and analysis. All authors discussed the results and contributed to the final version of the manuscript.

FUNDING

This research was partly funded by (a) the Leibniz ScienceCampus SOEP RegioHub (Bielefeld University and SOEP/DIW Berlin) and (b) the German Ministry for Family Affairs, Senior Citizens, Women and Youth (BMFSFJ) in the context of the National Discrimination and Racism Monitor (NaDiRa) as well as the Research Association Discrimination and Racism (FoDiRa) of the DeZIM Research Community (German Center for Integration and Migration Research).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fsoc.2022.910111/full#supplementary-material

REFERENCES

Ahmed, W., Vidal-Alaball, J., Downing, J., Seguí, F. L., et al. (2020). COVID-19 and the 5G conspiracy theory: social network analysis of Twitter data. *J. Med. Internet Res*. 22, e19458. doi: 10.2196/19458

Ajao, O., Hong, J., and Liu, W. (2015). A survey of location inference techniques on Twitter. *J. Inform. Sci*. 41, 855–864. doi: 10.1177/0165551515602847

Alex, B., Llewellyn, C., Grover, C., Oberlander, J., and Tobin, R. (2016). "Homing in on twitter users: evaluating an enhanced geoparser for user profile locations," in *LREC* (Protorovz), 3936–3944.

Amaya, A., Biemer, P. P., and Kinyon, D. (2020). Total error in a big data world: adapting the TSE framework to big data. *J. Survey Stat. Methodol*. 8, 89–119. doi: 10.1093/jssam/smz056

Arthur, R., and Williams, H. T. P. (2019). Scaling laws in geo-located Twitter data. *PLoS ONE* 14, e0218454. doi: 10.1371/journal.pone.0218454

Backstrom, L., Sun, E., and Marlow, C. (2010). "Find me if you can: Improving geographical prediction with social and spatial proximity," in *Proceedings of the 19th International Conference on World Wide Web* (Raleigh, NC), 61–70.

Bakerman, J., Pazdernik, K., Wilson, A., Fairchild, G., and Bahran, R. (2018). Twitter geolocation: a hybrid approach. *ACM Trans. Knowl. Discovery Data* 12 34, 1–34, 17. doi: 10.1145/3178112

Beauchamp, N. (2017). Predicting and interpolating state-level polls using twitter textual data. *Am. J. Pol. Sci*. 61, 490–503. doi: 10.1111/ajps.12274

Beisch, N., and Koch, W. (2021). 25 Jahre ARD/ZDF-onlinestudie: unterwegsnutzung steigt wieder und streaming/ mediatheken sind weiterhin treiber des medialen internets. *Media Perspektiven* 10, 486–503.

Blanford, J. I., Huang, Z., Savelyev, A., and MacEachren, A. M. (2015). Geo-located Tweets. enhancing mobility maps and capturing cross-border movement. *PLoS ONE* 10, e012902. doi: 10.1371/journal.pone.0129202

Blank, G. (2017). The digital divide among Twitter users and its implications for social research. *Soc. Sci. Comput. Rev*. 35, 679–697. doi: 10.1177/0894439316671698

Bundesinstitut für Bau-, Stadt- und Raumforschung (2022). INKAR - Indikatoren und Karten zur Raum- und Stadtentwicklung. BBSR Bonn.

Chandra, S., Khan, L., and Muhaya, F. B. (2011). "Estimating twitter user location using social interactions–a content based approach," in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing* (Boston, MA: IEEE), 838–843.

Cheng, Z., Caverlee, J., Barthwal, H., and Bachani, V. (2014). "Who is the barbecue king of texas?: a geo-spatial approach to finding local experts on Twitter," in *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Gold Coast, QLD; ACM), 335–344.

Cheng, Z., Caverlee, J., and Lee, K. (2010). "You are where you tweet: a content-based approach to geo-locating twitter users," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management-CIKM '10* (Toronto, ON: ACM Press), 759.

Choi, S. (2020). When digital trace data meet traditional communication theory: theoretical/methodological directions. *Soc. Sci. Comput. Rev*. 38, 91–107. doi: 10.1177/0894439318788618

Compton, R., Jurgens, D., and Allen, D. (2014). "Geotagging one hundred million Twitter accounts with total variation minimization," in 2014 *IEEE International Conference on Big Data (Big Data)* (Washington, DC: IEEE), 393–401.

Davis, Jr, C. A., Pappa, G. L., de Oliveira, D. R. R., and de Arcanjo, L. F. (2011). Inferring the location of twitter messages based on user relationships. *Trans. GIS* 15, 735–751. doi: 10.1111/j.1467-9671.2011.01297.x

Der Bundeswahlleiter (2021). Bundestagswahl 2021. Ergebnisse nach kreisfreien Städten und Landkreisen.

Destatis (2021). Kreisfreie Städte und Landkreise nach Fläche, Bevölkerung und Bevölkerungsdichte am 31.12.2020.

Dredze, M., Paul, M. J., Bergsma, S., and Tran, H. (2013). "Carmen: a twitter geolocation system with applications to public health," in *Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence* (Bellevue, WA: AAAI).

Elmongui, H. G., Morsy, H., and Mansour, R. (2015). "Inference models for Twitter user's home location prediction," in *2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)* (Marrakech: IEEE), 1–8,

European Commission (2016). Commission Regulation (EU) 2016/2066 of 21 November 2016 Amending the Annexes to Regulation (EC) No 1059/2003 of the European Parliament and of the Council on the Establishment of a Common Classification of Territorial Units for Statistics (NUTS). *Off. J. Euro. Union*. L 322, 1–61.

Gao, Y., Wang, S., Padmanabhan, A., Yin, J., and Cao, G. (2018). Mapping spatiotemporal patterns of events using social media: a case study of influenza trends. *Inte. J. Geograph. Inform. Sci*. 32, 425–449. doi: 10.1080/13658816.2017.1406943

Ghoorchian, K., and Girdzijauskas, S. (2018). "Spatio-temporal multiple geo-location identification on Twitte,". in *2018 IEEE International Conference on Big Data (Big Data)* (Seattle, WA: IEEE), 3412–3421.

Goldberg, D. W., Wilson, J. P., and Knoblock, C. A. (2007). From text to geographic coordinates: the current state of geocoding. *URISA J*. 19, 33–46.

Google Maps (2022). *Geocoding API*. Google Maps Platform.

Graham, M., Hale, S. A., and Gaffney, D. (2014). Where in the world are you? geolocation and language identification in Twitter. *Profess. Geographer*. 66, 568–578. doi: 10.1080/00330124.2014.907699

Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S., et al. (2010). Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci*. 368, 3875–3889. doi: 10.1098/rsta.2010.0149

Han, B., Cook, P., and Baldwin, T. (2013). "A stacking-based approach to twitter user geolocation prediction," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (Sofia), 7–12.

Han, B., Cook, P., and Baldwin, T. (2014). Text-based twitter user geolocation prediction. *J. Artif. Intell. Res*. 49, 451–500. doi: 10.1613/jair.4200

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., and Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartogr. Geogr. Inf. Sci*. 41, 260–271. doi: 10.1080/15230406.2014.890072

Hecht, B., Hong, L., Suh, B., and Chi, E. H. (2011). "Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vanvouver, BC), 237–246.

Hipp, J. R., Bates, C., Lichman, M., and Smyth, P. (2018). Using social media to measure temporal ambient population: does it help explain local crime rates? *Justice Q*. 36, 718–748. doi: 10.1080/07418825.2018.1445276

Hoang, T. B. N., and Mothe, J. (2018). Location extraction from tweets. *Inf. Process. Manag*. 54, 129–144. doi: 10.1016/j.ipm.2017.11.001

Hoffmann, S. (2021a). *Abbreviations*. Nominatim Blog.

Hoffmann, S. (2021b). *Nominatim 4.0.0 Released*. Nominatim Blog.

Huang, Y., Guo, D., Kasakoff, A., and Grieve, J. (2016). Understanding U.S. regional linguistic variation with Twitter data analysis. *Comput. Environ. Urban Syst*. 59, 244–255. doi: 10.1016/j.compenvurbsys.2015.12.003

Jungherr, A. (2018). *Normalizing Digital Trace Data*. New York, NY: Routledge.

Jurgens, D. (2013). "That's what friends are for: Inferring location in online social media platforms based on social relationships," in *Proceedings of the International AAAI Conference on Web and Social Media, Vol. 7* (Cambridge, MA: AAAI), 273–282.

Jurgens, D., Finethy, T., McCorriston, J., Xu, Y., and Ruths, D. (2015). "Geolocation prediction in twitter using social networks: a critical analysis and review of current practice," in *Proceedings of the International AAAI Conference on Web and Social Media, Vol. 9* (Oxford: AAAI), 188–197.

Karami, A., Kadari, R. R., Panati, L., Nooli, S. P., Bheemreddy, H., and Bozorgi, P. (2021). Analysis of geotagging behavior: do geotagged users represent the twitter population? *ISPRS Int. J. Geoinform*. 10, 373. doi: 10.3390/ijgi10060373

Karami, A., Lundy, M., Webb, F., and Dwivedi, Y. K. (2020). Twitter and research: a systematic literature review through text mining. *IEEE Access* 8, 67698–67717. doi: 10.1109/ACCESS.2020.2983656

Kong, L., Liu, Z., and Huang, Y. (2014). Spot: Locating social media users based on social network context. *Proc. VLDB Endowment* 7, 1681–1684. doi: 10.14778/2733004.2733060

Levy, B. L., Phillips, N. E., and Sampson, R. J. (2020). Triple disadvantage: neighborhood networks of everyday urban mobility and violence in U.S. cities. *Am. Sociol. Rev*. 85, 925–956. doi: 10.1177/0003122420972323

Li, R., Wang, S., Deng, H., Wang, R., and Chang, K. C.-C. (2012). "Towards social user profiling: unified and discriminative influence model for inferring home locations," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Beijing), 1023–1031.

Lopez, J. C. A. D., Collignon-Delmar, S., Benoit, K., and Matsuo, A. (2017). Predicting the brexit vote by tracking and classifying public opinion using Twitter data. *Stat. Politics Policy* 8, 85–104. doi: 10.1515/spp-2017-0006

Lwin, M. O., Lu, J., Sheldenkar, A., Schulz, P. J., Shin, W., Gupta, R., et al. (2020). Global sentiments surrounding the COVID-19 pandemic on Twitter: analysis of Twitter trends. *JMIR Public Health Surveillance* 6, e19447. doi: 10.2196/19447

Malik, M., Lamba, H., Nakos, C., and Pfeffer, J. (2015). "Population bias in geotagged Tweets," in *Proceedings of the International AAAI Conference on Web and Social Media, Vol. 9* (Oxford), 18–27.

Map Foundation (2021). *OpenStreetMap Foundation*. Available online at: https://wiki.osmfoundation.org

Martinez, L. S., Hughes, S., Walsh-Buhi, E. R., and Ming-Hsiang, T. (2018). "Okay, We get it. you vape": an analysis of geocoded content, context, and sentiment regarding e-cigarettes on Twitter. *J. Health Commun*. 23, 550–562. doi: 10.1080/10810730.2018.1493057

Matsuo, S., Shimoda, W., and Yanai, K. (2017). "Twitter photo geo-localization using both textual and visual features," in *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)* (Laguna Hills, CA: IEEE), 22–25.

McCormick, T. H., Lee, H., Cesare, N., Shojaie, A., and Spiro, E. S. (2015). Using twitter for demographic and social science research: tools for data collection and processing. *Sociol. Methods Res*. 46, 390–421. doi: 10.1177/0049124115605339

McGee, J., Caverlee, J., and Cheng, Z. (2013). "Location prediction in social media based on tie strength," in *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management* (San Francisco, CA: ACM), 459–468.

McGee, J., Caverlee, J. A., and Cheng, Z. (2011). "A geographic study of tie strength in social media," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management* (Glasgow), 2333–2336.

Menshikova, A., and van Tubergen, F. (2022). What drives anti-immigrant sentiments online? a novel approach using twitter. *Eur. Sociol. Rev*. doi: 10.1093/esr/jcac006

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., and Danforth, C. M. (2013). The geography of happiness: connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE* 8, e0064417. doi: 10.1371/journal.pone.0064417

Miura, Y., Taniguchi, M., Taniguchi, T., and Ohkuma, T. (2017). "Unifying text, metadata, and user network representations with a neural network for geolocation prediction," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vancouver, BC: Association for Computational Linguistics), 1260–1272.

Murthy, D., and Gross, A. J. (2017). Social media processes in disasters: Implications of emergent technology use. *Soc. Sci. Res*. 63, 356–370. doi: 10.1016/j.ssresearch.2016.09.015

Nguyen, Q. C., McCullough, M., Meng, H.-,w., Paul, D., Li, D., Kath, S., et al. (2017). Geotagged US tweets as predictors of county-level health outcomes, 2015– 2016. *Am. J. Public Health* 107, 1776–1782. doi: 10.2105/AJPH.2017.303993

Nominatim (2021). Add Wikipedia and Wikidata to Nominatim.

Nominatim (2022a). *Nominatim API*. Nominatim Documentation.

Nominatim (2022b). Open-source geocoding with OpenStreetMap data.

Nominatim (2022c). *Place Ranking in Nominatim*. Nominatim Documentation.

Nominatim (2022d). *Tokenizers*. Nominatim Documentation.

Ntompras, C., Drosatos, G., and Kaldoudi, E. (2022). A high-resolution temporal and geospatial content analysis of Twitter posts related to the COVID-19 pandemic. *J. Comput. Soc. Sci*. 5, 687–729. doi: 10.1007/s42001-021-00150-8

Onan, A. (2017). "A machine learning based approach to identify geo-location of Twitter users," in *Proceedings of the Second International Conference on Internet of Things, Data and Cloud Computing* (Cambridge United Kingdom: ACM), 1–6.

Ren, K., Zhang, S., and Lin, H. (2012). "Where are you settling down: geo-locating twitter users based on tweets and social networks," in *Information Retrieval Technology, Vol. 7675*, eds Y. Hou, J.-Y. Nie, L. Sun, B. Wang, P. Zhang, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D'Ṫerzopoulos, D. Tygar, M. Y> Vardi, and G. Weikum (Berlin; Heidelberg: Springer Berlin Heidelberg), 150–161.

Ribeiro, S., and Pappa, G. L. (2018). Strategies for combining Twitter users geo-location methods. *Geoinformatica* 22, 563–587. doi: 10.1007/s10707-017-0296-z

Rieder, Y., and Kühne, S. (2018). "Geospatial analysis of social media data - a practical framework and applications," in *Computational Social Science in the Age of Big Data. Concepts, Methodologies, Tools, and Applications. DGOF Schriftenreihe* (Cologne: Herbert van Halem Verlag), 423–446.

Roller, S., Speriosu, M., Rallapalli, S., Wing, B., and Baldridge, J. (2012). "Supervised text-based geolocation using language models on an adaptive grid," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (Jeju Island), 1500–1510.

Rout, D., Bontcheva, K., Preoţiuc-Pietro, D., and Cohn, T. (2013). "Where's@ wally? a classification approach to geolocating users based on their social ties," in *Proceedings of the 24th ACM Conference on Hypertext and Social Media* (New York, NY: ACM), 11–20.

Rüttenauer, T. (2022). Spatial regression models: a systematic comparison of different model specifications using monte carlo experiments. *Sociol. Methods Res*. 51, 728–759. doi: 10.1177/0049124119882467

Scheffler, T. (2014). "A German Twitter snapshot," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (Reykjavik: European Language Resources Association), 2284–2289.

Sen, I., Flöck, F., Weller, K., Weiß, B., and Wagner, C. (2021). A total error framework for digital traces of human behavior on online platforms. *Public Opin. Q*. 85, 399–422. doi: 10.1093/poq/nfab018

Shelton, T., Poorthuis, A., and Zook, M. (2015). *Social Media and the City: Rethinking Urban Socio-Spatial Inequality Using User-Generated Geographic Information*. SSRN Scholarly Paper 2571757, Social Science Research Network, Rochester, NY.

Sloan, L., and Morgan, J. (2015). Who Tweets with their location? understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. *PLoS ONE* 10, e0142209. doi: 10.1371/journal.pone.0142209

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., et al. (2013). Knowing the Tweeters: deriving sociologically relevant demographics from Twitter. *Sociol. Res. Online* 18, 74–84. doi: 10.5153/sro.3001

Stephens, M. (2020). A geospatial infodemic: mapping Twitter conspiracy theories of COVID-19. *Dialogues Hum. Geogr*. 10, 276–281. doi: 10.1177/2043820620935683

Stier, S., Breuer, J., Siegers, P., and Thorson, K. (2019). Integrating survey data and digital trace data: key issues in developing an emerging field. *Soc. Sci. Comput. Rev*. 38, 503–516. doi: 10.1177/0894439319843669

tagesschau.de (2021). Bundestagswahl: Saar-Landesliste der Grünen bleibt ausgeschlossen. tagesschau.de.

Tian, H., Zhang, M., Luo, X., Liu, F., and Qiao, Y. (2020). "Twitter user location inference based on representation learning and label propagation," in *Proceedings of The Web Conference 2020* (Taipei: ACM), 2648–2654.

Tromble, R., Storz, A., and Stockmann, D. (2017). *We don't know what we don't know: When and how the use of Twitter's public APIs biases scientific inference*. Available at SSRN 3079927.

Wang, Q., Phillips, N. E., Small, M. L., and Sampson, R. J. (2018). Urban mobility and neighborhood isolation in America's 50 largest cities. *Proc. Natl. Acad. Sci. U.S.A*. 115, 7735–7740. doi: 10.1073/pnas.1802537115

Wiedener, M. J., and Li, W. (2014). Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US. *Appl. Geograp* 54, 189–197. doi: 10.1016/j.apgeog.2014.07.017

Wing, B. P., and Baldridge, J. (2011). "Simple supervised document geolocation with geodesic grids," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11* (Portland, OR: Association for Computational Linguistics), 955–964.

Yildiz, D., Munson, J., Vitali, A., Tinati, R., and Holland, J. A. (2017). Using Twitter data for demographic research. *Demogr. Res*. 37, 1477–1514. doi: 10.4054/DemRes.2017.37.46

Zhang, Z., He, Q., and Zhu, S. (2017). Potentials of using social media to infer the longitudinal travel behavior: a sequential model-based clustering method. *Transport. Res. C Emerg. Technol*. 85, 396–414. doi: 10.1016/j.trc.2017.10.005

Zheng, X., Han, J., and Sun, A. (2018). A survey of location prediction on Twitter. *IEEE Trans. Knowl. Data Eng.* 30, 1652–1671. doi: 10.1109/TKDE.2018.2807840

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Sociology

TYPE Correction
PUBLISHED 02 August 2022
DOI 10.3389/fsoc.2022.995770

OPEN ACCESS

APPROVED BY
Frontiers Editorial Office,
Frontiers Media SA, Switzerland

*CORRESPONDENCE
H. Long Nguyen
long.nguyen@uni-bielefeld.de

SPECIALTY SECTION
This article was submitted to Sociological Theory, a section of the journal Frontiers in Sociology

RECEIVED 16 July 2022
ACCEPTED 18 July 2022
PUBLISHED 02 August 2022

CITATION
Nguyen HL, Tsolak D, Karmann A, Knauff S and Kühne S (2022) Corrigendum: Efficient and reliable geocoding of German Twitter data to enable spatial data linkage to official statistics and other data sources. *Front. Sociol.* 7:995770. doi: 10.3389/fsoc.2022.995770

Corrigendum: Efficient and reliable geocoding of German Twitter data to enable spatial data linkage to official statistics and other data sources

H. Long Nguyen*, Dorian Tsolak, Anna Karmann, Stefan Knauff and Simon Kühne

Faculty of Sociology, Bielefeld University, Bielefeld, Germany

KEYWORDS

Twitter, geocoding, spatial linkage, official statistics, regional analysis

A corrigendum on
Efficient and reliable geocoding of German Twitter data to enable spatial data linkage to official statistics and other data sources

by Nguyen, H. L., Tsolak, D., Karmann, A., Knauff, S., and Kühne, S. (2022). *Front. Sociol.* 7:910111. doi: 10.3389/fsoc.2022.910111

In the original article, there was an error in the Funding statement. The funder "Leibniz ScienceCampus SOEP RegioHub (Bielefeld University and SOEP/DIW Berlin)" was missing. The correct Funding statement appears below.

Funding

This research was partly funded by (a) the Leibniz ScienceCampus SOEP RegioHub (Bielefeld University and SOEP/DIW Berlin) and (b) the German Ministry for Family Affairs, Senior Citizens, Women and Youth (BMFSFJ) in the context of the National Discrimination and Racism Monitor (NaDiRa) as well as the Research Association Discrimination and Racism (FoDiRa) of the DeZIM Research Community (German Center for Integration and Migration Research).

The authors apologize for this error and state that this does not change the scientific conclusions of the article in any way. The original article has been updated.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Sociology

METHODS
published: 30 June 2022
doi: 10.3389/fsoc.2022.884640

Leveraging Dynamic Heterogeneous Networks to Study Transnational Issue Publics. The Case of the European COVID-19 Discourse on Twitter

Wolf J. Schünemann[1]*, *Alexander Brand*[1], *Tim König*[1] and *John Ziegler*[2]

[1] Institute of Social Sciences, Hildesheim University, Hildesheim, Germany, [2] Institute of Computer Science, Heidelberg University, Heidelberg, Germany

The ongoing COVID-19 pandemic constitutes a critical phase for the transnationalization of public spheres. Against this backdrop, we ask how transnational COVID-19 related online discourse has been throughout the EU over the first year of the pandemic. Which events triggered higher transnational coherence or national structuration of this specific issue public on Twitter? In order to study these questions, we rely on Twitter data obtained from the TBCOV database, i.e., a dataset for multilingual, geolocated COVID-19 related Twitter communication. We selected corpora for the 27 member states of the EU plus the United Kingdom. We defined three research periods representing different phases of the pandemic, namely April (1st wave), August (interim) and December 2020 (2nd wave) resulting in a set of 51,893,966 unique tweets for comparative analysis. In order to measure the level and temporal variation of transnational discursive linkages, we conducted a spatiotemporal network analysis of so-called Heterogeneous Information Networks (HINs). HINs allow for the integration of multiple, heterogeneous network entities (hashtags, retweets, @-mentions, URLs and named entities) to better represent the complex discursive structures reflected in social media communication. Therefrom, we obtained an aggregate measure of transnational linkages on a daily base by relating these linkages back to their geolocated authors. We find that the share of transnational discursive linkages increased over the course of the pandemic, indicating effects of adaptation and learning. However, stringent political measures of crisis management at the domestic level (such as lockdown decisions) caused stronger national structuration of COVID-19 related Twitter discourse.

Keywords: dynamic networks, heterogeneous information networks, COVID-19, European public sphere, discourse analysis, Twitter, transnationalization

OPEN ACCESS

Edited by:
Tobias Wolbring, University of Erlangen Nuremberg, Germany

Reviewed by:
Emanuel Deutschmann, University of Flensburg, Germany
Dimitri Prandner, Johannes Kepler University of Linz, Austria

***Correspondence:**
Wolf J. Schünemann
wolf.schuenemann@uni-hildesheim.de

Specialty section:
This article was submitted to Sociological Theory, a section of the journal Frontiers in Sociology

Received: *26 February 2022*
Accepted: *16 May 2022*
Published: *30 June 2022*

Citation:
Schünemann WJ, Brand A, König T and Ziegler J (2022) Leveraging Dynamic Heterogeneous Networks to Study Transnational Issue Publics. The Case of the European COVID-19 Discourse on Twitter. Front. Sociol. 7:884640. doi: 10.3389/fsoc.2022.884640

1. INTRODUCTION

Scholarly research across disciplines has shown great interest in the transnationalization of social communication through digital media. Against euphoric expectations of a "death of distance" (Cairncross, 2001) in the wake of the revolutionary period of Internet development and (again) the emergence of social networks, there is an ongoing dispute including many skeptical voices that

point to a panoply of structuring factors at the national and regional scale—be it linguistic or other cultural conditions, geographical proximity or institutional factors like media markets and political systems (Straubhaar, 1991, 2010; Taneja and Webster, 2016). However, the COVID-19 pandemic has tremendously altered the basic conditions of social life across the world. It likely constitutes a critical phase for the transnationalization of public spheres. While contact and travel restrictions have strongly affected physical mobility, especially palpable in otherwise "borderless Europe" (Opiłowska, 2021), digitalization has drastically reduced the gravity with which those changes could impact routines of social communication. The availability of digital means of communication and the increased use of digital technology during the pandemic have potentially compensated for many cuts into the social fabric, especially with respect to cross-border communication. Therefore, given that the pandemic has clearly boosted digital connectivity, one could also expect it to serve as a facilitator and driver for transnational communication.

When looking at the social effects of the pandemic from the perspective of transnationalization research, there is another fundamental alternative to be studied. While, on the one hand, news and discussion about the pandemic dominate political discourse worldwide and have thus exhibited if not produced so-called "overlapping communities of fate" (Held, 1997) arguably to a greater extent than ever before, there are, on the other hand, social and political reactions to the disease that have been interpreted as relapses into national egoisms. Central measures taken in response to the pandemic have shown a predominant national logic of crisis management, emphasizing the organizational needs of institutionally pre-disposed communities of place. Against this backdrop, we ask how transnational COVID-19 related online discourse has been. Do we observe new trends toward transnationally integrated social communication and discourse? Or do we see more nationally structured debates and communicative insulation driven by institutional nationalism in political crisis management? If overarching trends are inconsistent, which correlations between the course of the pandemic and crisis management on the one hand, and the transnationality of COVID-19 related online discourse on the other can be observed? How do national user communities differ in this regard?

While the theoretical discussion has been vivid over the last decades, methodology needs further development. New approaches especially need to live up to the great opportunities that Big Data and digital methods provide. With the methodological approach that we present in this paper, we make innovative use of a particularly rich resource of Twitter data (TBCOV multilingual COVID-19 Twitter dataset) and methods of digitally enhanced network analysis at scale. In particular, our approach allows for an integration of different kinds of discursive linkages (e.g., shared URLs, retweets, mentions, hashtags, named entities) into a Heterogeneous Information Network (HIN). HINs are defined as directed graphs which consist of multiple types of objects and connecting relations. These discursive linkages fall into two classes that we describe as topical and referential. To answer our research questions, we study HINs over time, allowing us to find valuable explanations for the variation observed in our data. Studying communication on social media platforms like Facebook or Twitter is a well-established research method in the social sciences (Edelmann et al., 2020; Özkula et al., 2022). This includes empirical research interested in questions of transnationality or transnationalization (Deutschmann, 2022). While being inspired by those pieces of related research, our approach turns toward a discourse-oriented methodology by concentrating on the variety of discursive linkages rather than social ties. In addition, our HIN-based approach allows for a flexible and temporally aware multi-dimensional modeling of discursive linkages, going beyond frequently applied network analytical methods.

As to the scope of our analysis, we deliberately focus our study on Europe for both pragmatic and theoretical reasons. For one, reducing the data allows us to conduct this resource-intensive research in reasonable time. Moreover, we can contribute to the ongoing debate about the European public sphere, which has been relevant for both the academic and political world for more than two decades. In the following section, we locate our study within the wider field of relevant research. In Section 3, we derive our hypotheses. Afterwards, we describe our data and methods in Sections 4, 5. Our results will be presented in Section 6 and discussed afterwards (7), before we summarize the most important findings in a short concluding section. Apart from its contribution to the ongoing debate on transnationalization within and outside of Europe, this paper advances the field of digital methods by introducing our novel, HIN-based methodology to the study of social networks and online discourse.

2. STATE OF RESEARCH

2.1. COVID-19 and Transnationality

For the longest time of human history, infectious diseases and transnationality have found themselves in a notoriously difficult relationship. It is common knowledge in medical science and public health that physical mobility is a key driver for the spread of infectious diseases like COVID-19—a crucial insight that has motivated contact restrictions, quarantine obligations and lockdowns as effective political measures. Social network research has traditionally contributed to explanations and predictions of the spread of infectious diseases (Klovdahl et al., 1994). This has more recently been transferred to the digital sphere based on digital communication data and computational methods (for an overview, see Aiello et al., 2020). In recent empirical research based on aggregated Facebook data, scholars have found strong correlations between social ties (i.e., social connectedness via Facebook friendship) and the regional spread of COVID-19 in the US as well as Italy (Bailey et al., 2020; Kuchler et al., 2020). Furthermore, Twitter and web news data have been used to predict COVID-19 outbreaks (Jahanbin and Rahmanian, 2020; Mellado et al., 2021). Given the local origin of a new virus and higher controllability of outbreaks at a local scale, transnational mobility is seen as responsible for the growing risks that infectious diseases constitute for an increasingly globalized world. While the expectation of a negative relationship

between infectious diseases and transnational mobility is thus mainly derived from medical science and public health studies, social science research has revealed additional facets of this relationship. Most fundamentally, researchers have put emphasis on the social construction of the risk of contagion (Bury, 1986; Conrad and Barker, 2010). Previous sociological work has shown how risk perceptions related to infectious diseases tend to be discursively coupled with social attitudes or convictions like colonial attitudes, xenophobic fears or racist convictions (Bhopal, 2014). This can lead to discrimination and stigmatization of outgroups with detrimental effects on transnational mobility in general and migration in particular (King, 2002; Bhambra, 2014). Empirical studies have found evidence for such discursive tendencies with respect to earlier diseases such as Aids/HIV, Ebola or Tuberculosis (Bancroft, 2001; Monson, 2017; von Unger et al., 2018, 2019), each accompanied by stereotypical fears toward (foreign) minorities detectable in different kinds of media discourses, including social media like Twitter and Facebook (Roy et al., 2020). Similar findings have been made recently with respect to COVID-19 and anti-Asian sentiment (Li and Nicholson Jr, 2021; Reny and Barreto, 2022). Judged from such perspectives alone, the expected effects of a global pandemic on transnationality can only be negative. Yet, the development of communication technologies in the age of digitalization has freed social communication (and connectedness) from its ontological relationship with physical mobility to a degree that it can now be regarded as an independent dimension of "transnational human activity" with different rules and expectations (Deutschmann, 2022). Therefore, the question about the effects of a global pandemic on transnational communication and discourse must be posed in a different way as it opens up new and relevant avenues for empirical research. To our knowledge, there has been no study taking up this demand by systematically studying the transnational quality of COVID-19 related online discourse so far.

2.2. The Ongoing Quest for the European Public Sphere

The emergence of a transnational public sphere has most prominently—and frequently—been studied with respect to the so-called European Public Sphere (Risse, 2010, 2015). The historical development and current state of Europe's political integration have driven normative and empirical expectations toward transnationalization. Empirical studies in the field have started out from different theoretical conceptions and adapted various methodologies (see Pfetsch and Heft, 2015). Besides more discourse-oriented studies (Koopmans and Zimmermann, 2010; Kantner, 2015), scholars have applied network analysis especially when studying communicative linkages in Internet communication and social networks (Koopmans and Zimmermann, 2010; Deutschmann et al., 2018; Ruiz-Soler, 2018; Schünemann, 2020; Stier et al., 2021; Wallaschek et al., 2022). While empirical scholars judged differently with respect to the fundamental question of whether there is such a thing as a European public sphere, there is some convergence around a common baseline observation. According to this insight, a European public sphere is not expected to appear "above and beyond the various national or issue-specific public spheres," but rather through the "Europeanization of national and other public spheres" (Risse, 2015, p.17). This is relevant also for our approach, as it lends additional justification to a less demanding operationalization of transnationality by measuring discursive linkages instead of actual social ties.

2.3. Transnational Communication and Digital Data

The Internet and social media are transnational by design. Coming from the perspective of the "networked public sphere," prominent scholars have predicted an extension of social communication across borders early on Benkler (2006) and Castells (2008). Internet technology and especially social media platforms would open up "electronic elsewheres" (Berry, 2010) as new places for social interaction (Papacharissi, 2015). Moreover, the structural transformations induced by digitalization would affect the concept of the public sphere as such, with a network of issue publics emerging instead of the single public constituted by traditional mass media (Bruns, 2008, p.69). For this process, Twitter plays a particularly important role as a central platform for the emergence of (*ad-hoc*) issue publics—at least in the Western Internet ecosystem (Bruns and Burgess, 2011). Methodologically, there is a broad range of measurements for transnationalization (Pfetsch and Heft, 2015). Traditionally, transnational communication flows have been assessed by network analysis (Koopmans and Zimmermann, 2010; Deutschmann, 2022) or discourse oriented studies (Koopmans and Statham, 2010; Kantner, 2015). More recently, scholars have turned toward digital trace data and computational methods (State et al., 2015; Taneja and Webster, 2016; Schünemann, 2020). However, most studies have looked at only one kind of linkage such as direct interactions, link-sharing or discourse in an isolated way and thus have not allowed for a combined perspective on different indicators of (trans-)national structuration.

2.4. Twitter Data and Empirical Research

Twitter is a unique data source for digital communication. Compared to other social media, data access for researchers is relatively easy and comprehensive (Özkula et al., 2022). However, there are important limitations that have to be kept in mind when using social media, and especially Twitter, data for social science research. These have been widely documented in the relevant research literature (Boyd and Crawford, 2012; Ruths and Pfeffer, 2014). The lack of representativeness in terms of the demographic characteristics of Twitter users has been discussed most broadly. Previous research has shown that social media users in general and Twitter users in particular tend to be younger, better educated and politically more liberal (Malik et al., 2015; Mellon and Prosser, 2017). While such bias is indeed likely to influence transnational communication and discourse, we would argue that this lack of representation affects our comparative study less than works that make inferences to the wider population. After all, it is precisely this subset of the population that is more likely to communicate transnationally across all countries.

TABLE 1 | KOF Globalization Index 2019 (KOFGI) social dimensions, Reuters social media usage 2020 (any purpose / general usage); and Reuters social media usage 2020 (news) by country.

Country	KOFGI (Score)	Reuters general (%)	Reuters news (%)	Country	KOFGI (Score)	Reuters general (%)	Reuters news (%)
Austria	525	10	5	Italy	477	18	9
Belgium	514	13	5	Latvia	493	–	–
Bulgaria	461	16	8	Lithuania	515	–	–
Croatia	500	12	4	Luxemburg	546	–	–
Cyprus	499	–	–	Malta	511	–	–
Czechia	496	8	4	Netherlands	516	16	7
Denmark	521	12	5	Poland	458	21	11
Estonia	499	–	–	Portugal	492	15	8
Finland	514	19	8	Romania	458	16	6
France	513	16	9	Slovakia	488	7	3
Germany	524	13	6	Slovenia	481	–	–
Greece	501	25	13	Spain	497	35	20
Hungary	473	13	4	Sweden	525	17	8
Ireland	523	24	14	United Kingdom	532	29	14

Of greater relevance to our study are Twitter's geographical, cultural and linguistic biases. While Twitter is used by a large community of users across the globe, there are cultural, regional and national differences that should be taken into account. Most obviously, activity on Twitter is very unequally distributed across the world, with dominant use in the United States followed by other OECD countries (Barnett and Park, 2014). With respect to our regional focus, for example, the Reuters Institute reported that in 2020, 29 percent of the British population were Twitter users, compared to 33% in Spain, 18 % in Italy, and only 13 % in Germany (Reuters Institute, 2020). These numbers also show the strong bias toward English-speaking, and especially Anglo-Saxon, countries. Furthermore, the user base in different countries uses the platform for different purposes. **Table 1** reports platform use across countries in our sample for both general purpose and news consumption. The latter are markedly lower, ranging at about half of the values for general usage. Moreover, different ratios (news in relation to any purpose) might be telling with respect to divergent usage patterns. For instance, the respective ratio is only at 0.31 for Hungary against 0.57 and 0.58 for Spain and Ireland, respectively. Cultural and national differences in how social media are used have been studied since their inception (Chu and Choi, 2010; Poblete et al., 2011; Sheldon et al., 2017; Hong and Na, 2018). International variation in social media usage patterns has been explained by cultural differences, e.g., between more individualist and more collectivist cultures (Chu and Choi, 2010; Shneor and Efrat, 2014; Sheldon et al., 2017). With respect to the method chosen for this paper, entity-based indicators for differences in usage are of particular interest here. So, for instance, frequent appearances of @-mentions and especially retweets have been interpreted as indications of a higher tendency to use Twitter for formal news dissemination. In contrast, a lesser degree of retweeting in a country sample would rather be read as showing a higher use of Twitter for conversational purposes (Poblete et al., 2011). Since country-specific general usage patterns likely affect the transnationality of COVID-19 related Twitter discourse, the respective statistics need to be taken into account (see Sections 6, 7 for results and discussion and **Supplementary Figure 3** for full statistics).

Returning to a macro-level of comparison, Twitter adoption itself is likely being influenced by the extent to which a national population is culturally globalized. The KOF Globalization Index (KOFGI) shall serve as a yardstick for assessing the extent to which a country is globalized in the following sections (Gygli et al., 2019). In order to provide an aggregate measure for the social dimensions of globalization, **Table 1** gives the respectively summed country scores of KOFGI for 2019. We can see at first glance that they do not significantly correlate with Twitter usage which is explainable by the fact that Twitter is only one social network among others and not the most central one for most populations.

Finally, there is a strong linguistic bias toward the English language in every global Twitter dataset. This, however, is not Twitter-specific, but rather reflects the special function of the English language for global communication—a kind of global language, especially online. As previous research has shown, English is the dominant language in cross-nationally linked online issue publics. For example, linguistic communities are more likely to be linked via English websites than direct ties, and content that is provided in other languages than English is unlikely to be recognized by international audiences at all (Hale, 2012). The effects of these phenomena on the results of our study on transnational COVID-19-related online discourse will be discussed in Section 7.

2.5. Heterogeneous Information Networks

So far, social science network research has not fully embraced the idea of heterogeneity. Networks in social science research traditionally grasp direct social ties between two or more actors

in a form of a sociogram or various kinds of actor-entity relations in an affiliation network. This actor-centered orientation of network analysis has particularly strong foundations in the tradition and theory of social action. Yet, other academic disciplines have increasingly shown that network analysis can be mobilized to study a broad variety of relational structures, including language and knowledge (Sowa, 2014). Previous research in the field of Computational Social Sciences also highlights the need to combine computational methods and social science theories when studying social media related questions (Fernandez et al., 2018). Nevertheless, prominent studies from the field of Heterogeneous Networks do not take theories from social sciences into account. This includes relevance measures based on meta paths that are used for searching similar nodes in HINs (Sun et al., 2011; Shi et al., 2012, 2014), but also methods to cluster or classify nodes into categories (Kong et al., 2012; Sun et al., 2013). While such approaches might be feasible in a context where only non-social relationships are considered, they might not be suitable for the study of social networks. Simply resorting to "information networks" as non-social does not resolve this shortcoming. Social media data, after all, clearly involves social interactions and might as well be modeled as HIN (Sun and Han, 2013). This is confirmed by recent work that leverages the Twitter network as HIN for recommendation and classification tasks (El-Kishky et al., 2022). With our approach, we build on theoretical conceptions from social sciences and communication studies. In contrast to other research in the field, we study discursive linkages instead of direct social ties. Unlike the related field of semantic network research developed in the fields of linguistics and Artificial Intelligence (Sowa, 2014), our approach does not concentrate on the level of concepts and linguistic structures, but integrates various relevant entities, including users and messages. Finally, whereas other network analytical research has remained static, we systematically include the temporal dimension. This latter feature is of crucial relevance given the procedural character of transnationalization and the dynamic character of the pandemic.

3. HYPOTHESES

With these considerations in mind, we expect transnational discursive linkages via Twitter to intensify with the severity of the pandemic. More precisely, when comparing the major phases of the pandemic throughout its first year, we expect shares of transnational linkages to go up during the so-called waves of the pandemic. Thus, for the macro-level perspective, we formulate our first hypothesis as follows:

> H1: We expect the share of transnational discursive linkages on Twitter to positively correlate with the severity of the pandemic.

Despite the inarguably transnational potential of social media communication, scholarly research has questioned the more substantial effects with respect to patterns of social connectedness and mass media publics that both still seem to be predominantly structured along national lines (Straubhaar, 2015; Bailey et al., 2020). While more cosmopolitically oriented elite actors—which are evidently overrepresented in most Twitter samples—practice transnational communication, the majority of users is still oriented toward the mainstream media and their national media logic. Moreover, not only are general social connectedness and media publics inherently structured along national lines, but the patterns observable in crisis management even across Europe have been critically discussed as exhibiting regrettable forms of neo-nationalism (Wang, 2021). Related to this discussion, it is important to keep in mind that especially political decisions on stringent measures have affected societies across the world at different times and to a different extent over the course of the pandemic. Such events are thus likely to produce peaks in society-specific communication, potentially inducing greater national structuration of public discourse and thus declines in transnational communication. Therefore, measured on the basis of daily events and its immediate effects, we expect the shares of transnational discursive linkages to decrease with the implementation of crisis management measures, such as lockdowns, in a certain country. Therefore, we formulate our second hypothesis as follows:

> H2: We expect transnational discursive linkages to negatively correlate with restrictive national measures.

Returning to the macro-perspective, overall effects of the pandemic on transnational discourse are not necessarily stable over time. Rather, we can expect processes of adaptation and learning over the course of the pandemic. Therefore, we can expect to observe variation between the two COVID-19 waves in our research period. Especially the first wave of the pandemic accompanied by the first national lockdowns might have produced some kind of shock-induced paralysis with people suddenly restricted to their homes, many social relations temporarily cut, and experiences of a particular state of exception. Therefore, we expect more national communicative activity during the first wave. In contrast, during the second wave, after having adapted to the Corona situation, including restrictive measures, and having established new digital ways to connect also in spheres where this has not been common beforehand, discursive linkages might have become more transnational over the development of the pandemic. Moreover, as our measurements are influenced by Twitter routines, one can assume that COVID-19 related communicative routines and codes have been established over time, facilitating discursive cohesion with the pandemic evolving. Taking these reflections into account, we formulate the following hypothesis:

> H3: We expect the share of transnational discursive linkages to positively correlate with the duration of the pandemic situation due to processes of learning and adaptation.

Finally, with respect to international variation, we expect countries that are less globalized with respect to socio-cultural indicators to show less alignment with global COVID-19 related discourse and thus have lower shares of transnational discursive linkages over all periods

studied. Therefore, we formulate our final hypothesis as follows:

H4: We expect the share of transnational discursive linkages to positively correlate with the extent to which a country is globalized in socio-cultural terms.

4. DATA

For our analysis, we chose the TBCOV Twitter dataset, a corpus of over 2 billion multilingual tweets posted between February 1st, 2020 and March 31st, 2021 (Imran et al., 2021). For now, TBCOV constitutes the most comprehensive dataset of worldwide Twitter communication on the pandemic. The TBCOV team collected tweets based on more than 800 multilingual query terms. There are crucial advantages of this dataset compared to similar resources (Dimitrov et al., 2020), namely that TBCOV is a multilingual dataset which is not restricted to English-language tweets. This makes it more balanced and less biased toward Anglo-Saxon communication flows. Moreover, tweets are geolocated with a multi-tier geolocation approach, using geotagged information, a lookup for user location entries and for elements with location information extracted from the body of the message via the Nominatim API. After processing, their dataset consisted of messages from 87 million unique users, across 218 countries, writing in 67 languages (Imran et al., 2021). We rehydrated the data for our subset of tweets located in one of the EU-27 countries or the United Kingdom. We further reduced the dataset by a selection of research periods representing the different phases of interest, namely April (1st wave), August (interim), and December (2nd wave). We understand the interim period as a relative reference period that helps us assess the effects of the pandemic waves on the transnationality of Twitter discourse. Our remaining dataset after rehydration through the Twitter APIv2 consisted of 51,893,966 tweets. Relevant discursive linkages, such as URLs, Hashtags and User Mentions, were extracted from the Twitter API, or, in the case of named entities, provided by the TBCOV dataset. Deviations in data from the full TBCOV dataset were mostly assignable to Twitter-initialized factors like bans.[1]

In order to determine the level of restrictions in the respective countries during our research period, we use the Covid Stringency Index. This index is a publicly available, day-wise composite measure based on indicators like school closings, work related restrictions and travel bans, scaled between 0 (no relevant restrictions) to 100 (strictest measures) offered by the Oxford COVID-19 Government Response Tracker project (OxCGRT) (Hale et al., 2021). It consists of a weighted summation of nine ordinally scaled indicators, whose numbers increase from recommendation to obligation of restrictions. The OxCGRT coded these indicators individually according to publicly available sources, e.g., news articles, press releases, and briefings.[2] A more detailed description of the indicator's development for EU-27 plus Great Britain is offered in the **Supplementary Figure 2**.

We obtained data on Twitter usage per country from the Digital News Report 2020 issued by the Reuters Institute for the Study of Journalism (Reuters Institute, 2020). Its yearly report is based on an international representative survey covering 21 of the 28 countries in our study (see **Table 1**). The dataset provides percentage values for Twitter usage for a) any purpose and b) news per country.

Finally, for assessing socio-cultural globalization per country, we rely on a combination of "social dimension" indicators taken from the KOF Globalization Index in its revised version for 2019 (Gygli et al., 2019). We obtained data for the six composed sub-indicators Interpersonal Globalization (de facto and de jure), Informational Globalization (de facto and de jure), Cultural Globalization (de facto and de jure), (each ranging from 0 to 100) and calculated a summarized score (see **Table 1**).

5. METHODS

5.1. Measuring Transnationality Through HIN-Based Methodology

Most network analytical studies on transnationality or transnationalization have analyzed actual activity. Thereby, researchers using digital communication data normally set a much higher threshold for relevant transnationality due to its exclusive orientation toward user interactions. We would argue that reading Twitter's default options for the creation of communicative linkages as ready-made relations upon which to construct a sociogram overestimates the real-world meaning of such platform-induced interactions. Furthermore, it disregards other, more subtle but still relevant, connective patterns in online discourse. Finally, the actual use of a platform's built-in connective features depends more strongly on pre-existent social network constellations and their reflections in platform membership or user hierarchies than the more balanced set of discursive linkages that we include in our HINs. In effect, this makes our measurement less prone to underestimate transnational alignments. Moreover, our approach is less bound to and biased by Twitter's affordance architecture.

By reorienting network analysis toward discourse research, our HIN-based approach puts emphasis on shared knowledge structures and discursive patterns instead of mediated user interactions. However, instead of representing webs of knowledge based only on one class of co-occurring linguistic or other signifiers (e.g., words), our approach allows to also include actors, documents (in our case tweets) and various kinds of automatically extractable entities as nodes in the overall network. While this fundamental feature makes our approach applicable to a wide range of research questions in the study of social communication beyond Twitter, transnationalization or the issue at hand, it is of course important to specify the linkages included

[1] While bans have multiple reasons, we found the most occurring non-dehydratable tweets based on bans of persons of interest like Donald Trump, Katie Hopkins or (Covid-related) disinformation accounts. With the ban of such accounts, tweets referencing them via e.g., mentions could not be rehydrated for our analysis. Additionally, retweets in our sample are restricted to API-delivered tweet contents due to computational restrictions.

[2] https://github.com/OxCGRT/covid-policy-tracker/blob/master/documentation/codebook.md; last accessed: 22-02-22.

in a HIN from case to case depending on the research questions. In the following sub-sections we present some basic reflections on the dimensionality of our linkages and make some conceptual clarifications before we outline our methodological approach in greater detail.

5.2. Discursive Linkage Dimensions

Similar to other network analyses, our HIN-based methodology bears the risk of flat ontologies with respect to the actual quality of the relations observed in the network. To determine what kind of connection a certain meta path constitutes in the real world is not always a trivial task. Relevant entities for our study certainly differ with respect to the kind of relationship they establish. Using the same named entity in a tweet as another user does somewhere at the other end of the world (or in the next village) constitutes a low-threshold linkage in comparison to sharing the same URL, not to mention actual user interactions. However, since such differences in frequency and likelihood cannot be easily translated into a differentiated metric, we have not weighted the entity-specific instances of our linkage types for our aggregate measure. In order to provide more differentiated information, we nevertheless present disaggregated measures and consider the individual types of linkages in our dataset. At a conceptual level, we distinguish between referential linkages via retweets or URLs and topical linkages via named entities, hashtags or @-mentions.

5.2.1. Referential Linkages

Referential linkages stand for a connection between two users due to the reference they make to the same content. While retweeting is the most frequently used in-built function for making reference to other content within the Twitter platform, sharing URLs constitute the standard practice of hypertext referencing within the much wider web-based media environment.

Retweets: In most Twitter-based communication studies using network analytical methods, retweets are conceived as direct links between the retweeting user and the user that has been retweeted (Ruiz-Soler, 2018; Stier et al., 2021). While such standard operationalization of social ties in Twitter research seems straight forward, it is important to keep in mind that Twitter ties, including retweets, are relatively weak (Takhteyev et al., 2012). Moreover, retweeting is heavily conditioned by both Twitter's affordance architecture and the pre-existent social relations partly represented in follower networks. Conceiving retweets as sharing of third party content as we do in this study has the advantage to not overestimate social ties that retweet interactions otherwise might suggest. In addition, retweets as entities can be better aligned to the overall taxonomy of linkages. We detected retweets based on the metadata available in the TBCOV dataset.

URLs: Hyperlinks represented by URLs as automatically extractable entities point to the relational core feature of web technology that platforms such as Twitter also rely on (Benkler, 2006). The analysis of hyperlinking patterns is standard practice in the study of online communication and has played a major role in earlier periods of internet development (Adamic and Glance, 2005; Hale, 2012). It is still a relevant approach and transferable to social media platform communication (Jacobson et al., 2016; Schünemann, 2020). Co-sharing of URLs establishes a connection between two users when referring to the same content in the wider universe of the web. Therefore, it allows to integrate referential linkages that are not Twitter-specific or dependent on the platform. Accordingly, hyperlinking has been taken as a proxy to measure awareness of media content across national or linguistic borders in previous research (Barnett et al., 2011; Taneja and Webster, 2016). We obtained the expanded URLs when rehydrating tweets via the Twitter APIv2.

5.2.2. Topical Linkages

Topical linkages group a second dimension of discursive connections. They indicate that two users in a pair of tweets deal with the same issue or topic. Entities assigned to this dimension are named entities, hashtags and @-mentions.

Named entities: At the conceptual level, named entities are the symbolic representations of various kinds of real-world objects or entities such as persons, locations or organizations, that can be automatically extracted from tweet text. Co-usage of named entities is a low-threshold discursive linkage establishing a connection between two users that in a pair of tweets speak about roughly the same things. Named Entity Recognition (NER) is a standard procedure in information extraction. We obtained named entities as metadata directly from the TBCOV dataset. NER is particularly error-prone and must be scrutinized accordingly. However, given the sheer amount of named entities extracted, we are optimistic that error rates are negligible with respect to our overall indicator. Nevertheless, it is important to keep the low threshold for matches in mind when interpreting absolute numbers of linkages of this type. As this holds true for this type of linkages regardless of the national/transnational quality, we still trust our measurement based on the relative weight of transnational linkages.

Hashtags: The use of hashtags is a very prominent built-in function of the Twitter platform by which users themselves can ascribe their tweet message to a broader topically oriented debate. While NER-based linkages can be regarded as the least deliberate discursive events that we include in our taxonomy, hashtags represent the opposite of this spectrum, given that users consciously relate their messages to ongoing debates. Thus, hashtag usage is completely interwoven with the platform's affordance architecture and the sociotechnical environment it constitutes. However, given their increased visibility in general public spheres and its platform-induced value for strategic communication, hashtags are of great relevance for online discourse analysis. Studying hashtag occurrence and co-occurrence has become a standard approach in related research and hashtags themselves are taken as markers of online issue publics (Steinskog et al., 2017; Eriksson Krutrök and Lindgren, 2018; Haunschild et al., 2019). In our taxonomy, co-usage of hashtags constitutes a platform-specific discursive linkage between two users using the same hashtag in a pair of tweets. Hashtags were obtained when rehydrating tweets via the Twitter APIv2. As certain hashtags had served as query terms for TBCOV's initial data collection, we have disregarded all linkages produced via these hashtags.

@-mentions: Almost everything that has been said about retweets

could also be repeated for @-mentions. @-mentioning is an in-built functionality of the platform, it is thus frequently used in Twitter communication and highly conditioned by Twitter's affordance architecture. As such, it is arguably less dependent on (though certainly influenced by) pre-existent follower networks than retweets, as users do not need to come across third party content but can simply type the user handle or the name of another user and wait for suggestions made by the algorithm to select the right handle. On the other hand, in a sociogram based on Twitter data, @-mentions would constitute an even weaker tie than retweets as they can have various meanings. In computational social science, however, @-mentions have been used as an indicator for direct communicative linkages (Stier et al., 2021). Against this backdrop, it might seem counter-intuitive that we subsume @-mention based linkages under topical instead of referential linkages. We argue, however, that as we do not include any direct communicative linkages between two users in our basic heuristics, co-mentioning of a third party can serve as an indicator for talking about the same things (i.e., persons or events). In this respect, @-mentions are arguably closer to named entities (type 'person') than to retweets. @-mentions were obtained when rehydrating tweets via the Twitter APIv2.

5.3. Model Descriptions

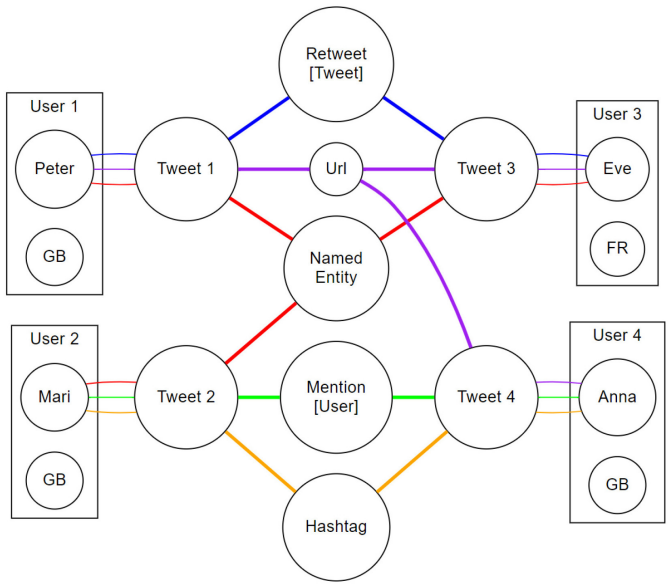
We apply Heterogeneous Information Networks (HINs) to COVID-19-related communication on Twitter. Following Sun et al. (2011) we understand a HIN as a directed graph which consists of multiple node and/or edge types, representing multiple types of objects or multiple types of relations between objects. In formal notation, we can describe a HIN as a graph $G = (V, E)$, consisting of nodes/vertices V and edges/links E, with an additional node type mapping of $\phi : V \rightarrow A$ (node types) and a link mapping of $\psi : E \rightarrow R$ (edge types). Further, meta paths P which describe paths on the graph have the form of $A_1 \overset{R_1}{\rightarrow} A_2 \overset{R_2}{\rightarrow} \ldots \overset{R_l}{\rightarrow} A_{l+1}$ (Sun and Han, 2013) following the given network scheme $T_G = (A, R)$. Using a hashtag, retweeting another message or sharing a URL are all regarded as discursive events and users are related via these discursive events they co-produce. Thereby, we establish discursive linkages between users. For instance, a user whose message has been retweeted by another user does not constitute a node in our network as such. Instead, we take the retweet information as central connector of a multi-hop linkage type whose instances connect a user to all other users retweeting the same message. This allows us to align all entity-specific linkage patterns to one taxonomy of similarly constructed multi-hop linkage types (meta paths) in our HIN (see **Figure 1**).

Among all possible meta paths P in our network, we consider only those that correspond to referential or topical linkages outlined above and refer to them as P'. Those meta paths follow the structure visualized in **Figure 1** and can be differentiated by their connecting central node which is part of the set $\{Retweet, Hashtag, Mention, Named\ Entitiy, Url\}$. Accordingly, we denote the different sets of meta paths as $P_{Retweet}, P_{Url}, P_{NamedEntity}, P_{Hashtag}, P_{Mention} \subseteq P'$ and name the starting and end nodes as v_s and v_e. Given C as set of countries present in the dataset and V_c as set of country nodes, we can further state that $v_s, v_e \in V_c$ for all of our analyzed meta paths. We get the actual country of a node via a mapping $\sigma : V_c \rightarrow C$. Therefore, to determine whether a given meta path p represents a national interaction (as opposed to a transnational), we define $\pi(p) = \delta(\sigma(v_s), \sigma(v_e))$ with δ being the Kronecker-Delta and $p \in P'$ as such indicator. Now, to calculate the transnationality score τ for a set of meta paths P' that start and end at country nodes, we leverage π as follows: $\tau(P') = 1 - \frac{\sum_{p \in P'} \pi(p)}{|P'|}$. Intuitively, this represents the fraction of meta paths that represent interactions between users of different countries. From the given definition one can conclude that $0 \leq \tau \leq 1$ holds true in all cases. Furthermore, for the temporal dimension of the discursive interaction described by a meta path, we resort to the publishing dates of the tweets in such a path. Meta paths are therefore taken into account within all time windows in which those dates fall. To test our hypothesis about a causal relationship between the stringency of crisis management measures and transnationalization, we rely on multilevel regression modeling. Here, we estimate six different models (one for each meta path type and a composite) with the stringency index as our central independent variable. Additionally, we include a month-wise term. This allows to control for general seasonal effects which may play a role regarding the transnational communicative patterns (e.g., summer vacations, Christmas). Additionally, we specify a country-wise random effect variable to control for unobserved heterogeneity and different approaches to restrictions. Finally, we control for meta path-specific effects in the composite model number six. To explore the effect of the level of national restrictions on τ, we estimated linear mixed effects models for each meta path of the structure: $\tau_{ip} \sim N(\mu_p, \sigma_p^2)$, which defines the assumption of a normal distributed dependent variable which can be modeled by estimating $\mu = \alpha_{c[ip]} + \beta_{1p}(\text{stringency index}) + \beta_{2p}(\text{month})$ for the linear combination of terms and additionally normal distributed separate intercepts $\alpha_{cp} \sim N\left(\mu_{\alpha_{cp}}, \sigma^2_{\alpha_{cp}}\right)$ for each country c and each meta path p. For the full model we build on a similar structure appending our general formula to explain $\tau_i \sim N(\mu, \sigma^2)$ with a meta path-specific term to $\mu = \alpha_{c[i]} + \beta_1(\text{stringency index}) + \beta_2(\text{month}) + \beta_3(\text{metapath})$, dropping the separation according to the meta path type in the random effects parameter $\alpha_c \sim N\left(\mu_{\alpha_c}, \sigma^2_{\alpha_c}\right)$ for each country c. Finally, in order to assess the impact of cultural globalization on transnational linkages in COVID-19 related Twitter discourse, we calculate Spearman's Rank correlations between the social dimensions of KOFGI and the aggregated indicator value for each country, summarized for the whole time period. We rely on a non-parametric correlation method due to non-normal distributions in the used variables.

6. RESULTS

6.1. Comparative Results: Time Periods

First, we consider the aggregate indicator for transnational discursive linkages (all meta paths) per time period (1st wave, interim, 2nd wave). In total, we found 5,216,112,060,389



$p_{Retweet}$
Country → User → Tweet → Retweet → Tweet → User → Country

p_{Url}
Country → User → Tweet → Url → Tweet → User → Country

$p_{Named\ Entity}$
Country → User → Tweet → Named Entity → Tweet → User → Country

$p_{Mention}$
Country → User → Tweet → Mention → Tweet → User → Country

$p_{Hashtag}$
Country → User → Tweet → Hashtag → Tweet → User → Country

FIGURE 1 | Structure of considered connections and corresponding meta paths for each type. National meta paths are constituted via connections between users from the same country, e.g., for $p_{Hashtag}$ from Mari to Anna (both located in Great Britain) and for $p_{Named\ Entity}$ from Peter to Mari. Transnational paths are e.g., p_{Url} from Peter to Eve (from Great Britain to France).

instances of meta paths that constitute the basis for the calculation of indicator values (1st wave: 2,213,876,499,502; interim: 1,053,010,119,563; 2nd wave: 1,949,225,441,324). A meta path-wise summary is depicted in **Supplementary Table 1**. **Figure 2** presents the share of transnational linkages per day for each of the phases in a line plot. Values are relatively fluid, mostly ranging from about 32–45% of all linkages (95% of all observations). A total minimum beyond this range is reached on a single day in August (interim period) with only about 27% total share of transnational linkages, a maximum at about 49% on a single day during the second wave. While there are no clear trends observable within one of the phases, there is a general upward movement with aggregate measures (mean and median) ranging higher for the interim than for the 1st wave and a more or less stable level of aggregate measures between interim and 2nd wave. Both the upward development toward the interim period and stagnation toward the second wave question hypothesis 1. The upward trend over the course of the year, however, lends support to hypothesis 3, as it corresponds to the expectation of adaptation and learning of communicative routines and the establishment of common discursive patterns during the pandemic which had not been effective in the first wave. The global share of transnational linkages is aggregated over all meta paths. The general impression can be differentiated by disaggregating the global indicator and by looking at the respective timelines for each meta path-specific subindicator. This is portrayed in **Figure 3**. The value ranges differ significantly between the subindicators. Hashtags produce the highest shares of transnational linkages, which was expected given the essential role of hashtags for the transnationally integrated affordance architecture Twitter provides as a global platform. The other topical linkages seem closer aligned to nationally structured discourses as transnational linkage shares are generally lower. As to the referential dimension, we see similar levels of values with wider ranges for retweets. Finally, transnational linkages realized via URLs are significantly lower, which indicates stronger dependence on nationally structured media logics and public spheres. Moreover, comparing the ranges of values over the different phases of the pandemic, an upward movement of transnational linkage shares can be identified as the clearest developmental pattern. It can be observed for all entities directly related to Twitter (hashtags, @-mentions and retweets) as well as named entities. For both retweets and named entities, however, the development appears to be mitigated with more or less stagnation of aggregate measures between the interim period and the 2nd wave. This might be explained by their closer alignment to either nationally structured discourse (named entities) or pre-existent Twitter follower networks, making changes in retweeting practice arguably more inert than in other kinds of linkages. Disaggregation also helps to understand that a steady upward development for the overall indicator seems to be impeded by named entities, which due to the masses of linkages produced via this entity have a huge weight in the overall indicator. The upward trend, observed as the clearest pattern in disaggregated results, lends additional support to our hypothesis 3. In this vein, adaptation and learning can indeed be observed over the course of the pandemic, with new communicative routines and discursive patterns successively established when dealing with the global state of exception—especially on a platform like Twitter. This overall pattern is somewhat contrasted by the timeline of

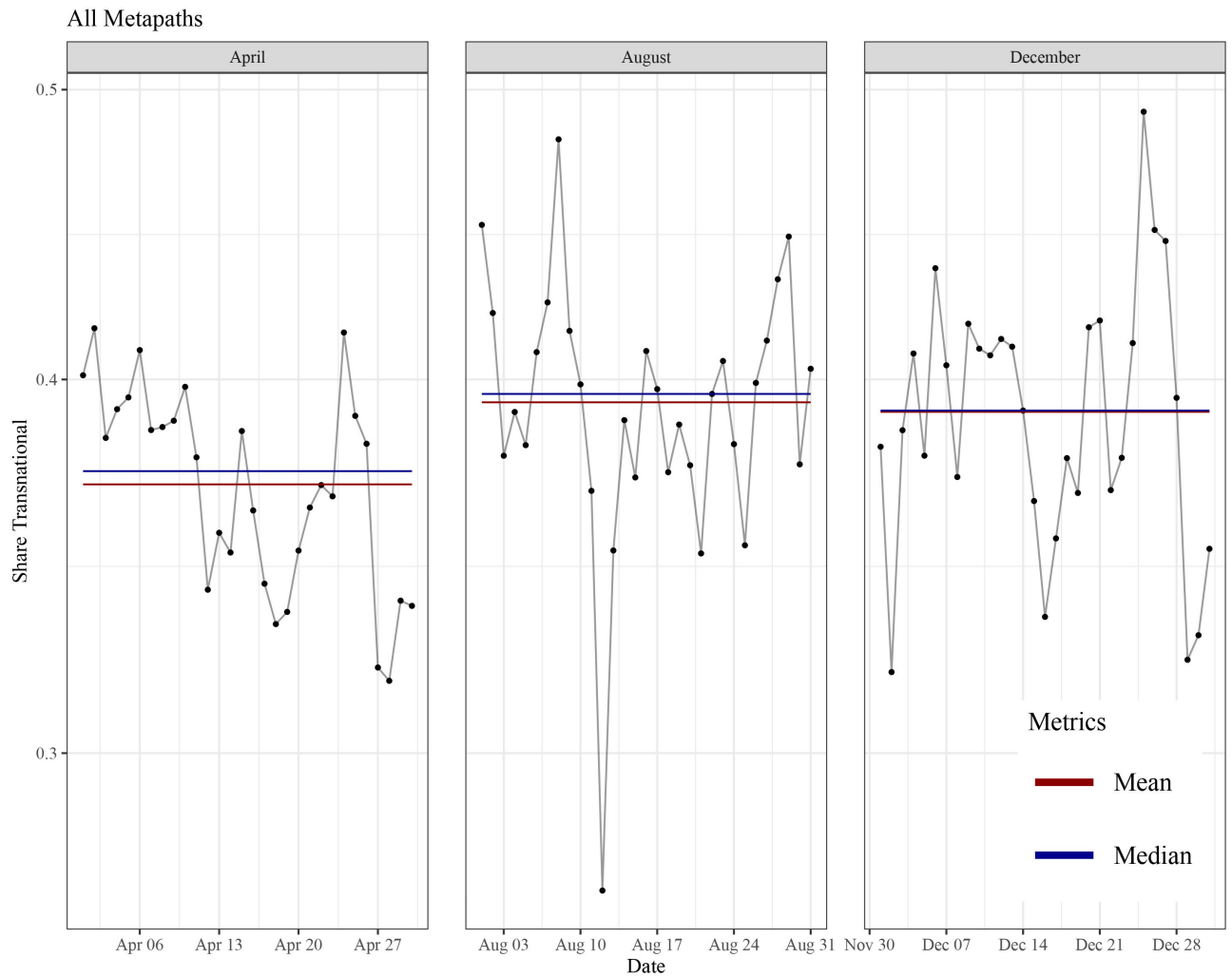


FIGURE 2 | Share of transnational linkages aggregated over all meta paths. The subfigures show the development of the share on a daily base (black line), monthly mean (red line) and monthly median (blue line). for the 3 months in the sample. The shares are presented on a logarithmic scale (y- axis) to better account for different magnitudes.

URL-based linkages. While it shows an upward development between the two waves, values are significantly lower during the interim period, producing a U-curve with the valley during the mid-term period. Interestingly, the subfigure would meet our intuitive expectation formulated in hypothesis 1, as it reveals that with the severity of the pandemic public attention rises for international sources and web content shared via social media, whereas this transnational issue attention is likely to be reduced during phases of relative calm.

6.2. Stringent Measures and Transnationality-Multilevel Model (Mixed Effects Models)

For our assessment of the effects of crisis management measures taken by political authorities, we exerted a multilevel analysis that models the relationship of stringent measures at the national level with the transnationality of Twitter discourse. **Table 2** shows the regression estimates with respect to the stringency of crisis management measures (compare **Supplementary Figures 1**, **2**) as independent variable and the meta path-specific indicators for transnationality as well as the aggregate indicator (all meta paths) as dependent variables. The fitted linear mixed effects model to explain the influence of stringency (10 point steps) on the share of transnational communication via hashtag paths (model 1) shows a statistically significant and negative effect (β = –1.11e-03, p = 0.006) of tightening restrictions. Furthermore, we see significant effects on a month-wise level with a negative effect (β = –3.12e-03, p = 0.041) for the interim period and for the 2nd wave (β = –1.33e-03, p = 0.141). For mentions (model 2), we observe a statistically significant and negative effect (β = –0.01, $p < 0.001$) of stringency and similar observations for hashtags with negative effects for the interim (β = –0.05, $p < 0.001$) and the 2nd wave (β = –0.03, $p < 0.001$). Regarding named entities (model 3), we

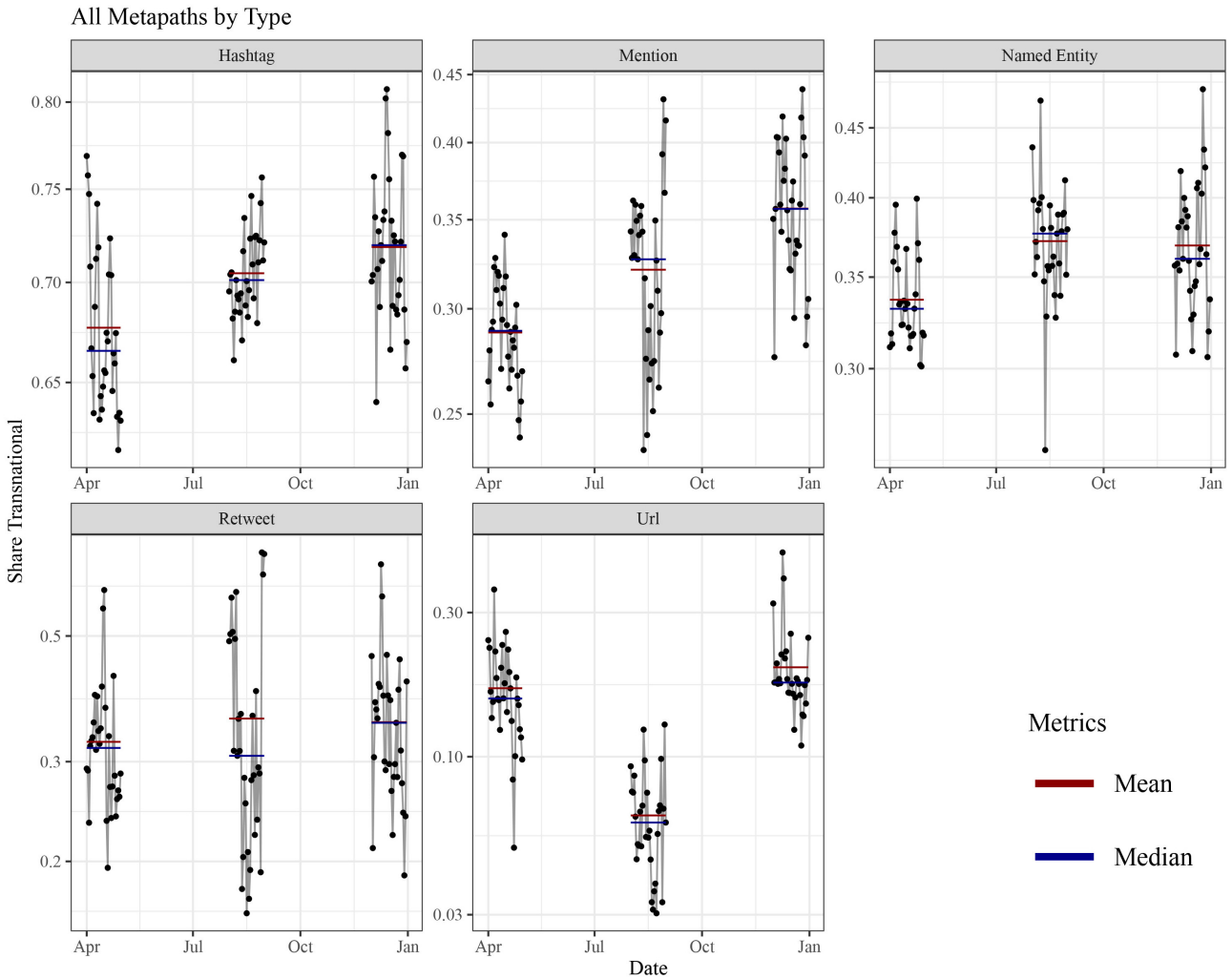


FIGURE 3 | Share of transnational linkages for all meta paths. The subfigures show the development of the share on a daily base (black line), monthly mean (red line) and monthly median (blue line), differentiated by type of path. The shares are presented on a logarithmic scale (y- axis) to better account for different magnitudes.

observe the same direction for stringency (β = –1.62e-03, p < 0.001) and for the interim (β = –9.96e-03, p < 0.001), but no significant difference for the 2nd wave compared to the level of transnational linkages during the 1st wave (β = –1.08e-03, p = 0.085) ceteris paribus. For retweets (model 4), we continue to see statistically significant and negative effects for stringency (β = –0.01, p < 0.001), interim (β = –0.07, p < 0.001) and 2nd wave (β = –0.05, p < 0.001), which also holds for URLs (model 5) with stringency effects (β = –0.01, p < 0.001), interim (β = –0.08, p < 0.001) and 2nd wave (β = –0.05, p < 0.001).

Fitting a model for all meta paths (model 6), we found a statistically significant and negative effect for stringency (β = –7.21e-03, p < 0.001), negative effects for the interim (β = –0.04, p < 0.001) and the 2nd wave (β = –0.03, p < 0.001). Controlling for meta path type, we observe significant negative effects of mention (β = –0.07, p < 0.001), retweet (β = –0.11, p < 0.001) and URL (β = –0.27, p < 0.001) in comparison to hashtag-based discursive linkages. The full model's total explanatory power is substantial (conditional R^2 = 0.80) and the part related to the fixed effects alone (marginal R^2) is of 0.20. All in all, the mixed effects model clearly lends support to hypothesis 2, as stringent measures of crisis management taken at the domestic level seem to have structuring effects toward national discourse communities at the expense of the transnationality of COVID-19-related discourse on Twitter. The strength of the observed transnationality seems to also depend on the type of meta path with negative effects compared to retweet-based paths consistent with **Figure 3**.

6.3. Country Comparison

Finally, we take a look at the country level by disaggregating our indicator per meta path and country (see **Figure 4**). We can observe a great variation in transnationality scores per country which seems to be related to the size of the population as well as the regional location. Over all meta path-specific indicators, smaller countries, especially the ones in Central and

TABLE 2 | Linear mixed effects models for each meta path and for a composition of all types.

	Hashtag		Mention		Named entity	
Coefficient	**Estimate**	**Conf. Int (95%)**	**Estimate**	**Conf. Int (95%)**	**Estimate**	**Conf. Int (95%)**
Stringency Score (+10 p)	–0.001***	–0.002 to –0.000	–0.011***	–0.012 to –0.009	–0.002***	–0.002 to –0.001
Period (Ref. 1st wave)						
Period (Interim)	–0.003**	–0.006 to –0.000	–0.046***	–0.053 to –0.039	–0.010***	–0.012 to –0.008
Period (2nd wave)	–0.001	–0.003 to 0.000	–0.027***	–0.031 to –0.023	–0.001*	–0.002 to 0.000
Intercept	0.946***	0.906 to 0.987	0.957***	0.873 to 1.040	0.954***	0.896 to 1.012
Random Effects						
σ^2	0.00		0.00		0.00	
τ_{00}	$0.01^{country}$		$0.05^{country}$		$0.02^{country}$	
ICC	0.98		0.98		1.00	
N	$28^{country}$		$28^{country}$		$28^{country}$	
Observations	2,576		2,576		2,576	
Marg. R^2 / Cond. R^2	0.000 / 0.982		0.003 / 0.979		0.000 / 0.996	
BIC	-14076.202		–9947.884		–15897.409	
	Retweet		**Url**		**All Meta paths**	
Coefficient	**Estimate**	**Conf. Int (95%)**	**Estimate**	**Conf. Int (95%)**	**Estimate**	**Conf. Int (95%)**
Stringency Score (+10 p)	–0.012***	–0.014 to –0.009	–0.011***	–0.015 to –0.007	–0.007***	–0.010 to –0.005
Period (Ref. 1st wave)						
Period (Interim)	–0.067***	–0.075 to –0.059	–0.079***	–0.093 to –0.065	–0.041***	–0.050 to –0.032
Period (2nd wave)	–0.054***	–0.059 to –0.049	–0.047***	–0.055 to –0.039	–0.026***	–0.031 to –0.021
Meta path (Ref. Hashtag)						
Meta path (Mention)					–0.074***	–0.079 to –0.069
Meta path (Named Entity)					0.002	–0.003 to 0.008
Meta path (Retweet)					–0.113***	–0.118 to –0.107
Meta path (Url)					–0.267***	–0.273 to –0.262
Intercept	0.941***	0.861 to 1.021	0.784***	0.691 to 0.876	1.007***	0.941 to 1.073
Random Effects						
σ^2	0.00		0.00		0.01	
τ_{00}	$0.04^{country}$		$0.06^{country}$		$0.03^{country}$	
ICC	0.97		0.92		0.75	
N	$28^{country}$		$28^{country}$		$28^{country}$	
Observations	2576		2576		12,880	
Marg. R^2 / Cond. R^2	0.009 / 0.965		0.007 / 0.920		0.204 / 0.797	
BIC	–8948.098		–6157.597		–22621.614	

*$p < 0.1$, **$p < 0.05$, and ***$p < 0.01$.

Eastern Europe, show higher shares of transnationality than larger countries. In contrast, variation is higher among the larger countries. Nevertheless, for all meta path-specific indicators, we find the lowest values for the United Kingdom, followed by Spain, France, Germany and Italy. This difference is not just produced by the amount of communication. Within our sample, we observe only a very small effect of the amount of communication (β = –4.069e-12, $p < 0.001$) if we include the number of paths in our final model (compared to **Table 2**). The reduction of the conditional R^2 of just 0.002 indicates a substantial effect of country beyond its amount of twitter communication. With respect to usage patterns, countries whose Twitter populations make comparatively frequent use of @-mentions and retweets such as Spain, France and the United Kingdom (for a full overview see **Supplemental Figure 3**), which can be regarded as an indication of a higher tendency to use Twitter for formal news dissemination (Poblete et al., 2011), also tend toward more nationally structured COVID-19 related Twitter discourses. However, this relation does not seem consistent (see, for instance, Germany, Italy or the Netherlands). With

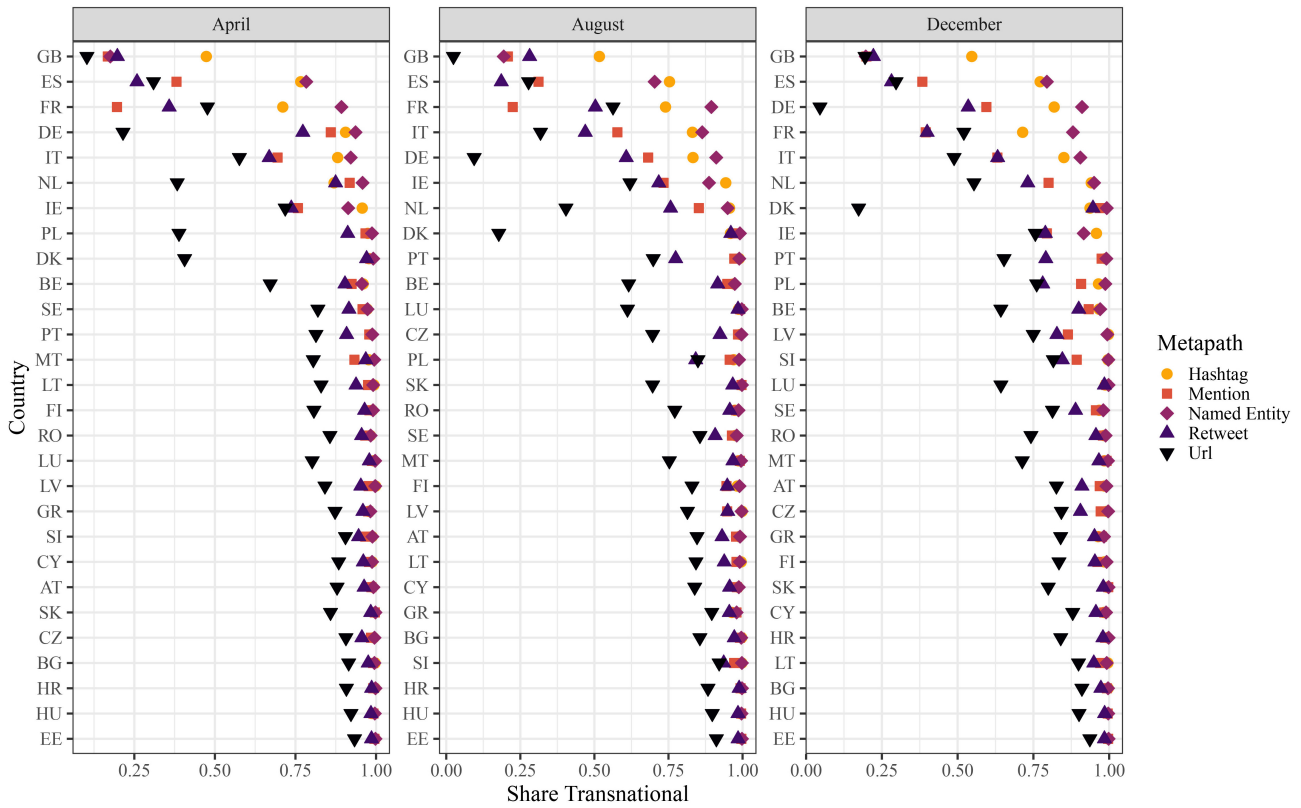


FIGURE 4 | Country-wise shares faceted by month. The point shapes and colors shows the respective type of meta path. The countries are ordered according to their average share of transnational communication for each month individually.

respect to the subindicators, the national comparison reveals some additional insights. As already discussed with respect to the aggregate figures, we see that URLs produce the lowest shares of transnational discursive linkages for almost all countries. Notable exceptions here are France and—to a lesser extent—Spain, where indicators for retweets and @-mentions score lower than for URLs. Furthermore, @mentions and retweets score remarkably lower in general than hashtags and named entities. Finally, what has been assumed for named entities—that they more closely represent a national discourse and are thus responsible for less transnational patterns—seems to hold true only for the United Kingdom and less so for Spain, whereas transnationality scores for named entities are among the highest when looking at the other countries.

As to cultural globalization, we expected a positive correlation between the extent to which a country is globalized in socio-cultural terms with transnational linkage shares measured by our HIN-based methodology. **Table 3** gives a Spearman's Rank Correlation of mean and median shares with the KOF Globalization Index. As the numbers show, effects are negative and mostly non-significant except for URLs. While this clearly puts H4 into question, it is important to note that the variables for Twitter usage (general purpose and news) from the Reuters Digital News Report show significant negative effects for almost all metapath-specific measurements. **Table 3** shows the respective values. The findings suggest that Twitter usage indeed has an impact on our measurement, yet in the sense that the more Twitter is used by a population, the more nationalized its Twitter discourse appears to be. In contrast, smaller Twitter populations in a given country tend to be closer aligned to a transnational discourse throughout all the phases covered. This general observation is plausible given the probable greater bias toward elite actors for smaller Twitter populations, whereas larger Twitter communities represent larger parts of the population and thus more fully represent a nationally mediated public discourse.

7. DISCUSSION

As presented in the results section, our study indicates that transnational discursive linkages have increased over the course of the pandemic, at least in Europe over the first months after its appearance from spring to summer 2020. The general experience of a global community of fate, which COVID-19 might have evoked, is thus partly supported by slight trends of discursive alignment across national borders via digital media. However, this overall statement needs to be differentiated in a number of relevant ways.

TABLE 3 | Country-wise rank correlation (Spearman's Rank Correlation) of mean and median shares with KOF Globalization Index 2019 (KOFGI) social dimensions, Reuters social media usage 2020 (any purpose/general usage); and Reuters social media usage 2020 (news) by country.

Indicator	Metapath	Spearman mean	Spearman median
KOFGI (Score)	Hashtag	–0.393	–0.394
	Mention	–0.336	–0.336
	Named entity	–0.354	–0.349
	Retweet	–0.263	–0.274
	Url	-0.354	–0.410*
Reuters general (%)	Hashtag	–0.562***	–0.572***
	Mention	–0.628***	–0.634***
	Named entity	–0.695***	–0.683***
	Retweet	–0.556***	–0.536**
	Url	–0.241	–0.242
Reuters news (%)	Hashtag	–0.634***	–0.650***
	Mention	–0.699***	–0.726***
	Named entity	–0.764***	–0.750***
	Retweet	–0.698***	–0.680***
	Url	–0.352	–0.348

*$p < 0.1$, **$p < 0.05$, and ***$p < 0.01$.

First, our research could not establish a general relationship between the transnationality of Twitter discourse and the severity of the pandemic, which had been our initial hypothesis (H1). What we have observed instead is an upward movement, starting at a comparatively low level for the first wave and then growing toward the interim period and stagnating (or even slightly declining) in the second wave. There is only one entity-specific subindicator—URLs—, for which we found the expected pattern of a U-curve with more transnational discursive linkages during the subsequent waves of the pandemic compared to less transnationally shared content during the interim period. This is telling, as among our multidimensional entity classes, URLs are the ones that arguably open a window to wider discursive universes and thus, in contrast to platform-specific entities, allow to go beyond the Twitter environment while not being as generic an indicator of discursive linkage as named entities. In previous research, sharing URLs has been perceived as an indication for more formal news dissemination in Twitter communication (Poblete et al., 2011). Thus, we can generally assume URLs to be bound more closely to a nationally structured public sphere and discourse, reflecting the prominent role of legacy media organizations. From this perspective, it makes sense that the transnational sharing of web sources is more frequent when the pandemic as a global event is salient than in periods of relative calm, when the global disease is only one topic among others. Another explanation for the unexpectedly high shares of transnational linkages during the interim period might be the fact that seasonal effects of the pandemic affect world regions differently so that, for instance, Twitter users in European countries might discuss dynamic developments on the Southern hemisphere, while going through a time of relative calm themselves.

Through our temporally sensitive design we have been able to observe correlations with political activities of crisis management. All stringent measures of crisis management were adopted and implemented at domestic level, reflecting an institutional nationalism apparently still dominant when it comes to public health or civil protection. With respect to the transnationality of COVID-19 related discourse, stringent measures had the expected effect of reinforcing national structuration of discourse and thus cause decreases in transnational linkage shares. Therefore, our multi-level analysis lends clear support to hypothesis 2.

Another important point to reflect upon when interpreting the results would be that April 2020, while serving as the natural starting period of our chronological observation, likely constitutes the most exceptional period covered in our research. With the first news on European incidents and casualties, and unprecedented political measures like general lockdowns imposed on populations across Europe and the world, people found themselves in a state of exception. It has likely taken time for routines of social exchange and communication to be reactivated—with the particular help of digital communication media. Thus, what we might observe in our analyses is how people learned to better cope with the novel situation caused by the pandemic and crisis management measures, including how to uphold connectivity and discourse across borders. Such ideas of adaptation and learning coincide with our third hypothesis. All in all, our empirical results lend support to H3: The share of transnational discursive linkages has increased over the course of the pandemic. Our combined indicators suggest that Twitter user populations across Europe have found more coherent ways to discursively deal with the pandemic and the socio-political effects it produced.

As to international variation and hypothesis 4, we do not find support for our basic expectation that transnational discursive linkages depend on the extent to which a national society is globalized. On the contrary, for URLs, we see a significant negative correlation (Spearman's Rank) between the KOFGI and our transnationalization measure. This might be explained with the fact that larger countries are likely to be more self-sufficient in professional news dissemination than smaller countries. This might be of particular importance during a pandemic with the general dependence on high quality information on public health. The other, non-significant KOFGI correlations indicate a similar direction. Moreover, Twitter usage seems to be an intervening factor, as transnationality of COVID-19 related Twitter discourse negatively correlates with Twitter usage both for any purpose and for news. This makes sense with respect to the differently skewed representation of national samples (Mellon and Prosser, 2017). Therefore, one should expect the general public discourse to be better represented in the COVID-19 related discourse of larger Twitter populations such as—most particularly—the United Kingdom, but also France or Spain. In contrast, smaller samples mostly include elite-level communication,

showing a closer alignment to global, transnational communicative patterns.

A number of limitations of our research and constraints in interpretability need to be considered. First of all, while the levels of transnational shares that we observe for our entity-specific indicators differ considerably, making judgements about whether they have generally reached a high or low level is not trivial. While for the most transnational entity type, hashtags, they range between impressive shares of 60 to 80 percent, they are remarkably lower for other entities. Especially for URLs, the share of transnational linkages is strikingly low. On some days during the interim period, transnational shares for URLs are close to zero. Moreover, when disaggregating by country, we see that ranges for some countries are even wider and minima are lower for large countries, especially for Great Britain. Thus, from a macro perspective, while we measure transnational discursive linkages and their tendency to increase, there are still strong indicators for national discourse structuration in the COVID-19 related Twitter discourse that can be explained by important macro-level factors such as language, culture and proximity.

All our findings need to be taken with a grain of salt given that our dataset—as many others in the field of Twitter research—suffers from a number of biases related to the self-selection effects of Twitter usage as presented in subsection 2.4. Of highest relevance to our comparative study is the lack of representation with respect to societies and languages. Even though we deliberately selected a multilingual dataset, the vast majority of tweets covered are in English. Data collection via hashtags and other pre-selected terms have likely favored English-speaking countries as well. Finally, also during data processing and entity extraction, it is very likely that algorithms such as geolocation or named entity recognition work better for English, and thus produce more and better results here than for all the other idioms.

A further limitation of our research design that affects possible explanations is the lack of a plausible baseline for our temporal comparison. This is a frequent issue for both research on transnationalization and Twitter research as such, given that appropriate longitudinal datasets are often not available. This is evidently true for our research as well, as there simply was no substantial COVID-19 related Twitter discourse before the beginning of 2020. Given the great variety of *ad-hoc* issue publics on Twitter, it would also not have made sense to compare indicators against some sort of random sample for which similar processing (including geolocation, entity extraction and annotation) would have not only been prohibitively time consuming, but unclear in its comparability. Instead, we chose a different path, conceiving the interim period between both waves of the pandemic as a relative reference period. While we find this decision still very plausible, it of course affects our interpretation with respect to hypothesis 3 on adaptation and learning. Does the upward movement of transnational linkage shares from the first wave to the interim indicate a COVID-19 effect on transnational discourse that can be understood without considering the exceptionality of this first wave? Or is this development just reflecting a form of normalization after the initial stage of paralysis? This question must remain open for future research as we do not have the proper benchmark or longitudinal data on which we could base the interpretation of our findings.

What is true for the comparison with respect to the past can also hold for the future, as we do not clearly see where all this leaves us in the long run. What can be taken for granted is that the pandemic and travel restrictions generally served as a driver for the expansion of digital communication, including cross-border communication. Whereas, in previous ages diseases reduced social communication due to their dependency on physical mobility, this fundamental connection seems somewhat resolved by digitalization. The pandemic and our adaptation of new rules for social life have given clear proof that digital platforms can provide substitutes for most forms of social communication and discourse. However, this should not make us neglect the potential effects that the drastic reduction of physical mobility, especially across borders, might cause in the long run. Given the fact that international travel has served as important driver for transnational connectivity over decades, it is likely that a substantial reduction can have the opposite longitudinal effect.

8. CONCLUSION

In this paper, we presented a novel HIN-based methodology for studying transnational discursive linkages in issue publics on Twitter. The COVID-19 pandemic served as a background context that motivated our issue-oriented interest. Thereby we contribute to the current research on the social effects of this extraordinary global crisis. We applied our method to a subset of TBCOV, a uniquely rich multilingual dataset of geolocated tweets. Focussing our regional scope on Europe helped us to avoid unrealistic expectations and relates our research to the ongoing quest for a European public sphere and the empirical research devoted to this question. Our findings suggest that the coherence of COVID-19 related Twitter discourse has not been a function of the severity of the pandemic, which would have supported the metaphoric understanding of the pandemic as building a community of fate, but that it interacts in more complicated ways with structuring factors that tend to conserve the pre-existent communities of place. What we observe is that transnational discursive coherence grows over the first months of the pandemic. However, this upward movement was cut, with our indicator remaining at a stable level between summer 2020 and the second wave in December. While adaptation to the pandemic context seems to increase transnational discursive linkages, a steady growth is arguably hampered by structural conditions. One factor that we studied more closely were the stringent measures of crisis management taken at domestic level. These had nationalizing effects, reducing the shares of transnational linkages significantly around such regulatory events. Moreover, we have found

interesting variations with respect to the linkage types included in our measurement as well as for the heterogeneous set of European countries. These insights into the complex geography of Twitter are also valuable for future researchers. While we discussed the limitations of our research in depth in Section 7, conclusions to be drawn from our study are particularly limited by its regional scope. Future studies should widen the scope beyond Europe or other regions, as only a global perspective would allow to reveal the structuring effects that regions itself (like Europe) have on the patterns of transnational exchange and discourse.

DATA AVAILABILITY STATEMENT

The datasets used in this paper are available via the original providers. TBCOV: https://github.com/CrisisComputing/TBCOV; COVID-19 policy response tracker: https://github.com/OxCGRT/covid-policy-tracker; KOF Globalization Index: downloadable file, URL: https://kof.ethz.ch/prognosen-indikatoren/indikatoren/kof-globalisierungsindex.html;
Reuters Digital News Report: full datasets upon request, for further information see: https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021/resources. The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

WS, AB, JZ, and TK contributed to conception and design of the study. WS was mainly responsible for theoretical foundations, including conceptions and metapath specifications, and wrote the first draft of the manuscript. AB and TK organized the database. AB and JZ performed the network analysis and wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This research has been funded by the Klaus Tschira Foundation in the framework of the EPINetz project. We acknowledge financial support by Stiftung Universität Hildesheim.

ACKNOWLEDGMENTS

We thank the Reuters Institute for the Study of Journalism at the University of Oxford for granting us access to the entire datasets prepared for the Digital News Reports 2020 and 2021.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fsoc.2022.884640/full#supplementary-material

Supplementary Figure 1 | Stringency of measures and transnational share of meta paths on a monthly comparison. Colors describe mean (red points and lines) and median (blue points and lines). Linetype and shape are used to differ between stringency index (dotted) and transnational share (solid).

Supplementary Figure 2 | Country-wise development of stringency index. The lines shows the index development per country, facetted by month.

Supplementary Figure 3 | Mean number of entities per tweet by country for all indicator (top subfigure) and facetted by entity type (bottom subfigure).

Supplementary Table 1 | Sum of meta paths per Type and Month.

REFERENCES

Adamic, L. A., and Glance, N. (2005). "The political blogosphere and the 2004 us election: divided they blog," in *Proceedings of the 3rd International Workshop on Link Discovery* (Chicago, IL).

Aiello, A. E., Renson, A., and Zivich, P. N. (2020). Social media- and internet-based disease surveillance for public health. *Annu. Rev. Public Health* 41, 101–118. doi: 10.1146/annurev-publhealth-040119-094402

Bailey, M., Johnston, D., Kuchler, T., Russel, D., State, B., and Stroebel, J. (2020). "The determinants of social connectedness in europe," in *Social Informatics, volume 12467 of Lecture Notes in Computer Science*, eds S. Aref, K. Bontcheva, M. Braghieri, F. Dignum, F. Giannotti, F. Grisolia, and D. Pedreschi (Cham: Springer International Publishing), 1–14.

Bancroft, A. (2001). Globalisation and hiv/aids: Inequality and the boundaries of a symbolic epidemic. *Health Risk Soc.* 3, 89–98. doi: 10.1080/713670174

Barnett, G. A., Chung, C. J., and Park, H. W. (2011). Uncovering transnational hyperlink patterns and web-mediated contents: a new approach based on cracking.com domain. *Soc. Sci. Comput. Rev.* 29, 369–384. doi: 10.1177/0894439310382519

Barnett, G. A., and Park, H. W. (2014). Examining the international internet using multiple measures: new methods for measuring the communication base of globalized cyberspace. *Quality Quant.* 48, 563–575. doi: 10.1007/s11135-012-9787-z

Benkler, Y. (2006). *The Wealth of Networks. How Social Production Transforms Markets and Freedom*. Yale: Yale University Press.

Berry, C. (Ed.). (2010). *Electronic Elsewheres: Media, Technology, and the Experience of Social Space, volume 17 of Public Worlds*. Minneapolis, MN: University of Minnesota Press.

Bhambra, G. K. (2014). Postcolonial and decolonial dialogues. *Postcolonial Stud.* 17, 115–121. doi: 10.1080/13688790.2014.966414

Bhopal, R. S. (2014). *Migration, Ethnicity, Race, and Health in Multicultural Societies, 2nd Edn*. Oxford: Oxford University Press.

Boyd, D., and Crawford, K. (2012). Critical questions for big data. *Inform. Commun. Soc.* 15, 662–679. doi: 10.1080/1369118X.2012.678878

Bruns, A. (2008). Life beyond the public sphere: towards a networked model for political deliberation. *Inform. Polity* 13, 71–85. doi: 10.3233/IP-2008-0141

Bruns, A., and Burgess, J. (2011). "The use of twitter hashtags in the formation of *ad-hoc* publics," in *6th European Consortium for Political Research General Conference. Reykjavik University of Iceland, Reykjavik Series: In 6th European Consortium for Political Research General Conference*.

Bury, M. R. (1986). Social constructionism and the development of medical sociology. *Sociol. Health Illness* 8, 137–169. doi: 10.1111/1467-9566.ep11340129

Cairncross, F. (2001). *The Death of Distance: How the Communications Revolution Is Changing Our Lives*. Boston, MA: Harvard Business School.

Castells, M. (2008). *The Rise of the Network Society, Volume Economy, Society and Culture/Manuel Castells, Vol. 1 of The Information Age, 2nd Edn*. Chichester, West Sussex; Malden, MA: Wiley-Blackwell.

Chu, S.-C., and Choi, S. M. (2010). Social capital and self-presentation on social networking sites: a comparative study of chinese and american young generations. *Chin. J. Commun.* 3, 402–420. doi: 10.1080/17544750.2010.516575

Conrad, P., and Barker, K. K. (2010). The social construction of illness: key insights and policy implications. *J. Health Soc. Behav.* 51, 67–79. doi: 10.1177/0022146510383495

Deutschmann, E. (2022). *Mapping the Transnational World: How We Move and Communicate across Borders, and Why It Matters*. Princeton, NJ: Oxford: Princeton University Press.

Deutschmann, E., Delhey, J. A., Verbalyte, M., and Aplowski, A. (2018). The power of contact: Europe as a network of transnational attachment. *Eur. J. Polit. Res.* 57, 963–988. doi: 10.1111/1475-6765.12261

Dimitrov, D., Baran, E., Fafalios, P., Yu, R., Zhu, X., Zloch, M., et al. (2020). "Tweetscov19-a knowledge base of semantically annotated tweets about the COVID-19 pandemic," in *Proceedings of the 29th ACM International Conference on Information Knowledge Management*, eds M. d'Aquin, S. Dietze, C. Hauff, E. Curry, and P. Cudre Mauroux (Galway: ACM), 2991–2998.

Edelmann, A., Wolff, T., Montagne, D., and Bail, C. A. (2020). Computational social science and sociology. *Annu. Rev. Sociol.* 46, 61–81. doi: 10.1146/annurev-soc-121919-054621

El-Kishky, A., Markovich, T., Park, S., Verma, C., Kim, B., Eskander, R., et al. (2022). Twhin: embedding the twitter heterogeneous information network for personalized recommendation. *arXiv preprint* arXiv:2202.05387. doi: 10.48550/arXiv.2202.05387

Eriksson Krutrök, M., and Lindgren, S. (2018). Continued contexts of terror: Analyzing temporal patterns of hashtag co-occurrence as discursive articulations. *Soc. Media Soc.* 4, 205630511881364. doi: 10.1177/2056305118813649

Fernandez, M., Asif, M., and Alani, H. (2018). "Understanding the roots of radicalisation on twitter," in *Proceedings of the 10th ACM Conference on Web Science* (Amsterdam), 1–10.

Gygli, S., Haelg, F., Potrafke, N., and Sturm, J.-E. (2019). The kof globalisation index-revisited. *Rev. Int. Organ.* 14, 543–574. doi: 10.1007/s11558-019-09344-2

Hale, S. A. (2012). Net increase? cross-lingual linking in the blogosphere. *J. Comput. Med. Commun.* 17, 135–151. doi: 10.1111/j.1083-6101.2011.01568.x

Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., et al. (2021). A global panel database of pandemic policies (oxford covid-19 government response tracker). *Nat. Hum. Behav.* 5, 529–538. doi: 10.1038/s41562-021-01079-8

Haunschild, R., Leydesdorff, L., Bornmann, L., Hellsten, I., and Marx, W. (2019). Does the public discuss other topics on climate change than researchers? a comparison of explorative networks based on author keywords and hashtags. *J. Informetr.* 13, 695–707. doi: 10.1016/j.joi.2019.03.008

Held, D. (1997). Democracy and globalization. *Glob. Governance* 3, 251–267. doi: 10.1163/19426720-00303003

Hong, S., and Na, J. (2018). How facebook is perceived and used by people across cultures. *Soc. Psychol. Pers. Sci.* 9, 435–443. doi: 10.1177/1948550617711227

Imran, M., Qazi, U., and Ofli, F. (2021). Tbcov: two billion multilingual covid-19 tweets with sentiment, entity, geo, and gender labels. *arXiv:2110.03664*. doi: 10.3390/data7010008

Jacobson, S., Myung, E., and Johnson, S. L. (2016). Open media or echo chamber: the use of links in audience discussions on the facebook pages of partisan news organizations. *Inform. Commun. Soc.* 19, 875–891. doi: 10.1080/1369118X.2015.1064461

Jahanbin, K., and Rahmanian, V. (2020). Using twitter and web news mining to predict covid-19 outbreak. *Asian Pac. J. Trop. Med.* 13, 378. doi: 10.4103/1995-7645.279651

Kantner, C. (2015). "National media as transnational discourse arenas: the case of humanitarian military interventions," in *European Public Spheres, Contemporary European Politics*, ed T. Risse (Cambridge: Cambridge University Press), 84–107.

King, N. B. (2002). Security, disease, commerce. *Soc. Stud. Sci.* 32, 763–789. doi: 10.1177/030631270203200507

Klovdahl, A. S., Potterat, J. J., Woodhouse, D. E., Muth, J. B., Muth, S. Q., and Darrow, W. W. (1994). Social networks and infectious disease: the colorado springs study. *Soc. Sci. Med.* 38, 79–88. doi: 10.1016/0277-9536(94)90302-6

Kong, X., Yu, P. S., Ding, Y., and Wild, D. J. (2012). "Meta path-based collective classification in heterogeneous information networks," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (Maui), 1567–1571.

Koopmans, R., and Statham, P. (2010). *Theoretical Framework, Research Design, and Methods. Communication, Society and Politics*. Cambridge: Cambridge University Press.

Koopmans, R., and Zimmermann, A. (2010). "Transnational political communication on the internet," in *The Making of a European Public Sphere, Communication, Society and Politics*, eds R. Koopmans, and P. Statham (Cambridge: Cambridge University Press), 171–194.

Kuchler, T., Russel, D., and Stroebel, J. (2020). The geographic spread of covid-19 correlates with the structure of social networks as measured by facebook. *National Bureau of Economic Research*, Working Paper 26990. Available online at: https://www.nber.org/papers/w26990

Li, Y., and Nicholson Jr, H. L. (2021). When "model minorities" become "yellow peril"–othering and the racialization of asian americans in the covid-19 pandemic. *Sociol. Compass* 15, e12849. doi: 10.1111/soc4.12849

Malik, M. M., Lamba, H., Nakos, C., and Pfeffer, J. (2015). "Populations bias in geotagged tweets," in *Proceedings of the 2015 International AAAI Conference on Web and Social Media* (Oxford).

Mellado, C., Hallin, D., Cárcamo, L., Alfaro, R., Jackson, D., Humanes, M. L., et al. (2021). Sourcing pandemic news: a cross-national computational analysis of mainstream media coverage of covid-19 on facebook, twitter, and instagram. *Digital J.* 9, 1261–1285. doi: 10.1080/21670811.2021.1942114

Mellon, J., and Prosser, C. (2017). Twitter and facebook are not representative of the general population: political attitudes and demographics of british social media users. *Res. Politics* 4, 205316801772000. doi: 10.1177/2053168017720008

Monson, S. (2017). Ebola as african: American media discourses of panic and otherization. *Afr. Today* 63, 3. doi: 10.2979/africatoday.63.3.02

Opiłowska, E. (2021). The covid-19 crisis: the end of a borderless europe? *Eur. Soc.* 23, S589-S600. doi: 10.1080/14616696.2020.1833065

Özkula, S. M., Reilly, P. J., and Hayes, J. (2022). Easy data, same old platforms? a systematic review of digital activism methodologies. *Inform. Commun. Soc.* 1–20. doi: 10.1080/1369118X.2021.2013918

Papacharissi, Z. (2015). *Affective Publics: Sentiment, Technology, and Politics. Oxford Studies in Digital Politics*. Oxford: Oxford University Press.

Pfetsch, B., and Heft, A. (2015). "Theorizing communication flows within a european public sphere," in *European Public Spheres, Contemporary European Politics*, ed T. Risse (Cambridge: Cambridge University Press), 29–52.

Poblete, B., Garcia, R., Mendoza, M., and Jaimes, A. (2011). "Do all birds tweet the same? characterizing twitter around the world," in *CIKM'11*, eds I. Ounis, I. Ruthven, and C. Macdonald (New York, NY: ACM), 1025.

Reny, T. T., and Barreto, M. A. (2022). Xenophobia in the time of pandemic: othering, anti-asian attitudes, and covid-19. *Politics Groups Identities* 10, 1–24. doi: 10.1080/21565503.2020.1769693

Reuters Institute (2020). Digital news report 2020. Available online at: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf

Risse, T. (2010). *A community of Europeans? Transnational Identities and Public Spheres. Cornell Paperbacks*. Ithaca, NY: Cornell University Press.

Risse, T. (Ed.). (2015). *European Public Spheres: Politics is Back. Contemporary European Politics*. Cambridge: Cambridge University Press.

Roy, M., Moreau, N., Rousseau, C., Mercier, A., Wilson, A., and Atlani-Duault, L. (2020). Ebola and localized blame on social media: analysis of twitter and facebook conversations during the 2014-2015 ebola epidemic. *Cult. Med. Psychiatry* 44, 56–79. doi: 10.1007/s11013-019-09635-8

Ruiz-Soler, J. (2018). The last will be the first. a study of european issue publics on twitter. *Partecipazione Conflitto* 11, 423–47.

Ruths, D., and Pfeffer, J. (2014). Social sciences. social media for large studies of behavior. *Science* 346, 1063–1064. doi: 10.1126/science.346.6213.1063

Schünemann, W. J. (2020). Ready for the world? measuring the (trans-)national quality of political issue publics on twitter. *Media and Communication* 8, 40–52. doi: 10.17645/mac.v8i4.3162

Sheldon, P., Rauschnabel, P. A., Antony, M. G., and Car, S. (2017). A cross-cultural comparison of croatian and american social network sites: exploring cultural differences in motives for instagram use. *Comput. Hum. Behav.* 75, 643–651. doi: 10.1016/j.chb.2017.06.009

Shi, C., Kong, X., Huang, Y., Philip, S. Y., and Wu, B. (2014). Hetesim: a general framework for relevance measure in heterogeneous networks. *IEEE Trans. Knowl. Data Eng.* 26, 2479–2492. doi: 10.1109/TKDE.2013.2297920

Shi, C., Kong, X., Yu, P. S., Xie, S., and Wu, B. (2012). "Relevance search in heterogeneous networks," in *Proceedings of the 15th International Conference on Extending Database Technology* (Berlin), 180–191.

Shneor, R., and Efrat, K. (2014). Analyzing the impact of culture on average time spent on social networking sites. *J. Promot. Manag.* 20, 413–435. doi: 10.1080/10496491.2014.930281

Sowa, J. F. (2014). *Principles of Semantic Networks: Explorations in the Representation of Knowledge*. Burlington, NJ: Elsevier Science.

State, B., Park, P., Weber, I., and Macy, M. (2015). The mesh of civilizations in the global network of digital communication. *PLoS ONE* 10, e0122543. doi: 10.1371/journal.pone.0122543

Steinskog, A., Therkelsen, J., and Gambäck, B. (2017). "Twitter topic modeling by tweet aggregation," in *Proceedings of the 21st Nordic Conference on Computational Linguistics* (Gothenburg: Association for Computational Linguistics), 77–86.

Stier, S., Froio, C., and Schünemann, W. J. (2021). Going transnational? candidates' transnational linkages on twitter during the 2019 european parliament elections. *West Eur. Polit.* 44, 1–27. doi: 10.1080/01402382.2020.1812267

Straubhaar, J. (1991). Beyond media imperialism: assymetrical interdependence and cultural proximity. *Crit. Stud. Mass Commun.* 8, 39–59. doi: 10.1080/15295039109366779

Straubhaar, J. (2010). *World Television: From Global to Local*. Thousand Oaks, CA: Sage Publications.

Straubhaar, J. (2015). Global, regional, transnational, translocal. *Media Ind. J.* 1, 309. doi: 10.3998/mij.15031809.0001.309

Sun, Y., and Han, J. (2013). Mining heterogeneous information networks: a structural analysis approach. *Acm Sigkdd Explorat. Newslett.* 14, 20–28. doi: 10.1145/2481244.2481248

Sun, Y., Han, J., Yan, X., Yu, P. S., and Wu, T. (2011). Pathsim: meta path-based top-k similarity search in heterogeneous information networks. *Proc. VLDB Endowment* 4, 992–1003. doi: 10.14778/3402707.3402736

Sun, Y., Norick, B., Han, J., Yan, X., Yu, P. S., and Yu, X. (2013). Pathselclus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. *ACM Trans. Knowl. Discovery Data* 7, 1–23. doi: 10.1145/2513092.2500492

Takhteyev, Y., Gruzd, A., and Wellman, B. (2012). Geography of twitter networks. *Soc. Netw.* 34, 73–81. doi: 10.1016/j.socnet.2011.05.006

Taneja, H., and Webster, J. G. (2016). How do global audiences take shape? the role of institutions and culture in patterns of web use. *J. Commun.* 66, 161–182. doi: 10.1111/jcom.12200

von Unger, H., Scott, P., and Odukoya, D. (2018). "Using skad to analyse classification practices in public health. methodological reflections on the research process," in *The Sociology of Knowledge Approach to Discourse, Routledge Advances in Sociology*, eds R. Keller, A. -K. Hornidge, and W. J. Schünemann (Oxon, MD; New York, NY: Routledge; Abingdon;), 169–185.

von Unger, H., Scott, P., and Odukoya, D. (2019). Constructing im/migrants and ethnic minority groups as 'carriers of disease': power effects of categorization practices in tuberculosis health reporting in the uk and germany. *Ethnicities* 19, 518–534. doi: 10.1177/1468796819833426

Wallaschek, S., Kaushik, K., Verbalyte, M., Sojka, A., Sorci, G., Trenz, H.-J., et al. (2022). Same same but different? gender politics and (trans-)national value contestation in europe on twitter. *Politics Governance* 10, 146–160. doi: 10.17645/pag.v10i1.4751

Wang, Z. (2021). From crisis to nationalism? *Chin. Polit. Sci. Rev.* 6, 20–39. doi: 10.1007/s41111-020-00169-8

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Sociology

BRIEF RESEARCH REPORT
published: 06 July 2022
doi: 10.3389/fsoc.2022.885784

Combining Survey and Social Media Data: Respondents' Opinions on COVID-19 Measures and Their Willingness to Provide Their Social Media Account Information

Markus Hadler[1,2], *Beate Klösch*[1], *Markus Reiter-Haas*[3] *and Elisabeth Lex*[3]*

[1] Department of Sociology, University of Graz, Graz, Austria, [2] Department of Sociology, Macquarie University, Sydney, NSW, Australia, [3] Institute of Interactive Systems and Data Science, Graz University of Technology, Graz, Austria

Research on combining social survey responses and social media posts has shown that the willingness to share social media accounts in surveys depends on the mode of the survey and certain socio-demographics of the respondents. We add new insights to this research by demonstrating that the willingness to share their Facebook and Twitter accounts also depends on the respondents' opinions on specific topics. Furthermore, we extend previous research by actually accessing their social media accounts and checking whether survey responses and tweets are coherent. Our analyses indicate that survey respondents who are willing to share their social media accounts hold more positive attitudes toward COVID-19 measures. The same pattern holds true when comparing their sentiments to a larger Twitter collection. Our results highlight another source of sampling bias when combining survey and social media data: a bias due to specific views, which might be related to social desirability.

Keywords: survey, social media, Facebook, Twitter, polarization, COVID-19, sentiment analysis, qualitative content analysis

OPEN ACCESS

Edited by:
Tobias Wolbring, University of Erlangen Nuremberg, Germany

Reviewed by:
Heinz Leitgöb, Catholic University of Eichstätt-Ingolstadt, Germany
Christoph Kern, University of Mannheim, Germany

***Correspondence:**
Elisabeth Lex
elisabeth.lex@tugraz.at

Specialty section:
This article was submitted to Sociological Theory, a section of the journal Frontiers in Sociology

Received: *28 February 2022*
Accepted: *30 May 2022*
Published: *06 July 2022*

Citation:
Hadler M, Klösch B, Reiter-Haas M and Lex E (2022) Combining Survey and Social Media Data: Respondents' Opinions on COVID-19 Measures and Their Willingness to Provide Their Social Media Account Information. Front. Sociol. 7:885784. doi: 10.3389/fsoc.2022.885784

INTRODUCTION

Combining survey and social media data has become more common over the last few years (Hill et al., 2019; Stier et al., 2019). Researchers have used this approach to study substantive questions such as political preferences and filter bubbles (Eady et al., 2019; Wolfowicz et al., 2021), and to discuss ethical challenges (Breuer et al., 2020, Sloan et al., 2020). This approach has also been used to address methodological issues such as using tweets to validate survey responses (Henderson et al., 2021) and researching possible biases in the willingness to share data (see Al Baghal et al., 2019; Mneimneh et al., 2021). Our paper focuses on the latter, thus expanding the research on biases in the willingness to share data. In other words, our research focuses on the differences between survey respondents who share their social media accounts and those who do not by considering attitudes and by accessing the respondents' social media accounts.

Previous research has shown that the willingness of survey respondents to provide additional data is limited. Revilla et al. (2019) provide an overview of various studies that consider the readiness of survey respondents to use additional tools such as geolocating, GPS tracking, and visual data capturing. The reported values varied widely. For example, the readiness to allow GPS tracking was in the 30% range, yet was around only 11% for the readiness to allow the respondent's child to wear a wristband that sends information to an online site. Studies considering the consent

to share social media accounts report a willingness of between 24 and 45% (Al Baghal et al., 2019; Mneimneh et al., 2021).

Given that only a limited number of survey respondents grant access to their social media accounts, combined data sets capture only a fraction of the data from the total number of survey respondents who actually use social media. Thus, using such samples to draw conclusions regarding an entire population is problematic, at the very least (Sen et al., 2021). If there is any systematic deviation, in the sense that certain groups are not willing to share their data, the linked dataset is biased. Al Baghal et al. (2019) show that the rate of agreement varies with the mode of the survey—with higher consent rates in face-to-face surveys—and with differences regarding gender and age, although those effects are inconsistent. Mneimneh et al. (2021) demonstrate that self-reported Twitter usage patterns can be important predictors of the willingness to share accounts. Henderson et al. (2021) were aware that consent to provide access is non-random. Therefore, they compared sample characteristics in their research, which indicates minor differences in terms of age, gender, education, race, and income.

The studies cited in this previous paragraph provide critical insights into this topic. Yet, our paper addresses novel aspects. While Al Baghal et al. (2019) and Mneimneh et al. (2021) base their analysis on the reported willingness to provide access, we also access the social media accounts of our respondents and thus can check whether respondents actually provide valid account names. Henderson et al. (2021) compared verified Twitter users to all Twitter users in their survey, but did not consider non-users or users whose accounts were not verified. We considered all these groups and conducted multivariate analyses testing the effects of various socio-demographics and attitudes on the respondents' willingness to share their account information. Specifically, we considered the effects of the respondents' opinions on COVID-19 measures.

Opinions on COVID-19 seem to be particularly suited for our purpose, as the pandemic gave rise to various conspiracy theories and associated skepticism toward science and scientific advice (Chayinska et al., 2021; Priniski and Holyoak, 2022). Our expectation was that respondents who oppose the COVID-19 measures are less likely to grant access to their social media accounts. One possible explanation[1] to consider is "social desirability." The social desirability bias assumes that respondents would like to be seen in a favorable light by the researcher and are thus likelier to underreport socially undesirable views and actions (Phillips and Clancy, 1972; Krumpal, 2013; Henderson et al., 2021). Social desirability effects are weak in anonymous settings, such as online surveys. They are more significant in face-to-face interviews and when the respondent is known to the researcher. Providing access to one's social media account can remove the anonymity of an online survey, especially when the social media account reflects the real name of a respondent or includes additional information that allows identification of the account holder. Given that we inform our respondents that we collect the social media information for a scientific purpose and that previous research indicates science skepticism among COVID-19 deniers, we expected, as stated above, a lower consent rate among respondents who oppose the COVID-19 measures.

In sum, these considerations lead to the following main research hypotheses: (a) The willingness of survey respondents to share social media accounts depends on their attitudes toward a specific topic, in our case, COVID-19 measures. We expect (b) a bias due to socially desirable responses, e.g., a higher willingness to share data among respondents who are in favor of COVID-19 measures. Finally, we will also investigate whether there is a difference between respondents who consent to share their data and respondents whose accounts can be accessed successfully.

The following section explains how the data was collected and used for our study. The data consists of an online survey of the adult population of Austria, Germany, and the German-speaking parts of Switzerland; the tweets of our survey respondents; and all tweets that match predefined search parameters related to COVID-19 during the survey period. The results section starts with a report on the different social media usage patterns of our survey respondents and their willingness to share Facebook and Twitter accounts. Afterwards, we compare the attitudes expressed in our survey to related tweets that were crawled during the survey period. Overall, our findings confirm our expectation that respondents who share their social media data hold more positive opinions toward the COVID-19 measures, and that merged data sets can be biased with regard to specific opinions.

METHODS

Our analyses are based on a public opinion survey, the tweets of survey respondents who provided their account information, and Twitter data during the same period of time. The cross-sectional survey was fielded online in the summer of 2020 in Austria, Germany, and the German-speaking parts of Switzerland. A total of 2,560 individuals participated. The individuals were selected based on representative quotas reflecting the official distribution of gender, age, and federal state/canton in the three countries. Having met these quotas, it can be assumed that the sample resembles the characteristics of the total population. However, it cannot be considered a random sample. Therefore, we do not draw any conclusions regarding the general population and focus only on the biases within our sample.

Of the survey respondents, 67% were from Germany, 22% from Austria, and the remaining 11% from Switzerland. Austrians were oversampled due to a specific interest in regional differences by some members of our research team. The survey included questions on attitudes toward the COVID-19 crisis, the use of various social media platforms, and socio-demographic information. Respondents were also asked if they were willing to provide the name of their Facebook and Twitter handles and were subsequently asked for them.

Linking survey data with Twitter data requires specific ethical considerations (Sloan et al., 2020). Our respondents

[1]Other possible explanations are incentives, privacy concerns, social trust, and trust in institutions, given that previous research on this subject has established links between COVID-19 and these aspects (see, e.g., Bian et al., 2020; Hafner-Fink and Uhan, 2020; Kreuter et al., 2020).

were informed about the content of this research, that their participation is voluntary, and that all information will remain confidential. Hence, the stored data does not include any information that would allow others to identify a specific person, including Facebook and Twitter handles or any tweets. The survey data, including more details on the fieldwork, are available publicly *via* Hadler et al. (2021)[2].

The data for the dependent variable—the willingness to provide account information—was first derived from the responses to the public survey. This data includes the following groups for both Facebook and Twitter, respectively: (a) respondents without an account, (b) account holders who are not willing to provide their account information, and (c) holders who provided their account information. For Twitter users among our survey respondents, we were able to identify another group: (d) respondents whose Twitter accounts were accessed successfully. As for Facebook, we were not able to access these accounts as we do not possess the required licenses mandated by Facebook's terms and conditions.

The independent variables in this study (see **Table 1**) are the socio-demographics of age, gender, and education. We included these basic socio-demographic variables because, in related research, they were considered to have had some effects (see Al Baghal et al., 2019; Henderson et al., 2021; Mneimneh et al., 2021). In addition, we captured attitudes toward three COVID-19 measures that were the subject of intense political and social debate prior to and during our fieldwork: (1) "Once there is a vaccine against the coronavirus, there should be a mandatory vaccination for all;" (2) "To contain the spread of the coronavirus, contact tracing data (e.g., *via* apps) should be collected;" and (3) "I only wear a face mask when it is required by the government, and not voluntarily." Responses were measured on a five-point scale, where 1 = absolutely disagree, and 5 = absolutely agree. Since the third item regarding the wearing of face masks was formulated inversely, we recoded it prior to the analysis. The three variables all correlate significantly with each other and have a Cronbach's alpha reliability score of $\alpha = 0.634$ when combined to a single scale. Given the moderate Cronbach's α, we also include the three items separately in our regressions and report these results as noted in our tables.

We were able to collect 221 tweets from the Twitter accounts provided by our survey respondents. We conducted a qualitative content analysis of these tweets, as they are a special type of text material that is limited to 280 characters and often contains answers to previous tweets, links, images, and more. Furthermore, we wanted to capture explicit opinions about the COVID-19 measures and the pandemic and not rely on automated or standardized methods. Therefore, the tweets were coded inductively using the qualitative content analysis approach, and agreement with the COVID-19 measures was assigned manually by two researchers independently using an ordinal 5-point scale, as in the survey. Finally, these scores were compared in terms of congruence with the survey data of the respective respondents.

[2]https://doi.org/10.11587/OVHKTR

TABLE 1 | Characteristics of the survey sample.

Variables	Mean (SD) or %
Social media usage: see **Table 2**	
Opinions toward COVID-19 measures (1 = disagree and 5 = agree):	
Once there is a vaccine against the coronavirus, there should be mandatory vaccination for all.	3.19 (1.52)
To contain the spread of the coronavirus, contact tracing data (e.g., *via* apps) should be collected.	3.10 (1.39)
I only wear a face mask when it is required by the government, and not voluntarily.	2.99 (1.51)
Index	3.10 (1.12)
Socio-demographic variables	
Female	50.4%
Age	44.34 (13.80)
Education	
Compulsory school	35.2%
Vocational training	11.6%
High school degree	23.9%
University degree	29.3%

n = 2,560; online survey conducted in Austria, Germany, and Switzerland in 2020. See methods section for details.

Our third source is a collection of tweets posted on COVID-19-related topics during the survey period. We used the keyword and account list from Chen et al. (2020) to collect German tweets from the full-archive search API (Application Programming Interface) using Academic Research access *via* twarc2. We filtered the tweets according to three predefined word stems that resemble the three prevention measures. Specifically, we used the German word stem "impf" for vaccination ($n = 12{,}336$ tweets), "trac" for contact tracing ($n = 1{,}391$), and "mask" for mask wearing ($n = 35{,}044$). As this selection includes all relevant tweets during this period, we use the commonly used data-sciences term "collection" for this source. Subsequently, we conducted a sentiment analysis using the Python library TextBlob with the German language extension, which includes a polarity lexicon for sentiment extraction. The extracted sentiments range on a scale of −1 for negative sentiment to +1 for positive sentiment. Similar to Al Baghal et al. (2021), we averaged the sentiment per Twitter account and excluded accounts that did not express any relevant sentiment or did not contain sufficient information for extracting a sentiment.

RESULTS

Social Media Usage of Our Survey Respondents and Access to Their Social Media Accounts

Table 2 provides, first, a descriptive overview of the social media usage of our survey respondents and their willingness to provide their account information. These results are based on a set of

TABLE 2 | Usage of different platforms and the willingness to provide access (survey data).

	Facebook		Twitter	
Account holders (% of all respondents)	1,774	69.3%	404	15.8%
Active users (% of holders)	700	39.5%	141	34.9%
Account provided (% of holders)	617	34.8%	119	29.5%
Successfully accessed (% of holders)	N.A.*		79	19.6%
n	2,560		2,560	

**Account access restricted by Facebook's terms and conditions.*

questions that were asked toward the end of our survey: (1) "Do you have a private Facebook account? (yes/no);" (2) "How would you describe the way you use Facebook? (actively posting vs. more passively reading);" (3) "We would like to find out who is using Facebook and for which purposes. If you provide us access to your account, we assure to keep your personal information confidential (access granted vs. no access);" and (4) for those who granted access, "What is your username?" The same set of questions was also used for Twitter. Second, **Table 2** depicts the number of accounts we accessed successfully on Twitter.

The data shown in **Table 2** confirms the already known difference between the usage of Facebook and Twitter in German-speaking countries in the sense that the former is used far more often by our respondents (69 vs. 16%). However, the proportion of account holders who consider themselves active users and of survey respondents who are willing to provide their account information are much more similar for both platforms. Around 40% of the Facebook account holders consider themselves active users, while around 35% of the Twitter account holders consider themselves active users. As for the willingness to provide their account information, 35% of the Facebook account holders provided their information, and around 30% of the Twitter account holders provided theirs. As for actual access to their social media accounts, we did not access the Facebook accounts due to the specific terms and conditions of the platform. As for Twitter, we were able to access the accounts of 79 respondents (20% of the account holders among our respondents). Forty respondents provided an incorrect Twitter name or a protected account.

Attitudes Toward the COVID-19 Measures Across Different Social Media Platforms and Users

One of the main goals of our paper is to analyze the association between attitudes toward the COVID-19 measures and the willingness to share one's social media information in a public opinion survey. **Table 3** provides several statistics for all users of a platform compared to our respondents and users who granted access. Furthermore, we provide additional details for Twitter users among our survey respondents, as we were able to differentiate between accounts that we accessed successfully and accounts that we could not access. The bottom part of **Table 3** presents statistics based on tweets: first, the sentiments shown in the tweets of our survey respondents and second, the sentiments shown in the overall collection of tweets during the survey period.

Before we discuss the survey results and social media content, we also have to ensure that there is a match between an individual response in the survey and the respondent's postings on social media. This initial step is necessary to ensure that our comparison of survey responses and Twitter sentiments is valid. Of the 79 Twitter accounts that were successfully accessed, 20 accounts (i.e., survey participants) posted about the COVID-19 pandemic. Overall, a total of 221 tweets from these 20 accounts were analyzed using qualitative content analysis. All tweets were original postings (i.e., no re-tweets). The manual classification of the tweets by two independent researchers regarding the user's opinion toward the pandemic and accompanying measures shows a binary inter-annotator agreement of $\alpha = 0.7$. The assessment of the coherence between (a) the survey answers of our respondents, (b) their tweets on the COVID-19 measures, and (c) their overall COVID-19 Twitter sentiments shows a match in all but eight total cases (out of 42 pairwise comparisons[3])—that is, a match of 81%. This agreement indicates a relatively high level of congruence regarding the opinions toward the pandemic and the related measures between the survey data and the Twitter data, which lends support to our following comparison of survey results and sentiment analysis.

Table 3 shows that the Twitter account holders among our survey respondents are generally more in favor of the COVID-19 measures than the overall sample, whereas Facebook account holders express more of an average sentiment. The mean value across the three COVID-19 measures—vaccination, mask wearing, and contact tracing (on a scale of 1–5, 1 = absolutely disagree)—is 3.09 for the overall survey respondents, 3.27 for Twitter account holders, and 3.06 for Facebook account holders. A positive bias is visible for respondents who are willing to share their Facebook account (mean = 3.18), respondents whose Twitter accounts were actually accessed (mean = 3.40), and even more for respondents who actually posted on Twitter on this topic (mean value = 3.63).

Alongside the survey data, we also analyzed Twitter data. The sentiment analysis considered all German tweets published during the same time period as our survey and that included either positive or negative sentiments regarding the three COVID-19 measures. To deal with the fact that a single account can post multiple tweets and thus lead to an imbalance in numbers (Al Baghal et al., 2021), we based the statistics in **Table 3** on the average value for each account. The results show that the overall sentiment regarding the three COVID-19 measures is 0.08. The actual tweets of our 20 survey respondents who posted on COVID-19-related matters indicate a mean value of 0.16, which is larger than the average within the Twitter collection. Hence, we observe the same bias in the tweets as in the survey data—the tweets of respondents who shared their account information are more positive toward the COVID-19 measures than those of the larger Twitter collection.

[3]Forty-two comparisons reflect the total number of possible pairwise comparisons, i.e., a respondent had valid answers in at least two variables.

TABLE 3 | Different usage groups and their opinions on COVID-19 measures.

Dataset and variables	Statistics	Mean	Min	Max	1. Quartile	2. Quartile—Median	3. Quartile	*n*
All survey respondents	Vacc.	3.19	1	5	2	4	5	2,497
	Mask	2.99	1	5	2	3	4	2,523
	CT	3.10	1	5	2	3	4	2,502
	Index	3.09	1	5	2.33	3.33	4	2,541
Survey respondents who have a Facebook account	Vacc.	3.12	1	5	2	3	5	1,732
	Mask	2.95	1	5	2	3	4	1,752
	CT	3.10	1	5	2	3	4	1,735
	Index	3.06	1	5	2.33	3.33	4	1,761
Survey respondents who provided us with their Facebook account name	Vacc.	3.27	1	5	2	4	5	603
	Mask	2.95	1	5	2	3	4	609
	CT	3.34	1	5	2	4	4	608
	Index	3.18	1	5	2.33	3.33	4	612
Survey respondents who have a Twitter account	Vacc.	3.28	1	5	2	4	5	396
	Mask	3.27	1	5	2	4	5	401
	CT	3.28	1	5	2	4	4	399
	Index	3.27	1	5	2.33	3.33	4	403
Survey respondents whose Twitter accounts were accessible	Vacc.	3.24	1	5	2	4	4	78
	Mask	3.38	1	5	2	4	4	79
	CT	3.56	1	5	3	4	4	79
	Index	3.40	1	4.67	2.67	3.67	4	79
Survey respondents who tweeted about COVID-19	Vacc.	3.53	1	5	2	4	5	19
	Mask	3.60	1	5	3	4	4.75	20
	CT	3.75	1	5	4	4	4	20
	Index	3.63	2	4.67	3.08	3.83	4.33	20
Twitter accounts that express sentiment regarding COVID-19 measures in the survey time period	Vacc.	0.21	−1	1	−0.19	0.23	0.7	4,465 (8,344 tweets)
	Mask	0.02	−1	1	−0.17	−0.03	0.25	11,537 (26,029 tweets)
	CT	0.16	−1	1	−0.18	0.23	0.44	673 (865 tweets)
	Index	0.08	−1	1	−0.15	0.06	0.33	14,752 (36,769 tweets)
Twitter accounts of survey respondents that express any sentiment regarding COVID-19	Index	0.16	−0.08	0.5	0.02	0.15	0.28	19 (220 tweets)

Results from the survey data are based on the questions: "Once there is a vaccine against the coronavirus, there should be mandatory vaccination for all." "To contain the spread of the coronavirus, contact tracing data (e.g., via apps) should be collected." "I only wear a face mask when it is required by the government, and not voluntarily." With 1 = disagree and 5 = agree. Additionally, we calculated all measures by excluding the value "3" which equals "neither/nor" and might match the neutral sentiment in the sentiment analysis of Twitter. The substantive findings remain the same.

Results from Twitter are based on sentiment on the three prevention measures during the survey period. The sentiments are measured as per tweet in a range from −1 for the maximum negative sentiment to +1 for the maximum positive sentiment. The numbers presented in ***Table 3*** *are based on the average of all sentiments of an account. Accounts that do not express any sentiment and tweets with a neutral sentiment are excluded.*

Multivariate Analyses of Factors Influencing the Willingness to Share Social Media Content

So far, we have presented various descriptive analyses of attitudes toward COVID-19 measures and the willingness to share social media content. **Table 4** presents the results of three multinomial regression models that estimate the effect of these COVID-19-related attitudes on the willingness to share account information, controlling for a few socio-demographic variables. The first two regression models use "respondents who are willing to provide their account information" as the reference group being compared to respondents who do not have an account as well as those who did not provide their account information in the survey. The third regression model allows another differentiation for Twitter, as it uses "respondents whose Twitter accounts were actually accessed" as the reference group being compared to "respondents without an account," "respondents who have an account but did not grant access," and "respondents whose accounts were not accessible" (due to incorrect account names or protected accounts).

For Facebook, respondents without an account are older than the reference group, but, otherwise, do not differ significantly from the reference group in terms of gender, education, and attitudes toward COVID-19 measures. For the respondents who do have a Facebook account but are not willing to share their social media information, the regression identifies a significant effect with COVID-19 attitudes. Facebook account holders who oppose the COVID-19 measures are less willing to provide their account information than Facebook users who support the

TABLE 4 | Social media usage and the willingness to provide account information (Multinomial regression).

	Facebook (ref: account provided) *B*-values (*p*-values)		**Twitter (ref: account provided) *B*-values (*p*-values)**		**Twitter (ref: account accessible) *B*-values (*p*-values)**		
	No Account	**Account, but not provided**	**No Account**	**Account, but not provided**	**No Account**	**Account, but not provided**	**Account provided but not accessible**
Intercept	−0.756 (0.016)	0.647 (0.023)	1.389 (0.009)	1.445 (0.019)	1.337 (0.041)	1.397 (0.052)	−2.117 (0.055)
Pro COVID-measures	−0.069 (0.175)	−0.162 (0.001)	−0.279 (0.003)	−0.105 (0.327)	−0.343 (0.003)	−0.170 (0.184)	−0.185 (0.328)
Female	−0.162 (0.144)	0.141 (0.165)	0.726 (0.000)	−0.065 (0.776)	0.939 (0.000)	0.147 (0.590)	0.618 (0.125)
Age	0.033 (0.000)	0.008 (0.056)	0.029 (0.000)	0.000 (0.983)	0.040 (0.000)	0.011 (0.279)	0.032 (0.031)
Compulsory school	−0.192 (0.165)	−0.205 (0.107)	0.206 (0.402)	−0.232 (0.413)	0.088 (0.768)	−0.348 (0.292)	−0.350 (0.486)
Vocational training	−0.234 (0.226)	−0.188 (0.287)	0.191 (0.592)	−0.453 (0.288)	0.123 (0.781)	−0.521 (0.298)	−0.200 (0.783)
High school degree	0.342 (0.028)	0.152 (0.288)	−0.029 (0.906)	−0.155 (0.581)	−0.054 (0.856)	−0.180 (0.583)	−0.085 (0.865)
University degree	Ref.		Ref.		Ref.		
Cox-Snell	0.042		0.048		0.051		
Nagelkerke	0.047		0.074		0.075		
X^2 (*df*)	108.039*** (12)		125.215*** (12)		132.230*** (18)		
n	2,529		2,529		2,529		

Pro-COVID-19 measures (low value = disagreement); age in years; education (reference category = University degree).

We have also taken into account the opinions on the three respective COVID-19 measures separately. This analysis shows that the item on mask wearing is only significant regarding the difference between the reference group and respondents who do not have an account. As for Twitter, the item on contact tracing becomes significant and indicates that respondents, who oppose contact tracing, are also less likely to share their information.

We also considered excluding the middle category (3 = "neither nor," on a scale from 1 to 5) for the variables regarding the COVID-19 measures in the survey data, as we did with tweets with a neutral sentiment. The results are very similar throughout.

$^{***}p < 0.001$.

COVID-19 measures. Socio-demographics, on the other hand, do not have significant effects for this group.

As for Twitter, the results indicate that respondents without an account are older and more often female when compared to the reference group, and are also more often against the COVID-19 measures. Regression model 2 does not indicate any significant differences between respondents who are willing to provide their Twitter accounts and respondents who are not willing to provide their accounts in terms of the variables considered when using the COVID-19 index. Comparing the groups of Twitter account holders whose accounts were accessible and those who (accidentally) provided an incorrect account name shows that older respondents more often reported an incorrect or protected account. Furthermore, when including the three COVID-19 measures separately, the effect of contact tracing becomes significant and indicates that respondents, who oppose contact tracing, are also less likely to share their information Hence, this specific aspect is also associated with the willingness to grant access.

Finally, the explained variances remain low in our models, with Nagelkerke values between 0.047 and 0.075. Given that a sampling error can be divided into a random part and a systematic bias (Sen et al., 2021; see also their supplemental material), a perfect random selection of respondents would be reflected in the absence of any systematic bias (i.e., showing no significant effects of any independent variables and a very low explained variance). Several of our variables are significant and thus indicate a systematic bias between account holders who share their information, those who do not share their information, and, to a certain extent, those who provide incorrect Twitter handles.

DISCUSSION

Our paper set out to investigate the association between attitudes toward the contentious topic of COVID-19 measures and the willingness of survey respondents to provide access to their social media accounts. The overall willingness to provide account information was around 30%, which is more or less in line with the numbers reported in previous studies (Al Baghal et al., 2019; Mneimneh et al., 2021). The overall willingness to provide information did not differ across the socio-demographic variables.

As for attitudes toward the COVID-19 measures, we found that respondents who oppose the measures are less willing to provide their Facebook account information. As for Twitter, the survey showed that Twitter users are generally more in favor of the COVID-19 measures than the other respondents. However, the same negative (albeit not significant) effect of less willingness to provide account information was also visible and in line with the Facebook findings. Furthermore, a separate analysis of the three measures showed that the item on contact tracing would be significant for Twitter as well. In sum, the contentious topic of COVID-19 measures is associated with the willingness to provide social media account details. These findings support our research hypotheses (a) and (b), that opinions on a specific topic are another source of possible biases when asking survey respondents for consent to share their social media accounts. Furthermore,

our analysis also shows that younger Twitter account holders reported incorrect account information or a private account less often than male respondents. In line with our research hypothesis (c), we also see a bias between the sample of "consent given" and its subsample "account accessed successfully." Our results thus add three additional aspects to the results discussed in previous papers (see, for example, Al Baghal et al., 2019; Henderson et al., 2021; Mneimneh et al., 2021).

A tentative explanation for this bias is social desirability (Phillips and Clancy, 1972; Krumpal, 2013; Henderson et al., 2021). Social desirability plays a role when survey respondents are no longer anonymous, which is the case, for example, whenever their social media handle allows them to be identified. In line with our research hypothesis (b), respondents who are in favor of the COVID-19 measures and thus aligned with scientific views are more willing to share their social media accounts with scientists. This explanation, however, remains tentative, as we did not include any questions on the reasons why respondents are willing to share their data. Yet, in a related paper (Klösch et al., 2022), we found that attitudes toward the environment are also associated with the willingness of our respondents to share data. Thus, it seems to be the case that attitudinal dimensions are related to the survey respondents' willingness to share their social media accounts.

Our research has other limitations. In terms of total survey error and its version for online data (Sen et al., 2021), our Twitter collection is limited to tweets that used the three predefined terms and our online access panel to registered respondents. Thus, our results should not be used to draw conclusions regarding the overall population. The matching of survey responses and tweets at the individual level was based only on a limited number of respondents, who, in addition, expressed rather positive views on the COVID-19 measures. This limitation is caused by, first, a shrinking sample size, as not all survey respondents use social media; second, the fact that not all respondents shared their account information; and third, the fact that not all accounts can be accessed. Furthermore, opinions can only be compared if they are expressed. Hence, we can only grasp the views of social media users if they expressed their opinion on COVID-19 measures—a group that constituted 20 account holders in our case.

In sum, we were able to demonstrate that attitudes toward a specific topic are associated with the survey respondents' willingness to grant access to their social media accounts, which adds another dimension to existing research on this topic. We were also able to show that the responses in the survey and the tweets are mostly coherent. Given the limitations in terms of sample size and Facebook account access, future research should revisit this topic using additional social media platforms and larger collections of actual users.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: "(Hadler et al., 2021), 'Polarization in public opinion: Combining social surveys and big data analyses of Twitter (SUF Edition)', https://doi.org/10.11587/OVHKTR."

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This work was supported by TU Graz Open Access Publishing Fund.

ACKNOWLEDGMENTS

We thank the rectorates of University of Graz and Graz University of Technology for supporting this research. We would also like to thank the two reviewers whose comments and suggestions helped improving and clarifying this manuscript.

REFERENCES

Al Baghal, T., Sloan, L., Jessop, C., Williams, M. L., and Burnap, P. (2019). Linking Twitter and survey data: the impact of survey mode and demographics on consent rates across three UK studies. *Soc. Sci. Comput. Rev.* 38, 0894439319828011. doi: 10.1177/0894439319828011

Al Baghal, T., Wenz, A., Sloan, L., and Jessop, C. (2021). Linking Twitter and survey data: asymmetry in quantity and its impact. *EPJ Data Sci.* 10, 32. doi: 10.1140/epjds/s13688-021-00286-7

Bian, Y., Miao, X., Lu, X., Ma, X., and Guo, X. (2020). The emergence of a COVID-19 related social capital: the case of China. *Int. J. Sociol.* 50, 419–433. doi: 10.1080/00207659.2020.1802141

Breuer, J., Bishop, L., and Kinder-Kurlanda, K. (2020). The practical and ethical challenges in acquiring and sharing digital trace data: negotiating public-private partnerships. *New Media Soc.* 22, 2058–2080. doi: 10.1177/1461444820924622

Chayinska, M., Ulug, O. M., Ayanian, A. H., Gratzel, J. C., Brik, T., Anna, K., et al. (2021). Coronavirus conspiracy beliefs and distrust of science predict risky public health behaviours through optimistically biased risk perceptions in Ukraine, Turkey, and Germany. *Group Processes Intergroup Relat.* 1368430220978278. doi: 10.1177/1368430220978278

Chen, E., Lerman, K., and Ferrara, E. (2020). Tracking social media discourse about the COVID-19 pandemic: development of a public coronavirus Twitter data set. *JMIR Public Health Surveill.* 29, e19273. doi: 10.2196/19273

Eady, G., Nagler, J., Guess, A., Zilinsky, J., and Tucker, J. A. (2019). How many people live in political bubbles on social media? Evidence from linked survey and Twitter data. *SAGE Open*. doi: 10.1177/2158244019832705

Hadler, M., Klösch, B., Lex, E., and Reiter-Haas, M. (2021). *Polarization in Public Opinion: Combining Social Surveys and Big Data Analyses of Twitter (SUF Edition)*. Available online at: https://data.aussda.at/dataset.xhtml?persistentId=doi:10.11587/OVHKTR

Hafner-Fink, M., and Uhan, S. (2020). Life and attitudes of Slovenians during the COVID-19 pandemic: the problem of trust. *Int. J. Sociol.* 51, 76–85. doi: 10.1080/00207659.2020.1837480

Henderson, M., Jiang, K., Johnson, M., and Porter, L. (2021). Measuring Twitter use: validating survey-based measures. *Soc. Sci. Comput. Rev.* 39, 1121–1141. doi: 10.1177/0894439319896244

Hill, C. A., Biemer, P., Buskirk, T., Callegaro, M., Córdova Cazar, A. L., Eck, A., et al. (2019). Exploring new statistical Frontiers at the intersection of survey science and big data: convergence at "BigSurv18". *Surv. Res. Methods* 13, 123–135. doi: 10.18148/srm/2019.v1i1.7467

Klösch, B, Hadler, M., Reiter-Haas, M., and Lex, E. (2022). "Social desirability and the willingness to provide social media accounts in surveys. The case of environmental attitudes," in *4th International Conference on Advanced Research Methods and Analytics (CARMA)*.

Kreuter, F., Georg-Christoph, H., Keusch, F., Bähr, S., and Trappmann, M. (2020). Collecting survey and smartphone sensor data with an app: opportunities and challenges around privacy and informed consent. *Soc. Sci. Comput. Rev.* 38, 533–549. doi: 10.1177/0894439318816389

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Qual. Quant.* 47, 2025–2047. doi: 10.1007/s11135-011-9640-9

Mneimneh, Z. N., McClain, C., Bruffaerts, R., and Altwaijri, Y. A. (2021). Evaluating survey consent to social media linkage in three international health surveys. *Res. Soc. Adm. Pharm.* 17, 1091–1100. doi: 10.1016/j.sapharm.2020.08.007

Phillips, D. L., and Clancy, K. J. (1972). Some effects of "social desirability" in survey studies. *Am. J. Sociol.* 77, 921–940. doi: 10.1086/225231

Priniski, J. H., and Holyoak, J. K. (2022). A darkening spring: how preexisting distrust shaped, COVID-19 skepticism. *PLoS ONE* 17:e0263191. doi: 10.1371/journal.pone.0263191

Revilla, M., Couper, M. P., and Ochoa, C. (2019). Willingness of online panelists to perform additional tasks. *Methods Data Analyses* 13, 29. doi: 10.12758/mda.2018.01

Sen, I., Flöck, F., Weller, K., Weiß, B., and Wagner, C. (2021). A total error framework for digital traces of human behavior on online platforms. *Public Opin. Q.* 85, 399–422. doi: 10.1093/poq/nfab018

Sloan, L., Jessop, C., Al Baghal, T., and Williams, M. (2020). Linking survey and Twitter data: informed consent, disclosure, security, and archiving. *J. Empir. Res. Hum. Res. Ethics* 15, 63–76. doi: 10.1177/1556264619853447

Stier, S., Breuer, J., Siegers, P., and Thorson, K. (2019). Integrating survey data and digital trace data: key issues in developing an emerging field. *Soc. Sci. Comput. Rev.* 38, 503–516. doi: 10.1177/0894439319843669

Wolfowicz, M., Weisburd, D., and Hasisi, B. (2021). Examining the interactive effects of the filter bubble and the echo chamber on radicalization. *J. Exp. Criminol.* doi: 10.1007/s11292-021-09471-0

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Big Data

TYPE Original Research
PUBLISHED 11 August 2022
DOI 10.3389/fdata.2022.880554

OPEN ACCESS

EDITED BY
Heinz Leitgöb,
Catholic University of Eichstätt-Ingolstadt, Germany

REVIEWED BY
Tobias Wolbring,
University of Erlangen Nuremberg, Germany
Katharina Meitinger,
Utrecht University, Netherlands

*CORRESPONDENCE
Anna-Carolina Haensch
anna-carolina.haensch@stat.uni-muenchen.de

SPECIALTY SECTION
This article was submitted to Data Science, a section of the journal Frontiers in Big Data

RECEIVED 21 February 2022
ACCEPTED 15 July 2022
PUBLISHED 11 August 2022

CITATION
Haensch A-C, Weiß B, Steins P, Chyrva P and Bitz K (2022) The semi-automatic classification of an open-ended question on panel survey motivation and its application in attrition analysis.
Front. Big Data 5:880554.
doi: 10.3389/fdata.2022.880554

The semi-automatic classification of an open-ended question on panel survey motivation and its application in attrition analysis

Anna-Carolina Haensch[1]*, Bernd Weiß[2], Patricia Steins[3], Priscilla Chyrva[2,3] and Katja Bitz[4]

[1]Department of Statistics, Ludwig-Maximilians-University Munich, Munich, Germany, [2]GESIS-Leibniz-Institute for the Social Sciences, Mannheim, Germany, [3]School of Social Sciences, University of Mannheim, Mannheim, Germany, [4]Faculty of Economics and Social Sciences, Eberhard Karl University of Tübingen, Tübingen, Germany

In this study, we demonstrate how supervised learning can extract interpretable survey motivation measurements from a large number of responses to an open-ended question. We manually coded a subsample of 5,000 responses to an open-ended question on survey motivation from the GESIS Panel (25,000 responses in total); we utilized supervised machine learning to classify the remaining responses. We can demonstrate that the responses on survey motivation in the GESIS Panel are particularly well suited for automated classification, since they are mostly one-dimensional. The evaluation of the test set also indicates very good overall performance. We present the pre-processing steps and methods we used for our data, and by discussing other popular options that might be more suitable in other cases, we also generalize beyond our use case. We also discuss various minor problems, such as a necessary spelling correction. Finally, we can showcase the analytic potential of the resulting categorization of panelists' motivation through an event history analysis of panel dropout. The analytical results allow a close look at respondents' motivations: they span a wide range, from the urge to help to interest in questions or the incentive and the wish to influence those in power through their participation. We conclude our paper by discussing the re-usability of the hand-coded responses for other surveys, including similar open questions to the GESIS Panel question.

KEYWORDS

text analysis, support vector machine (SVM), survey methodology, semi-automated analysis, machine learning, survey research

1. Introduction

Open-ended questions in surveys have become more prominent in recent years thanks to the increased use of web surveys. Responses can now be captured digitally, significantly reducing the cost and human effort involved in capturing the responses. However, a primary concern regarding the inclusion of open-ended questions is the

increased burden on respondents and researchers. Respondents cannot choose from a pre-defined set of answers but have to access the possible range of answers and choose a suitable answer. From the researcher's perspective, the analysis of responses to open-ended answers requires manual coding, which, when relying solely on human coders, will be costly and impractical, especially if the number of responses is high (Züll and Menold, 2019). Therefore, during the last years, researchers have tried to automate parts of the process with the help of computer-assisted content analysis. This encompasses both dictionary-based as well as supervised machine learning-based procedures (Schonlau, 2015; Schonlau and Couper, 2016; Schonlau et al., 2017; Schierholz and Schonlau, 2021). The later ones are potentially very powerful when mapping respondents' responses to substantially relevant categories, but are yet not widely used in the survey context. *In this article, we will demonstrate how useful supervised learning is for categorizing a large number of responses to open-ended questions, in our case, a question on respondent's motivation to participate in a panel. This article also serves as an illustrative example of how to apply supervised learning in the survey context.* The survey from which we take our data is the GESIS Panel (GESIS, 2021), a German probability-based mixed-mode access panel. In the following, we will describe the pre-processing steps of the text corpus, the semi-automated classification. Since we want to generalize beyond our use case, we will also discuss alternative options regarding the pre-processing and coding steps and look at our semi-automated classification's performance. Finally, we will illustrate the benefit of semi-automated classification by conducting a descriptive evaluation of the respondent's motivation from the GESIS Panel and a more advanced analysis of panel dropout. An evaluation of an open-ended question on panel motivation has not yet been conducted at this granularity; to date, survey motivation has been measured much more coarsely (Porst and Briel, 1995; Brüggen et al., 2011). Therefore, our analysis can provide an unbiased yet clear view of respondents' motivations by combining the openness of the question and the automatic categorization.

2. Open-ended questions in survey research

2.1. Advantages and disadvantages of open-ended questions

Survey questions that do not provide a set of response options but demand respondents to formulate a response in their own words are known as open-ended questions (Krosnick and Presser, 2010). Open-ended questions are recommended when there is an unknown range of possible answers for the subjects of interest. For instance, we are interested in respondents' motivations to participate in a survey. However, we only have some isolated examples for the possible range of answers, and it is unclear if they are valid for panel participants (in comparison to cross-sectional surveys). Another reason for open-ended questions is an excessively long list of possible answers. An example for such an item with a list of several hundred answer categories would be occupation coding (Schierholz and Schonlau, 2021). Open-ended questions also have the advantage of avoiding being directive in a particular direction through the provided options. Without prompts, respondents have to reflect on the question on a deeper level than choosing a random answer. On the other hand, the need for deeper reflection increases respondents' burden, which can lead to more "don't know" responses or item nonresponse than closed questions do (Krosnick and Presser, 2010). Dillman et al. (2009) recommended using open-ended questions only rarely to not overburden participants. Another reason why open-ended questions are only rarely used is that they also put a burden on the researchers analyzing the data. Analyzing open-ended questions requires the following steps: (1) development of a categorization scheme, (2) coder training, (3) coding, and (4) testing of reliability. Coding can be done either manually or (semi-)automatically. We will move on to these coding options in the next section.

2.2. Manual and semi-automated coding

When textual responses to open-ended questions need to be categorized, researchers have two options: manual coding and automatic coding. Manual coding means that a human coder decides which class to assign an answer to, while automatic coding relies upon statistical learning models that assign substantial, a-priory defined categories to textual responses. Manual coding is expensive and time-consuming since it requires human coders; ideally, at least two persons independently code in order to assess inter-coder reliability (Leiva et al., 2006; Schonlau, 2015). Therefore, automatic or semi-automatic coding is an attractive option for large data sets. Completely automatic coding will not be discussed here; the performance quality is, in general, not good enough for the categorization of short texts such as open-ended questions in surveys (Jónsson and Stolee, 2015). However, a subset of the data, called training data, is coded by human coders in semi-automatic coding. A statistical learning model is then trained on this subset. This model is then used to predict the class of uncategorized text responses. Therefore, a disadvantage of semi-automatic coding is the need for expertise to perform the modeling and prediction, but this is offset by lower cost and faster execution (once the modeling is complete). Plus, several applications show that semi-automatic encoding utilizing machine learning algorithms can effectively code and classify different kinds of text data. For instance, Grimmer and Stewart (2013) compared different ways of coding political texts

automatically and discussed the advantages and disadvantages of this approach; Gentzkow et al. (2019) did the same for economic data. Open-ended questions in surveys are also just text data, and machine learning algorithms have been used to classify these answers. Kern et al. (2019) give an overview of these applications of statistical learning methods in survey research in general and also discusses open-ended questions (Joachims, 2001; Schonlau and Couper, 2016). Joachims (2001) used a support vector machine (SVM) for the classification of open-ended answers and achieved good performance. Schonlau and Couper (2016) developed a semi-automatic approach where answers to open-ended questions are classified automatically by multinomial gradient boosting when the probability of correct classification is high and manually by a human otherwise. A paper by He and Schonlau (2020) explored using double coded data for classification.

Another practical example of these (semi-)automatic methods for open-ended questions is occupation coding. It refers to the coding of text responses to an open-ended question about the respondent's profession. For example Gweon et al. (2017) proposed three automatic coding algorithms and improved coding accuracy for occupation coding, and Schierholz (2019) compared statistical learning algorithms in occupation coding.

Another example of text data in surveys is responses to exploratory questions, e.g., web probing. Exploratory questions are follow-up questions that ask respondents to provide additional information about a survey item (Beatty and Willis, 2007; Meitinger et al., 2018).

In the following, we introduce our exemplary application area: research on respondents' motivation to participate (or not) in surveys, exemplified using the GESIS Panel. Then, we will explain the central role that survey motivation plays in survey methodological research and give a short overview of previous analyses with data from open-ended questions on survey motivation.

3. Collecting and coding of data on survey motivation

3.1. Survey motivation

A key concern for panel infrastructures such as the GESIS Panel is maintaining their group of panelists and motivating them to participate in survey waves repeatedly. Even if initial recruitment was successful, throughout multiple panel waves, panel attrition might decrease the number of respondents (Hill and Willis, 2001; Behr et al., 2005; Lynn, 2018), leading to nonresponse bias and variance inflation. Theoretical and empirical research on response behavior has thus been an integral part of survey research for the last decades (Keusch, 2015). Apart from societal level factors such as survey fatigue and attributes of the survey design, respondents' personality traits, topic interest, attitudes toward survey research, and previous participation behavior are examined as a possible influence on response behaviors (Keusch, 2015). Survey motivation is an intermediate step between these external/internal factors and the response behavior. For example, in the frame of the leverage-salience theory (Groves et al., 2000), different survey attributes can have very different effects among possible respondents. The achieved influence of a particular feature is a "function of how important it is to the potential respondent, whether its influence is positive or negative, and how salient it becomes to the sample person during the presentation of the survey request" (Groves et al., 2000, p.301). Although this and other theories (Singer, 2011) of survey participation establish a direct link between survey motivation and survey participation, we know surprisingly little about how people describe their participation motivation when not prompted with pre-defined categories. A short overview of previous research on survey motivation building upon open-ended questions and accompanying classification of survey motivation is given in the next section.

3.2. Overview over existing classification schemes

Two studies by Porst and Briel (1995) and Singer (2003) looked at participation motivation in surveys; both use very similar classification schemes to categorize answers to the open-ended question. Singer (2003) included vignettes in a monthly RDD survey and asked respondents how willing they would be to participate in the described survey, and a second open-ended question: "Why would (or Why wouldn't) you be willing to participate in the survey described?." She then divided the reasons given into three broad categories—altruistic, egoistic, and characteristics of the survey. The author explains that the alleged overlap between survey characteristics and the other two categories is resolved through the respondents' emphasis on themselves and their altruistic motive or the survey characteristics. These categories closely resemble those developed by Porst and Briel (1995) for German panelists. They differentiate three broad categories: altruistic, survey-related, and personal, and develop a classification with finer details ranging from four to seven sub-categories. For example altruistic reasons are divided into the motive "surveys important/meaningful for politics, society, economy, science," "surveys important/meaningful for ZUMA" (ZUMA was the Centre for Survey Research and Methodology, now part of GESIS), "surveys important/meaningful (without specification)," "social responsibility." We will employ a similar classification scheme in the analysis of the open-ended question

on panel motivation in the GESIS Panel, which we will present next.

4. Survey motivation in the GESIS Panel

We use data from the GESIS Panel (Bosnjak et al., 2018), a probability-based mixed-mode panel that has been in operation since 2014. In order to compensate for panelist dropout, there have been two refreshment samples, so that the panel in October 2020 consisted of three cohorts and a total of more than 5,000 respondents (GESIS, 2021). Bi-monthly, panelists are invited to respond to a survey that lasts approximately 20 min. About 75% of respondents answer in web mode and 25% in mail mode. With each survey invitation, they receive a prepaid incentive of five euros.

4.1. The question on survey motivation

Until 2020, the GESIS Panel had six waves per year, and a question on survey motivation was included in every sixth and last wave of a year. Figure 1 shows the design and wording of the survey question. The panelists are asked for what reasons they participate in the GESIS GesellschaftsMonitor surveys. GESIS GesellschaftsMonitor is the name with which the GESIS Panel presents itself to its participants. The panelists are then prompted to give their most important reason, second most important reason, and third most important reason in three separate lines. This questionnaire design has major advantages in coding, since it leads to unidimensional answers with very few exceptions ($< 1\%$).

4.2. The semi-automatic categorization of an open-ended question on survey motivation

4.2.1. Manual coding

We first developed our coding scheme before beginning with the manual and semi-automated coding. This was done iteratively by a team of two authors; we started with a coding scheme from Porst and Briel (1995) but adapted it to the GESIS Panel survey, adding categories that were more specific to the GESIS Panel survey and were often mentioned. We aimed for at least 40 observations for each category, collapsing categories that did meet that criterion. We aimed toward categories that were as distinct as possible from other categories. For some of the multivariate analyses, we collapsed the finer categories into broader categories due to sample size. The final list of categories is available in Table 1 and the entire coding scheme (in English and German) in the Appendix. The assignment of finer to broader categories is given in Figure 2. We will give a brief overview of the categories. The first group of possible categories refers to answers that express interest or curiosity in the survey topic. Some respondents indicate that they learn from the survey about current topics or that they learn something about themselves when they reflect and answer the questions ("Learning"). Other respondents participate because they want to share their opinion ("Tell opinion"); some even want to influence policy or research through their participation ("Influence"). Since the GESIS Panel has an incentive of 5 Euro, many people mention it ("Incentive"). Respondents often mention that they enjoy taking the survey ("Fun") or that they participate because it has become part of their routine ("Routine"). This is, of course, an answer that one would not find in one-time surveys. Some persons feel obliged to participate out of a sense of duty ("Dutifulness"). A lot of people want to help through their participation and they often, but not always, specify the addressees of their assistance: researchers, politicians or even the society in general ("Help science," "Help politicians" "Help society" and "Help in general"). There were also some survey-related reasons to participate that some respondents mentioned: the brevity, the anonymity, and the professionalism of the GESIS Panel in particular ("Brevity," "Anonymity," "Professionalism," and "Other survey characteristics"). More people than we expected from previous research mentioned their recruitment or even specific traits of their recruiting interviewer for the GESIS Panel. Therefore, we added these categories ("Recruiter" and "Recruitment"). Many respondents simply mentioned that participation in general or the survey are important; this is a comprehensive and common class ("Importance in general"). Some persons cannot think of a reason or give other answers that are very rare, e.g., "I have been pushed to participate by my parents". These responses are summarized in a residual class ("No reason/Other"). Examples from the survey for each class are available in the Appendix as part of the Coding Scheme.

A random subset of the data (n = 5,000), about one-fifth of the data, was manually coded by one of the co-authors and a student assistant independently after two co-authors settled on the final coding scheme. While manually coding 5,000 answers might seem like a very high number in comparison to other studies (Schonlau and Couper, 2016 coded around 500 answers), one should, however, keep in mind that we also used 21 categories. A sufficient number of observations is required for each of these categories. As in many other cases, intercoder disagreement occurred repeatedly (Popping and Roberts, 2009; Schonlau, 2015). We used Cohen's κ-coefficient to calculate the measure of agreement between coders (Fleiss et al., 2003). Cohen's κ-coefficient was high, around 0.91. The remaining disagreements can be resolved in several ways, such as: (1) the two coders discuss the disagreement and reach consensus (D'Orazio et al., 2016), (2) a third person (an expert) with more experience determines the code, or (3) a third coder is used. The

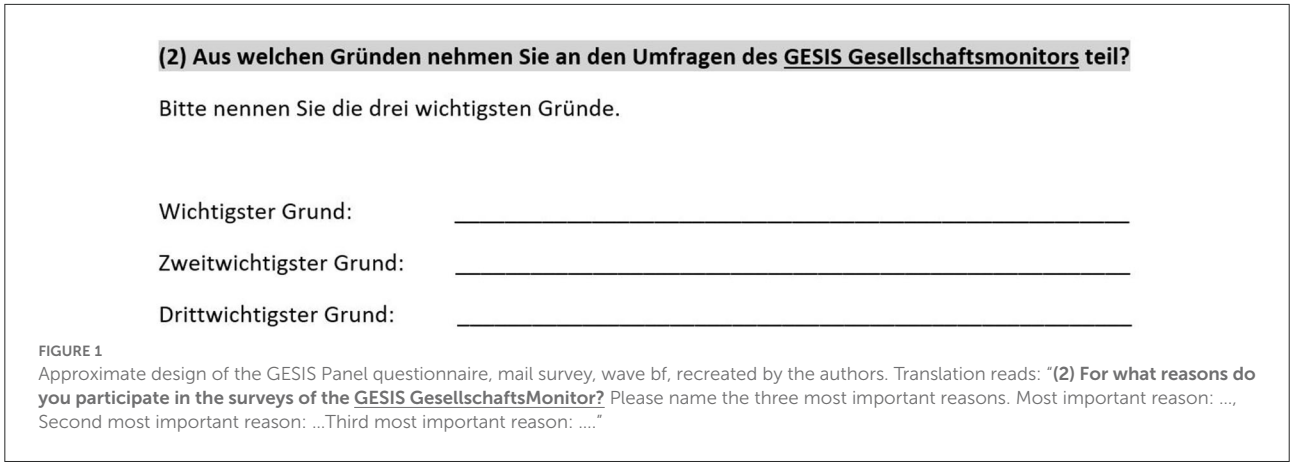


FIGURE 1
Approximate design of the GESIS Panel questionnaire, mail survey, wave bf, recreated by the authors. Translation reads: "**(2) For what reasons do you participate in the surveys of the GESIS GesellschaftsMonitor?** Please name the three most important reasons. Most important reason: ..., Second most important reason: ...Third most important reason: ...."

TABLE 1 List of categories for respondents' motivation to participate in the GESIS Panel.

Categories		
1. Interest	10. Help science	19. Other survey characteristics
2. Curiosity	11. Help politicians	20. Importance in general
3. Learning	12. Help society	21. No reason/Other
4. Tell opinion	13. Help	
5. Influence	14. Brevity	
6. Incentive	15. Anonymity	
7. Fun	16. Professionalism	
8. Routine	17. Recruiter	
9. Dutifulness	18. Recruitment	

A more detailed coding scheme with examples can be found in the Appendix.

third coder can then break the tie between the first two coders. In our case, we resolved disagreements by the second option; the expert was a more senior member of the team of authors.

4.2.2. Pre-processing

Processing and cleaning text data for semi-automated classification can require varying amounts of efforts and techniques, however, a set of typically used techniques has already been established: this set includes spellchecking (Quillo-Espino et al., 2018), lowercasing (Foster et al., 2020), stemming (Jivani, 2011; Bao et al., 2014; Singh and Gupta, 2017), lemmatization (Bao et al., 2014; Banks et al., 2018; Symeonidis et al., 2018; Foster et al., 2020), stopword removal (Foster et al., 2020), and different ways of text enrichment/adding of linguistic features (Foster et al., 2020). We will systematically review these options below and justify our choices (for an overview, see Table 2).

First, we perform an automated spelling correction through the hunspell R package (Ooms, 2020) using the dictionary "DEde2," i.e., words that are not part of this German dictionary are replaced with a word that is as similar as possible in spelling to the unknown word. Then, after a first check of the performance of the correction, we expanded the dictionary with words that were incorrectly improved because they were not part of the dictionary but very frequent in our corpus (e.g., "GESIS"). After adding these, two of the authors used a random sample ($n = 100$) to re-check performance and concluded that only about 6% of the improved words were changed so that they no longer corresponded to the meaning that the respondent had intended.

For subsequent pre-processing steps, we used the popular quanteda R package (Benoit et al., 2018), which offers a multitude of text functions like lemmatization and stemming, trimming, upper- and lowercasing. It also allows transforming text snippets into tokens such as uni- and multigrams and also allows functions for computing text statistics, fitting models, and producing visualizations from a text corpus. Text functions such as stemming, lemmatization or even lowercasing reduces the size of the text matrix, making it more dense. Stemming for example is especially desirable for languages where a single stem can generate dozens of words in case of verbs (e.g., French or Turkish). Similarly, lemmatization groups together inflected forms together as a single base form. However, in some cases,

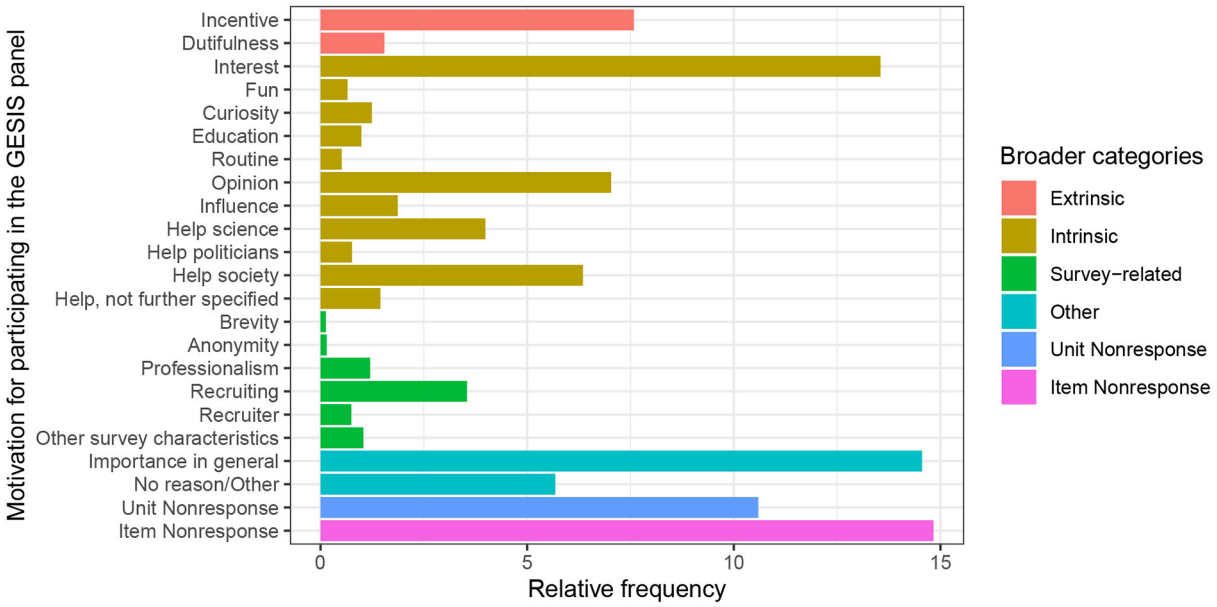


FIGURE 2
Most important reason given by panelists in wave "bf" (2014) of the GESIS Panel. Semi-automated classification of reasons described in Section 4.2.3.

the actual word might make a difference, compared to its stem. In our case, differences between, e.g., "I am" and "I was" may carry important information, since referring to the past was often done when referring to the recruiter or the recruiting experience compared to the present used for expressing interest in the questions. We tested both lemmatization and stemming as well as stopword removal, but concluded that not doing either of the steps increased performance. We did remove hyphens, separators and punctuations.

4.2.3. Semi-automated coding: SVM and other options

Semi-automatic text classification is possible through statistical learning. Many statistical learning algorithms are now available in statistical software like R and Python, and it is not possible to give a complete overview here (see e.g., Hao and Ho, 2019, for a Python overview). However, we do want to point to some of the most popular choices that have been applied to classifying answers to open-ended questions: these include tree-based methods like random forests and boosting (Schonlau and Couper, 2016; Kern et al., 2019; Schierholz and Schonlau, 2021), support vector machines (SVM) (Joachims, 2001; Bullington et al., 2007; He and Schonlau, 2020, 2021; Khanday et al., 2021), multinomial regression (Schierholz and Schonlau, 2021) and naïve Bayes classifiers (Severin et al., 2017; Paudel et al., 2018).

From the multitude of possibilities, we explored *via* 5-fold cross-validation with 70% of the hand-classified data the different algorithmic options in the R package `LiblineaR` (Helleputte, 2021) based on the C/C++ library 'LIBLINEAR' (i.e., L2-regularized L2-loss SVM, L2-regularized L1-loss SVM, SVM by Crammer and Singer, L1-regularized L2-loss SVM, L1-regularized and L2-regularized logistic regression). Liblinear is able to handle large-scaled data sets, especially for text classification, i.e., data sets where some features are scarce (Fan et al., 2008; Helleputte, 2021). We also explored different cost parameter values (0.1, 1, and 10) combined with these algorithms. We also explored random forests and naïve Bayes classifiers, but they did not yield higher performance rates. After examining the cross-validation (5-fold) results, we chose to use an L1-regularized L2-loss SVM (Liblinear type 5) with cost parameter 10. We retrained the model with this parameter setup and the combined training and test set (70% of the hand-classified data). The accuracy rate for the validation set (30% of the hand-classified data) was 0.93 [0.91, 0.94], and the unweighted median macro F1 measure over all categories was 0.83. We then continued to automatically classify the complete dataset, including the $\sim$ 20.000 answers not classified by hand. The categories with the weakest performance (F1 measure around 0.6) were Learning, Help in general, and Recruiter. In general, smaller categories or categories that are close to others (e.g., "Help in general" and "Help politicians," "Help society") were more difficult to categorize for our trained model. Here, one could also think about collapsing different categories for better performance measures, but also at the cost of being less specific and losing information. Therefore, we decided against this step for our analysis. Larger (and easy to catch) categories such as

TABLE 2 Steps in semi-automated coding, choice of methods and alternatives.

Step	Options
Manual coding of test set	
Sampling	**Random sampling**, random sampling with min. sample numbers for each class, iterative procedure with coding of observations with low predictive certainty
Number of coders	1, **2**, 3, ... + **additional expert coders to resolve differences in coding**.
Resolving differences in coding	(1) Reaching consensus between original coders (2) **expert decides** (3) majority vote by third coder
Pre-processing of text data	
Spellchecking	**Yes**/no
Lowercasing	Yes/**no**
Stemming or	Yes/**no**
lemmatization	Yes/**no**
Stopword removal	Yes/**no**
Tokenization	**Unigrams**, Bigrams, Trigrams ...
Inclusion of word/sentence embeddings	**Yes**/no
Inclusion of non-text data	Yes/**no**
Semi-automated categorization	
Statistical learning algorithm	Tree-based methods (e.g., boosting or random forests), **support vector machine (SVM)**, multinomial regression, Naive Bayes classifier
Additional human coding for observations with low predictive probability	Yes/**no**
Checking/Validation	
Evaluation parameters	**Accuracy. Precision, Recall, F1, Detection Rate, Detection Prevalence, Balanced Accuracy (either macro or micro). Confusion matrix.**

Our decisions for our example are in bold font.

Interest and Incentive have micro F1 measures of > 0.99 in our classification.

4.3. Analyses

4.3.1. Univariate analysis of respondents' motivation

We are now moving on to further analysis steps after the semi-automated classification of the not hand-coded observations. First, we are interested in the empirical distribution of reasons given by panelists of the GESIS Panel for their participation. This is important information, especially for panel management and maintenance, as the knowledge of participants' motivation can, for example, be used in further waves for adaptive design measures. In Figure 2, the relative frequencies of the most important reason to participate in the panel, which have been semi-automatically classified, are shown for the panelists of wave "bf" (2014) of the GESIS Panel. As can be seen in Figure 2, the different categories have very unequal relative frequencies. The biggest substantive categories are "Incentive," "Interest" and "Importance in general." Many panelists openly state that they are participating because of the 5 Euro incentive given in each wave for participation. Over 10% participate because they are interested in the topics covered in the survey. A substantive part of respondents also state that they participate because they perceive their responses as important, but do not further elaborate on why they think so or for whom they think their participation is important. All in all, around 10% of the panelists state that they want to help, and usually, they also indicate the recipient of their help that they have in mind: science, politicians, or, simply, society in general. Other categories have been mentioned by just a few respondents, e.g., that it is part of their routine or that the person recruiting them was nice. In Figure 2, the substantive amount of item nonresponse (almost 15%) and general unit nonresponse (around 10%) is also depicted. Results for other waves are not shown here, but do not differ much.

4.3.2. Multivariate analysis: Survey motivation and panel attrition

Secondly, we will use the semi-automatically coded motivations for panel participation as a predictor variable in panel attrition analysis. As mentioned earlier, previous

research (Brüggen et al., 2011) on participant motivation also often distinguishes between intrinsic and extrinsic motivation. An extrinsically motivated person would participate with the prospect of incentives or moral obligation. In contrast, an intrinsically motivated person would participate motivated by pleasure, curiosity, or interest, or they want to help and reveal their opinion. In the study by Brüggen et al. (2011), it was found that those with intrinsic motivations had the highest response rate, and those with extrinsic and self-focused motivation (incentives) had the lowest. According to this result, dropout should generally be higher among extrinsically motivated individuals, which is also consistent with the findings of Porst and Briel (1995). However, our goal is not to formally test these theories but rather to motivate our empirical demonstration.

We take a look at the difference between the correlation of extrinsic and intrinsic motivation and panel dropout in the GESIS Panel between the years 2013 and 2018. We are, therefore, interested in the estimates of the model parameters for extrinsic and intrinsic in a logistic regression model with panel dropout as the dependent variable. This research question can be analyzed utilizing a discrete event-history model. A panelist of the GESIS Panel can drop out in two ways: either by requesting to be excluded from the panel management or by not participating in three subsequent waves of the panel. The data set used to estimate the model parameters consists of 6.031 panelists with 14.635 points of observation (2.4 years of being part of the panel on average). The data set is in person-period format, i.e., one row per person per period observed. The panel dropout indicator is set to 1 for the year in which the panelist dropped out of the panel (rows with dropout: 1,196, rows without dropout: 13,439). For periods thereafter, the panelist is not part of the sample anymore. We group the participation motivation categories that we presented earlier into broader categories, since the original categories include very small ones, potentially creating computational problems. We group them into the broader categories "Extrinsic reasons," "Intrinsic reasons," "Survey-related reasons," and "Other reasons." These broader categories were already indicated through the colors used in the barplot in Figure 2. Apart from the independent variable we are interested in, we also included other sociodemographic variables (gender, age, and education) and wave indicators as intercepts in our analysis. We present average marginal effects for easier interpretation than log odds ratios (Mood, 2010), the average marginal effect is the average change in probability when the independent variable increases by one unit. In Figure 3, we can see that persons who indicate either extrinsic or intrinsic reasons are much less likely to drop out in the year after than persons who do not answer the question. The difference between both, however, is small; we are therefore not able to replicate the findings by Porst and Briel (1995) and Brüggen et al. (2011).

5. Conclusion

With this article, we have made several contributions. First, we developed and presented a precise categorization scheme for reasons to participate in a panel survey, the GESIS Panel. We built on existing studies by Porst and Briel (1995) and Brüggen et al. (2011), but extended them significantly and tailored them to suit the requirements of panel studies better. While certain subcategories will probably differ in their magnitude for different surveys, our coding scheme can serve as a starting point for other survey researcher, just like Porst and Briel (1995) did serve as our—albeit much broader—starting point.

Second, we have also demonstrated that semi-automatic classification is a suitable tool to classify large sets of responses to open-ended questions in surveys. A potential application area of semi-automated classification are potentially other panel surveys with repeated questions or survey with response numbers $> 10,000$. Below this number, we estimate that the effort used to train and calibrate the semi-automated classification would—depending on the previous experience of researchers, of course—not be less than hand-coding the entire dataset. One also needs to keep in mind, that a certain number (at least 40 in our experience) of observations are needed for each category, a number that can be hard to achieve with small training sets or uneven distributions of categories. Some characteristics of the questionnaire design and the answers worked in our favor for semi-automatic classification: The answers were generally one-dimensional and short and concise by presenting the question with different fields for different reasons. Difficulties can, however, be encountered with answers that are too short and therefore ambiguous. This is not a problem that only occurs when using semi-automated classification, this is also a problem when using only hand-coding. As an example, we noticed several times that persons simply answered "Opinion" as a reason, it is however unclear to humans and machines alike whether the persons like to give their opinion or whether the person hopes that their opinion has an impact. Another limitation is also that a few categories (especially rare ones) were harder to predict than others, this might bias our analysis in a small way. To better handle spelling errors, we automatically corrected the answers with Hunspell before further pre-processing steps. We tested the use of sentence embeddings (Conneau et al., 2018) in addition to word counts, but in our case, we did not see any significant additional improvement in performance. Overall, however, the performance can be rated as very good.

Third, after semi-automatic classification, we used the newly generated measurements for descriptive analyses. Apart from a widely-shared sense among respondents that surveys are important, incentives and interest in the topics of the GESIS Panel are some of the most important motivators. Factors that have not been discussed prominently in literature are the wishes of many participants to help politicians and those in power

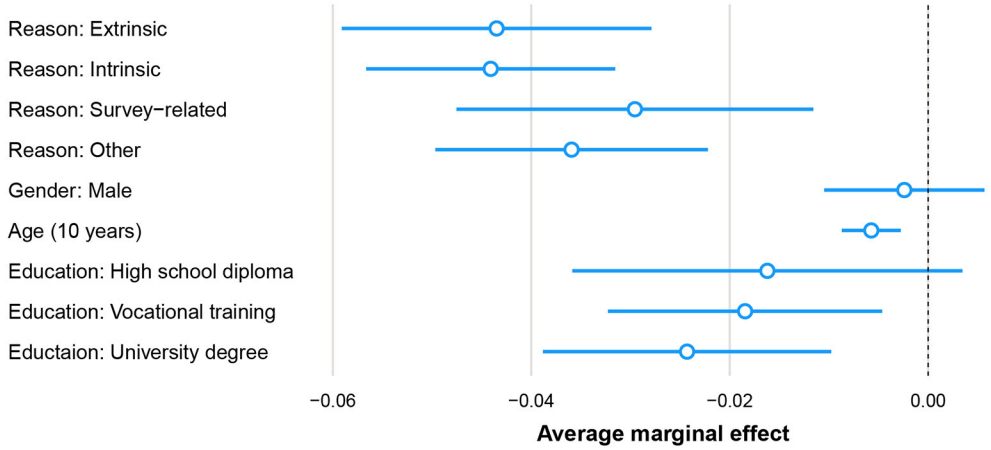


FIGURE 3
Logistic regression of panel dropout on independent variable most important participation reason, categorized in four broader categories "Extrinsic reasons," "Intrinsic reasons," and "Survey-related reasons," and "Other reasons." Reference category for reason: No reason given, reference category for gender: female, reference category for education: no formal education diploma. The AME for unit nonresponse is 0.4140 (SE 0.0183), not depicted since it is outside of the x-axis scale.

understand what people think. It is also apparent that people enjoy sharing their opinion.

Fourth, we were also able to use the newly generated measurement with categories of participation in a simple analysis of (partial) correlations between survey motivation and panel attrition. While a full causal analysis would go beyond the scope of this article, we did notice that intrinsic or extrinsic motivations were clearly associated with less panel dropout than survey-related reasons (or item nonresponse and unit nonresponse). Again, this result can be seen as a good indicator of criterion validity. Other than Brüggen et al. (2011), however, we did not see any noticeable difference between intrinsic and extrinsic motivations regarding the association with panel dropout.

Several other research paths lead from here: regarding panel management, it may be worth tailoring cover letters to potential respondents to their motivations to increase participation rates further and lessen dropout. In addition, a thorough causal analysis would be important to examine the influence of motivation on the willingness to participate in surveys in more detail.

Another open question is how to enable further use of coded text responses or trained models. It would be helpful to have more comprehensive comparisons and explorations of general advantages and disadvantages of different semi-automated classification methods and algorithms. Another issue is the fact that answers to open-ended questions are not generally part of scientific use files since they can contain personal information, which would potentially allow the identification of respondents. A possible solution might be strategies that have been employed in other cases where data confidentiality has to be guaranteed: the creation of synthetic data sets (Drechsler, 2011) from the original data feature matrix. This could allow other researchers to train models and use these on their data, while at the same time minimizing disclosure risks.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: We used data from the GESIS Panel (extended edition). The dataset can be accessed at the GESIS Leibniz Institute for the Social Sciences. https://search.gesis.org/research_data/ZA5664. Requests to access these datasets should be directed to https://search.gesis.org/research_data/ZA5664.

Author contributions

A-CH and BW contributed to conception and design of the study. A-CH and KB hand-coded a subset of the text data together with student assistants. A-CH wrote the draft of the analysis script and wrote the draft of the manuscript. A-CH, PS, PC, and KB extended and performed the statistical analysis. BW contributed in-depth feedback. All authors contributed to manuscript revision and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdata.2022.880554/full#supplementary-material

References

Banks, G. C., Woznyj, H. M., Wesslen, R. S., and Ross, R. L. (2018). A review of best practice recommendations for text analysis in r (and a user-friendly app). *J. Bus. Psychol.* 33, 445–459. doi: 10.1007/s10869-017-9528-3

Bao, Y., Quan, C., Wang, L., and Ren, F. (2014). "The role of pre-processing in twitter sentiment analysis," in *International Conference on Intelligent Computing* (Taiyuan: Springer), 615–624.

Beatty, P. C., and Willis, G. B. (2007). Research synthesis: the practice of cognitive interviewing. *Public Opin Q.* 71, 287–311. doi: 10.1093/poq/nfm006

Behr, A., Bellgardt, E., and Rendtel, U. (2005). Extent and determinants of panel attrition in the european community household panel. *Eur. Sociol. Rev.* 21, 489–512. doi: 10.1093/esr/jci037

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., et al. (2018). quanteda: an r package for the quantitative analysis of textual data. *J. Open Source Softw.* 3, 774. doi: 10.21105/joss.00774

Bosnjak, M., Dannwolf, T., Enderle, T., Schaurer, I., Struminskaya, B., Tanner, A., et al. (2018). Establishing an open probability-based mixed-mode panel of the general population in Germany: the GESIS Panel. *Soc. Sci. Comput. Rev.* 36, 103–115. doi: 10.1177/0894439317697949

Brüggen, E., Wetzels, M., De Ruyter, K., and Schillewaert, N. (2011). Individual differences in motivation to participate in online panels: the effect on reponse rate and reponse quality perceptions. *Int. J. Market Res.* 53, 369–390. doi: 10.2501/IJMR-53-3-369-390

Bullington, J., Endres, I., and Rahman, M. (2007). "Open ended question classification using support vector machines," in *MAICS 2007* Chicago, IL.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. (2018). What you can cram into a single vector: probing sentence embeddings for linguistic properties. *arXiv [Preprint]*.

Dillman, D. A., Smyth, J. D., and Christian, L. M. (2009). *Internet, Mail, and Mixed-Mode Surveys. The Tailored Design Method.* Hoboken, NJ: Wiley.

D'Orazio, V., Kenwick, M., Lane, M., Palmer, G., and Reitter, D. (2016). Crowdsourcing the measurement of interstate conflict. *PLoS ONE* 11, e0156527. doi: 10.1371/journal.pone.0156527

Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control.* New York, NY: Springer.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). Liblinear: a library for large linear classification. *J. Mach. Learn. Res.* 9, 1871–1874.

Fleiss, J. L., Levin, B., and Paik, M. C. (2003). *Statistical Methods for Rates and Proportions.* Hoboken; New Jersey, NY: John Wiley & Sons, Inc.

Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., and Lane, J. (2020). *Big Data and Social Science-Data Science Methods and Tools for Research and Practice.* Boca Raton, FL: CRC Press.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *J. Econ. Lit.* 57, 535–574. doi: 10.1257/jel.20181020

GESIS (2021). *Gesis Panel-Standard Edition. GESIS Datenarchiv, Köln.* ZA5665 Datenfile Version 41.0.0.

Grimmer, J., and Stewart, B. M. (2013). Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Polit. Anal.* 21, 267–297. doi: 10.1093/pan/mps028

Groves, R. M., Singer, E., and Corning, A. (2000). Leverage-saliency theory of survey participation: description and an illustration. *Public Opin. Q.* 64, 299–308. doi: 10.1086/317990

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M., and Steiner, M. (2017). Three methods for occupation coding based on statistical learning. *J. Off. Stat.* 33, 101–122. doi: 10.1515/jos-2017-0006

Hao, J., and Ho, T. K. (2019). Machine learning made easy: a review of scikit-learn package in python programming language. *J. Educ. Behav. Stat.* 44, 348–361. doi: 10.3102/1076998619832248

He, Z., and Schonlau, M. (2020). Automatic coding of open-ended questions into multiple classes: whether and how to use double coded data. *Survey Res. Methods* 14, 267–287. doi: 10.18148/srm/2020.v14i3.7639

He, Z., and Schonlau, M. (2021). A model-assisted approach for finding coding errors in manual coding of open-ended questions. *J. Survey Stat. Methodol.* 10, 365–376. doi: 10.1093/jssam/smab022

Helleputte, T. (2021). *LiblineaR: Linear Predictive Models Based on the LIBLINEAR C/C++ Library.* R package version 2.10–12.

Hill, D. H., and Willis, R. J. (2001). Reducing panel attrition: a search for effective policy instruments. *J. Hum. Resour.* 36, 416–438. doi: 10.2307/3069625

Jivani, A. G. (2011). A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl.* 2, 1930–1938.

Joachims, T. (2001). "A statistical learning model of text classification for support vector machines," in *SIGIR Forum (ACM Special Interest Group on Information Retrieval) 24* New Orleans, LA.

Jónsson, E., and Stolee, J. (2015). "An evaluation of topic modelling techniques for twitter," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)* (Beijing), 489–494.

Kern, C., Klausch, T., and Kreuter, F. (2019). Tree-based machine learning methods for survey research. *Survey Res. Methods* 13, 73–93. doi: 10.18148/srm/2019.v13i1.7395

Keusch, F. (2015). Why do people participate in web surveys? applying survey participation theory to internet survey data collection. *Manag. Rev. Q.* 65, 183–216. doi: 10.1007/s11301-014-0111-y

Khanday, A. M. U. D., Khan, Q. R., and Rabani, S. T. (2021). "Svmbpi: support vector machine-based propaganda identification," in *Cognitive Informatics and Soft Computing* (Singapore: Springer), 445–455.

Krosnick, J. A., and Presser, S. (2010). *Question and Questionnaire Design, Vol. 2.* Bingley: Emerald Group Publishing Limited.

Leiva, F. M., Ríos, F. J. M., and Martínez, T. L. (2006). Assessment of interjudge reliability in the open-ended questions coding process. *Quality Quant.* 40, 519–537. doi: 10.1007/s11135-005-1093-6

Lynn, P. (2018). "Tackling panel attrition," in *The Palgrave Handbook of Survey Research* (Cham: Springer), 143–153.

Meitinger, K., Braun, M., and Behr, D. (2018). Sequence matters in online probing: the impact of the order of probes on response quality, motivation of respondents, and answer content. *Survey Res. Methods* 12, 103–120. doi: 10.18148/srm/2018.v12i2.7219

Mood, C. (2010). Logistic regression: why we cannot do what we think we can do, and what we can do about it. *Eur. Sociol. Rev*. 26, 67–82. doi: 10.1093/esr/jcp006

Ooms, J. (2020). *hunspell: High-Performance Stemmer, Tokenizer, and Spell Checker*. R package version 3.0.1.

Paudel, S., Prasad, P. W. C., Alsadoon, A., Islam, M. R., and Elchouemi, A. (2018). "Feature selection approach for twitter sentiment analysis and text classification based on chi-square and naïve bayes," in *ATCI 2018: International Conference on Applications and Techniques in Cyber Security and Intelligence ATCI 2018* (China: Springer International Publishing), 281–298.

Popping, R., and Roberts, C. W. (2009). Coding issues in modality analysis. *Field Methods* 21, 244–264. doi: 10.1177/1525822X09333433

Porst, R., and Briel, C. V. (1995). "Wären Sie vielleicht bereit, sich gegebenenfalls noch einmal befragen zu lassen?" in *Oder: Gründe für die Teilnahme an Panelbefragungen, Vol. 1995/04 of ZUMA-Arbeitsbericht* (Mannheim: Zentrum für Umfragen, Methoden und Analysen -ZUMA).

Quillo-Espino, J., Romero-González, R. M., and Lara-Guevara, A. (2018). Advantages of using a spell checker in text mining pre-processes. *J. Comput. Commun*. 6, 43–54. doi: 10.4236/jcc.2018.611004

Schierholz, M. (2019). *New Methods for Job and Occupation Classification* (Ph.D. Thesis). University of Mannheim Mannheim.

Schierholz, M., and Schonlau, M. (2021). Machine learning for occupation coding–a comparison study. *J. Survey Stat. Methodol*. 9, 1013–1034. doi: 10.1093/jssam/smaa023

Schonlau, M. (2015). "What do web survey panel respondents answer when asked "do you have any other comment?"," in *Survey Methods: Insights From the Field*.

Schonlau, M., and Couper, M. P. (2016). Semi-automated categorization of open-ended questions. *Surv. Res. Methods* 10, 143–152. doi: 10.18148/srm/2016.v10i2.6213

Schonlau, M., Guenther, N., and Sucholutsky, I. (2017). Text mining with n-gram variables. *Stata J*. 17, 866–881. doi: 10.1177/1536867X1801700406

Severin, K., Gokhale, S. S., and Konduri, K. C. (2017). "Automated quantitative analysis of open-ended survey responses for transportation planning," in *2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing and Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)* (San Francisco, CA: IEEE), 1–7.

Singer, E. (2003). Exploring the meaning of consent: participation in research and beliefs about risks and benefits. *J. Off. Stat*. 19, 273.

Singer, E. (2011). Toward a benefit-cost theory or survey participation: evidence, further tests, and implications. *J. Off. Stat*. 27, 379–392.

Singh, J., and Gupta, V. (2017). A systematic review of text stemming techniques. *Artif. Intell. Rev*. 48, 157–217. doi: 10.1007/s10462-016-9498-2

Symeonidis, S., Effrosynidis, D., and Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert. Syst. Appl*. 110, 298–310. doi: 10.1016/j.eswa.2018.06.022

Züll, C., and Menold, N. (2019). *Offene Fragen*. Wiesbaden: Springer Fachmedien Wiesbaden.

frontiers | Frontiers in Big Data

TYPE Original Research
PUBLISHED 14 September 2022
DOI 10.3389/fdata.2022.908636

OPEN ACCESS

EDITED BY
Dimitri Prandner,
Johannes Kepler University of Linz, Austria

REVIEWED BY
Heinz Leitgöb,
Catholic University of Eichstätt-Ingolstadt, Germany
Robert Moosbrugger,
Johannes Kepler University of Linz, Austria

*CORRESPONDENCE
Roland Verwiebe
verwiebe@uni-potsdam.de

SPECIALTY SECTION
This article was submitted to Data Science, a section of the journal Frontiers in Big Data

RECEIVED 30 March 2022
ACCEPTED 16 August 2022
PUBLISHED 14 September 2022

CITATION
Seewann L, Verwiebe R, Buder C and Fritsch N-S (2022) "Broadcast your gender." A comparison of four text-based classification methods of German YouTube channels. *Front. Big Data* 5:908636. doi: 10.3389/fdata.2022.908636

"Broadcast your gender." A comparison of four text-based classification methods of German YouTube channels

Lena Seewann, Roland Verwiebe*, Claudia Buder and Nina-Sophie Fritsch

Faculty of Economics and Social Sciences, University of Potsdam, Potsdam, Germany

Social media platforms provide a large array of behavioral data relevant to social scientific research. However, key information such as sociodemographic characteristics of agents are often missing. This paper aims to compare four methods of classifying social attributes from text. Specifically, we are interested in estimating the gender of German social media creators. By using the example of a random sample of 200 YouTube channels, we compare several classification methods, namely (1) a survey among university staff, (2) a name dictionary method with the World Gender Name Dictionary as a reference list, (3) an algorithmic approach using the website gender-api.com, and (4) a Multinomial Naïve Bayes (MNB) machine learning technique. These different methods identify gender attributes based on YouTube channel names and descriptions in German but are adaptable to other languages. Our contribution will evaluate the share of identifiable channels, accuracy and meaningfulness of classification, as well as limits and benefits of each approach. We aim to address methodological challenges connected to classifying gender attributes for YouTube channels as well as related to reinforcing stereotypes and ethical implications.

KEYWORDS

text based classification methods, gender, YouTube, machine learning, authorship attribution

Introduction

Every day, thousands of people around the world share their homes, thoughts, and activities on social media platforms such as YouTube. This online self-representation provides an extensive and accessible resource for research in various disciplines, focusing on different aspects of YouTube as a platform. Among other things, YouTube is discussed as a cultural phenomenon (Boxman-Shabtai, 2018; Burgess and Green, 2018). Especially among younger age groups, the more than 30 million YouTube channels worldwide have become a primary source of social, cultural, and political information, whose relevance is significantly higher than that of traditional media formats such as newspapers and TV (Mitchell et al., 2018; Litvinenko, 2021). Other authors study the functions of the platform algorithm and examine the impact it has

on consumers and producers (Rieder et al., 2018; Bishop, 2020; Bryant, 2020). Most of the existing studies deal with various aspects of the presence and activity of YouTubers within the platform. The range of topics is wide and includes economic (mis)success (Postigo, 2016; Soha and McDowell, 2016; Duffy, 2020), political activism among YouTubers (Ekman, 2014; Sobande, 2017), the popularity and content of YouTube channels (García-Rapp, 2017; Ladhari et al., 2020), or the use of emotional labor, i.e., creating closeness and authenticity through which the attention and attachment of viewers is to be obtained (Berryman and Kavka, 2018; Raun, 2018; Rosenbusch et al., 2019). This research employs a wide range of methods. A large number of studies use qualitative interviews (Choi and Behm-Morawitz, 2017; Sobande, 2017; Bishop, 2019), video ethnographic methods or netnographic methods (García-Rapp, 2017; Mardon et al., 2018), qualitative content analysis, or discourse analysis (Montes-Vozmediano et al., 2018; Scolari and Fraticelli, 2018; Lewis et al., 2021). Quantitative analysis, webscraping, or Machine Learning-based methods are less frequently used in current YouTube research (Zeni et al., 2013; Schwemmer and Ziewiecki, 2018; Kalra et al., 2019; Obadimu et al., 2019), although the already quantified digital setting of the platform seems to lend itself to such an approach (Munger and Phillips, 2022).

This observation marks the starting point for this paper, in which we aim to use a random sample of German YouTube channels and apply four different classification methods in order to assess the gender of channel creators. We concentrate on the classification of gender for the following reasons: (1) gender shapes how people make sense of themselves, their social relationships, their networks, and their professional activity. A person's gender, once it becomes visible online, is important to describe people's behavior and is relevant to explaining mechanisms of inequality on social media platforms and beyond–this might even mimic or exaggerate gender inequalities that already exist in the offline world (Molyneaux et al., 2008; Wagner et al., 2015; Muñoz Morcillo et al., 2019). (2) We need to examine the contexts in which the absence of women in digital narratives, and the often stereotyped expression of them, form a constitutive part and reproduce a patriarchal system of imaginaries associated with prestige, reason, and power (Regueira et al., 2020). We also have to highlight which contexts provide new potential to change existing social hierarchies, tackle traditional boundaries to promoting mariginalized groups, and therefore could even play a role in turning women's or individuals with non-binary gender identities' talk into voice (Sreberny, 2005; Molyneaux et al., 2008). (3) However, when we focus on previous research, it becomes apparent that despite the existence of an enormous potential introduced by this data source, the social structure of the YouTube community is still ambiguous and not properly explored. The lack of personal information (such as gender, age, education, or ethnicity) is intriguing, because we know that individual characteristics have a significant impact on who creates online content in the first place, but equally important is which content is produced and why it is (not) widely circulated (Haraway, 2006; Regueira et al., 2020; van Dijk, 2020).

From a practical point of view, the YouTube API easily provides access to comprehensive data (such as content of videos, number of views, inter-personal comments etc.). Using text strings from channel names and descriptions we aim to distinguish male, female, and multi-agent presentations[1] We will evaluate and compare four classification methods in terms of the performance and meaningfulness of classification, as well as the resource efficiency of each approach.

The remaining sections of the paper are structured as follows: chapter two offers a summary of previous research on YouTube by focusing on frequently used methods in this realm. Thereafter, in chapter three we describe our dataset and provide precise information about the classification methods we use. We compare our reference data set gained through a multi-platform research to (i) a classification survey, (ii) a dictionary-based method, (iii) an algorithmic classification approach using gender-api.com, and (iv) a machine learning approach that uses Multinomial Naïve Bayes (MNB). In chapter four we present the performance of each method, focusing on accuracy, precision, and recall, as well as the Brier score and combined weighted classifier (Performance) (Kittler et al., 1998; Yan and Yan, 2006; Filho et al., 2016; Kalra et al., 2019; Weissman et al., 2019). Moreover, we discuss limits and benefits, and give some examples for misclassifications that showcase the meaningfulness of the acquired results (Meaningfulness) (similar as in studies such as Wu et al., 2015; Hartmann et al., 2019; Grimmer et al., 2021). The chapter concludes with remarks on benefits and challenges of using YouTube data. In chapter five we discuss our results and offer some concluding remarks.

State of the art

Advances in computational methods have opened up new possibilities in using social media data for social science research in the last decades. As a result, a multitude of text-based classification methods have been established in recent years. In the following, we want to give a short insight into the current use of a variety of text-based classification methods within the social

1 We refer to gender rather than to the biological sex, including expressions of gender roles and norms, as well as gender-specific representations in text-based descriptions, which we found on the web, viewing women, men and non-binary individuals as social categories and culturally constructed subjective identities (Oakley, 2016; Leavy, 2018). Here, we understand that social categories refer to the common identification with a social collectivity that creates a common culture among participants concerned, thus effecting individuals' self-perception and (online) behavior.

sciences, and illustrate their application to different topics and methodological challenges as they present themselves today.

A large number of studies, especially in the beginning of web-related research, have employed classification surveys to assess information that is difficult to access for automated methods, such as viewing experiences and emotions (Hoßfeld et al., 2011; Biel and Gatica-Perez, 2013). MoorMoor et al. (2010, p. 1539), for example, used several questionnaires to assess flaming on YouTube, defined as displaying hostility by insulting, swearing, or using other offensive language. Konijn et al. (2013) conducted a survey in a mixed-method study (that also featured experimental designs) of social media preferences and moral judgments among a younger YouTube audience. With respect to employed methods, the study of Fosch-Villaronga et al. (2021) is relevant as well, with the results of a survey among Twitter users showing that platform algorithms can (re-)produce inaccurate gender inference. The initial popularity of classification surveys has seen a decrease since the establishment of automatic classification approaches, which do not rely on the costly involvement of respondents. However, to this day, especially when complex classes such as gender identities are concerned, the use of surveys is an established method in classifying social media data.

Dictionary-based text classifications mark a shift toward automatic classification approaches. Their use is often resource intensive, because it requires the establishing of copious dictionaries that researchers share across generic domains or obtain from administrative or survey data (Hartmann et al., 2019, p. 23). The method is most widely used in research where multiple generic lexicons exist, such as in sentiment analysis (Feldman, 2013; Devika et al., 2016; Zad et al., 2021) which has produced an extensive literature. In other fields, dictionary methods have gained less prominence, since they underperform in comparison to algorithmic approaches and machine learning models (e.g., González-Bailon and Patoglou, 2015; Hartmann et al., 2019). Algorithmic classification approaches that use APIs of websites like, gender-api.com, genderize.io, NamSor, or Wiki-Gendersort are also quite common in recent studies (Karimi et al., 2016). These services offer automatic classification by comparing various types of character strings to large privately owned databases. In recent research this cost-effective method is used, for example, to explore the gender gap in scientific publications, or the identification of gender diversity in groups of knowledge production (Larivière et al., 2013; West et al., 2013; Fox et al., 2016; Giannakopoulos et al., 2018; Sebo, 2021). Although their emphasis lies in the analysis of gender-inference based on names, similar third-party APIs also exist for the detection of nationalities (e.g., https://nationalize.io/) or age (e.g., https://agify.io/).

Finally, more complex supervised machine learning approaches tackle a broader variety of goals when applied to YouTube data, such as classifying the content of YouTube videos (Kalra et al., 2019), or estimating the political ideology of channels (Dinkov et al., 2019). Most of them use video content (Kalra et al., 2019; Ribeiro et al., 2020) or viewer comments (Hartmann et al., 2019) as their main data source. When text from YouTube and other social media platforms is concerned, most approaches use Random Forest Classifiers (Kalra et al., 2019), Naïve Bayes (Hartmann et al., 2019), Support Vector Machines (Pratama and Sarno, 2015), K-nearest neighbor (Agarwal and Sureka, 2015) or a combination of multiple approaches (Park and Woo, 2019). For example, in a recent study Hartmann et al. (2019) compared 10 text classification approaches across 41 social media datasets and found that Naïve Bayes is well-suited for YouTube data and known to be fast, easy to implement, and computationally inexpensive (Filho et al., 2016; Kowsari et al., 2019). Various studies point out that classifying sociodemographic information in this way also holds questions of research ethics, such as the dangers of gender stereotyping through incorrect inferences of social media data, as was pointed out by Fosch-Villaronga et al. (2021). Some of the adverse consequences they highlight are statistical or legal discrimination, stigmatization, reinforcing of gender binarism, or self-identity issues.

It becomes clear that the range of available methods to tackle social media text classification is wide and ever-growing as more researchers take these information sources into account. However, it also becomes increasingly hard to gain an understanding of the benefits and limitations that these methods can offer. A number of methodological studies dedicated to the systematic comparison of methods already exists (González-Bailon and Patoglou, 2015; Jindal et al., 2015; Hartmann et al., 2019; Kowsari et al., 2019). However, this methodological literature is pioneered in disciplines such as computer science, often concerned with specific technical challenges. Adaption of these methods to social scientific research, and a systematic understanding of the quality and bias within these classifications, in terms of social issues, is still lacking, and marks the motivation for the study at hand.

Materials and methods

Dataset

The data we utilize in this paper consists of 200 German YouTube channels that we collected in March of 2020 using the YouTube API. Channels were selected with the help of a free of charge website (www.channelcrawler.com) that has since been made subject to a fee. In 2020, the website allowed identification of YouTube channels with more than 1,000 views in total. In order to randomize our sample, 200 channels which had uploaded a video most recently on March 17th of that year were chosen. This procedure avoided picking channels based on their topic, prominence, or number of views, which is common in some (qualitative) studies (Jerslev, 2016; Fägersten, 2017;

García-Rapp, 2017; Duguay, 2019; Wegener et al., 2020). The API provided us with information such as the YouTube channel name, the channel description, the number of views, as well as information on the 61,071 videos uploaded by the channels. We restricted our sample to 200 cases, because we used two resource-intensive and time-consuming research strategies (a multi-platform research and one online survey), for which we had a limited number of staff at the University of X.

Information we could not gather using the YouTube API was filled in using a multi-platform research strategy (Jordan, 2018; Van Bruwaene et al., 2020). At this stage, we looked up sociodemographic characteristics (such as gender, age, education, and ethnicity) on the web, including Facebook and Instagram profiles, Twitter accounts, Google and Wikipedia records, or other YouTube channels. This multi-platform research strategy was done by one female and one male researcher in May 2020. Each person classified 100 cases of the reference data set. In order to check for interrater reliability, we selected 40 cases which were processed by these researchers; results revealed no inconsistencies. Both proceeded in three steps. First, we used the $N = 200$ YouTube channels to extract available sociodemographic information from the channel descriptions, profile pictures, or other video content. This first step allowed us to classify gender in about two-thirds of all cases, age and ethnicity for roughly one-third, and education for roughly one-quarter of all channels hosts. In a second step, we looked at the Facebook, Instagram, and Twitter pages linked on these YouTube channels to find information that was missing. In a final step, we used Google and Wikipedia data to identify additional sociodemographic characteristics, if necessary. This course of action allowed us to fill in all information available, and therefore serves as our reference data set. One important distinction was whether the channel featured an individual YouTuber, or a form of multi-agent-channel (pair, group, organization). For those channels that were representative of an individual, the variables assigned included the gender of the YouTuber (female, male, non-binary) as well as age, ethnicity and educational level (if available).

The final reference data set consists of 26 (13%) female and 129 (65%) male individuals, as well as 39 (20%) multi-agent channels[2]. One channel (<1%) within our dataset featured a person with a self-declared non-binary gender identity (specifically identifying as a demiboy, a person with mostly male characteristics). Five channels (2%) could not be assigned due to missing information on gender categorizations and therefore classified as NA (not available). Aside from gender, the sample presents itself as diverse also in terms of age, ethnicity and video content. In the final reference data set, the average age of the YouTubers is 29 years, ranging from 11 to 63 years. In terms of ethnicity, a migration background was estimated in 15% of the cases (based on country of birth and surnames). The dataset consists of a large range of channels, including political channels, car enthusiasts, religious channels, gaming, beauty and lifestyle channels, local news, travel, and channels linked to TV shows. On average, each channel had uploaded 306 videos and collected 5 million views overall. However, the inequality within this distribution is significant, amounting to a Gini-coefficient of 0.94 with regards to views[3].

Classification methods

In this paper, we use four different classification methods to infer the gender of YouTubers from text information, and to compare the quality and limits of these approaches for social science research.

First, we conducted a classification survey, in which respondents were asked to identify gender identities based on text presented to them. An online questionnaire was generated using SoSci Survey (Leiner, 2019) and distributed among ten members of staff at the University of *X. The respondents varied by age, gender, education, and familial status, and were told to each classify about 20 randomly selected YouTube channels. In a first step, respondents were shown the name of the channel and its description. They were asked to categorize the channels by type of YouTuber into one of four categories: individual, group, organization, or other. When a channel was classified as "individual," respondents were asked to classify the gender by the following question: "Based on the name and the description of the channel, can a statement be made about the gender of the person?". Answer options included the following categories: Women, men, non-binary, no statement possible. The questionnaire also assessed the gender composition of multi-agent channels, and estimated the age and education background of YouTubers, categories which are not the central to the present paper.

Second, we used a dictionary based approach (Jaidka et al., 2020) to classify the channels. In our case, gender classification was made accessible by inferring the given names of YouTubers

2 The sociodemographic composition of YouTube creators is rarely studied. However, our sample seems to be in line with existing studies. For example, Debove et al. (2021, p. 4 ff.) found the percentage of women, men, and institutions among their sample of French science channels to be 12, 64, and 21%. Wegener et al. (2020) have roughly 17% female, 53% male, and 30% institutional creators in their study of top-rated German YouTube channels and Regueira et al. (2020) observed 10% women, 60% man, and 30% institutional creators in top-listed Spanish YouTube channels.

3 This relatively high viewership and distribution inequality are related to the fact that our random sample includes a very popular YouTube channels of a German TV Show ("Berlin Tag und Nacht"). However, most other channels in our sample present few views. Nevertheless, other studies on YouTube have shown that an unequal distribution of viewership is typical for this platform (Tang et al., 2012; Zhou et al., 2016).

from their channel names and channel descriptions. As a reference list, we used the second edition of the World Gender Name Dictionary (Raffo, 2021), which includes 26 million records of given names, including 62,000 names for Germany. The dictionary classification method compares all words of the channel names and channel descriptions against this database and counts the number of female and male names identified in the text. To classify multi-agent channels as a third category we defined a list of German key words for "we," "team," "institute," "organization," "firm," "company," "group," "us," "our." Thus, we were able to classify and count the number of female names, male names, and multi-agent identifiers. When no identifiers whatsoever were found, the channels remained unclassified. When both female and male names were present, but no multi-agent identifier, the majority category guided classification. In cases where the same amount of female and male names were found, but no multi-agent identifier, the channel remained unclassified.

Third, we applied an algorithmic classification approach using gender-api.com. The R implementation of this algorithmic classification enabled us to predict the gender of a YouTube channel creator. The method estimates the gender based on a character strings and given names (Wais, 2016), referring to a database. This database of gender-api.com is built on continuous scanning of public records, registry data, and public profiles and their gender data on major social networks, and offers 6,084,389 records in total. The website is free of charge if queries do not exceed a certain level per month. It displays the number of data records examined in order to calculate the response and releases probabilities, indicating the certainty of the assigned gender.

Fourth, we deployed a machine learning approach using Naïve Bayes Classifiers for small samples (see Filho et al., 2016; Hartmann et al., 2019)[4]. The text preprocessing for this step consisted of transforming all words to lower cases, removing URLs and separators (such as hyphens), as well as punctuation and single digits. Common stopwords (such as "or," "and," "he," "she," "we") were retained, since previous studies find that the removal of stop words lowers gender classification accuracy (Yan and Yan, 2006). These words also proved key to identifying multi-agent YouTube channels, similar to our dictionary approach[5]. Another important source of information was the gender specific use of Emojis, which is in line with recent studies (Wolny, 2016). Although the diverse Unicode representations of Emojis took some effort to account for in text preprocessing, these symbols proved very important in our classification. Finally, we did not conduct stemming of words in accordance with previous critiques that suggest the loss of important information (e.g., Dave et al., 2003; Bermingham and Smeaton, 2010). To estimate the out-of-sample accuracy, we split the dataset into a training set and a hold-out test set (80 vs. 20% of the data)[6]. The MNB model was trained on the training set, and the performance estimated on the test set. Laplace smoothing (a = 1) was applied as a regularization method, to avoid the zero-observation problem. Furthermore, since the categories in our dataset are not equally distributed, the prior probability of categories was factored into the model. However, there is still a large margin of error in randomly splitting a test and training sample. To better estimate the generalized performance of our method on YouTube data we applied the aforementioned procedure to five different splits of test/training data within our sample, and estimated the average performance across all five splits, also called outer fold cross validation (Parvandeh et al., 2020). This procedure also helped us to compare the output of machine learning algorithms to the other classification methods and identify particularly challenging cases.

Information used in classification

As Table 1 illustrates, the classification methods described above rely on different sources of information. To begin with, the classification survey presented the channel name and description to the respondents, and asked them whether they could estimate the author's gender on the basis of this information. The dictionary method also considers the channel name and description, whereas with gender-api.com we based their classification only on the channel name. API approaches can be extended using the channel description as well, but these longer texts also introduce a lot of noise that can be misinterpreted as names. Finally, the MNB model uses the channel descriptions as bag of words, and finds commonalities in words used across genders. In this case, the addition of

4 The NB is a probability-based approach that calculates the probability of a certain document to be part of a specific class given its features. In our case, we calculate the probability of a channel description to belong to one of four gender categories given the words and emojis is based on the following equation: $P(c_k|x) = P(c_k) \times \frac{P(x|c_k)}{P(x)}$. $P(c_k|x)$ being the conditional probability of the occurrence of a category given the existence of a vector of features x. $P(c_k)$ is the general probability of the occurrence of the category, $P(x|c_k)$ being the conditional probability of a certain word belonging to a category and $P(x)$ being the probability of the occurrence of the feature x. Naïve Bayes assumes that all features are independent from one another (Lewis, 1998).

5 Our machine learning model was not able to deal with non-binary cases due to the limited number of cases. We therefore decided to classify this single object as NA (not available), whilst formatting the data set. As final outcome categories for the machine learning model, we keep "female," "male," "multi-agent" and "NA".

6 In our study, the sample size of the dataset was relatively small, which is not ideal for machine learning approaches, but also not uncommon for social scientists working with data donations through surveys (Molyneaux et al., 2008; Muñoz Morcillo et al., 2019; Chen et al., 2021; Debove et al., 2021).

TABLE 1 Information regarded in single classification methods.

	Reference data set	Classification survey	Dictionary method	Gender Api	MNB
Channel name	●	●	●	●	.
Channel description	●	●	●	.	●
Channel profile picture	●	.	.	.	.
Video content	●	.	.	.	.
Information from other platform (e.g., Twitter)	●	.	.	.	.

Source: own illustration.

the channel name to these methods would be possible, but additional information is likely minimal unless the channel names follow a certain pattern, or are given more weight in comparison to the description[7].

Evaluation

We evaluated the performance of each text classification method in terms of four parameters: (1) First, we discussed each classification method in terms of the degree to which these approaches come to the same classification result as our reference data. Four measures were evaluated and explained using the following examples (Yan and Yan, 2006; similar as in Filho et al., 2016; Kalra et al., 2019; Weissman et al., 2019): *Accuracy*, which displays the ratio of correctly predicted women within all observations. Accuracy is well-suited to evaluate the overall performance of the methods, but also has some limitations (Kowsari et al., 2019). In datasets where the categories are unbalanced (one including more cases than the others), it is wise to include precision and recall as well. *Precision* (also known as positive predictive rate) measures how many of the channels we predicted as female, were actually female. Precision is most important when false positives are to be avoided, for example, when men should not wrongly be classified as women. To see how precise the overall method was, we used macro-averaged precision (Murphy, 2012, p. 183), which average the precision over all classes. *Recall* [also known as sensitivity or true positive rate (Murphy, 2012, p. 181)] quantifies how many, out of all actual women, were labeled as female. Recall is especially important when false negatives are to be avoided, for example, when we want to minimize the women overlooked by our classification. Again, Macro-recall averages the performance between all classes to evaluate the models as a whole. Finally, the *Brier score* is reported (Brier, 1950) for those methods that compute probabilities for a channel belonging to each of the classes. The Brier score takes into account how close the predictive probability was to the correct outcome, in our case if we assigned a female led channel the probability of being 60 or 95% female. The more accurate the prediction is, the closer the Brier score is to zero. The Brier score also has the advantage of handling predictions of multi-class classifications, making it useful in the application of gender prediction (including multi-agent channels). (2) Second, we also assessed the limits and benefits of the four methods to the study of YouTube data. This should help researchers to evaluate whether the method is realizable for them, and which trade-offs exist between those methods. (3) Third, the meaningfulness of the achieved gender classifications is an additional aspect we considered. This includes discussions of misclassifications, hard to reach groups and similar issues (see Fosch-Villaronga et al., 2021). (4) Fourth, we tackled ethical challenges that arise in our study, such as the reproduction of stereotypes and consent to participate in research.

Results

Performance

The classification survey approach shows the second-best overall performance in Table 2, with an accuracy, precision and recall around 60%. Since hand-coded survey classifications are useful as training data for machine learning approaches, they can play a big role in determining the quality of follow up methods used (e.g., Brew et al., 2010). In our case, no probability-based Brier score was available for this method, since each channel was classified only once. However, allowing for multiple classifications, and assessing the interrater-reliability and Brier

7 Machine learning models would be capable of integrating alternative procedures such as image classification, speech recognition, or face recognition for obtaining gender assignment estimates (Hinton, 2012; Balaban, 2015). However, problems with these methods remain far from being solved. For example, 2-D image representations of human faces exhibit large variations due to illumination, facial expression, pose, the complexity of the image background, and aging variations (Kasar et al., 2016). Moreover, image examples available for training face recognition machines are limited which makes the task of characterizing subjects difficult.

TABLE 2 Accuracy, precision and recall of classification methods.

		Classification survey	Dictionary method	Gender API	MNB ($n = 40$)	Average MNB (5 folds, $n = 200$)	Weighted vote
Male	Accuracy	0.688	0.598	0.593	0.675	0.718	**0.779**
	Precision	0.972	0.838	0.747	0.869	**0.872**	0.801
	Recall	0.535	0.477	0.569	0.667	0.746	**0.876**
Female	Accuracy	0.925	0.754	0.879	0.800	0.869	**0.894**
	Precision	0.824	0.274	0.533	0.222	0.228	**0.619**
	Recall	0.538	0.538	**0.615**	0.667	0.500	0.500
Multi-Agent	Accuracy	0.905	0.764	-	0.900	**0.849**	0.834
	Precision	0.702	0.390	-	0.571	0.520	**0.609**
	Recall	0.868	0.421	-	0.800	**0.620**	0.368
NA	Accuracy	0.678	0.819	0.573	0.975	**0.950**	0.709
	Precision	0.047	0.030	0.200	1.000	0.250	**0.525**
	Recall	0.500	0.200	0.326	0.500	0.250	**0.478**
Total sample	Accuracy	0.598	0.467	0.522	0.675	0.698	**0.709**
	Macro-Precision	0.636	0.383	0.494	0.658	**0.578**	0.525
	Macro-Recall	0.610	0.409	0.503	0.666	**0.485**	0.478
	Brier score	-	-	0.158	0.040	0.061	-

Source: own calculations; N = 200. Accuracy is the ratio of correctly predicted cases within all observations. Precision is the ratio of all correctly predicted cases within all prediction in one class. Recall is the ratio of correctly predicted cases within all cases that actually belong to said class. The Brier-score shows the accuracy of the probabilistic prediction. Bold values represent the highest scores in each row.

score based on the probability of classifications could increase the performance of this method.

The dictionary method based on the World Gender Name Dictionary performs the worst, with an overall accuracy, precision, and recall around 40%. This is not surprising, considering that multiple evaluation studies have found dictionary approaches to underperform in the past (e.g., González-Bailon and Patoglou, 2015; Hartmann et al., 2019). However, based on our experience, the accuracy might be improved given a reduced approach. As will be discussed in more detail below, the World Gender Name Dictionary consists of a large sample including rare names, which lead to misclassifications when applied to texts with non-name words. The performance of the dictionary method could increase, but only if common names are included and the text is preprocessed in advance.

The application of Gender API underperforms the hand-coded survey, with its overall accuracy, precision, and recall around 50%. This is surprising considering that Gender-API operates on the basis of a relatively large database, when compared to our dictionary and machine learning approach. However, in our case only the channel names were processed by the API, while the MNB and survey method included the channel description. Future research could evaluate whether the addition of channel descriptions contributes to the Gender APIs performance, or adds distracting information that worsens the scores.

Overall, it becomes clear that the MNB machine learning approach performance is the best of the four single classifiers. Taking into account the slight variation between the test sample ($n = 40$), and the average performance across all five folds ($n = 200$), the model's accuracy, precision, and recall all score around 66%. The Brier score of about 0.05 also attests high accuracy of the predictions based on probabilities. Multinomial Naïve Bayes is known to perform well with classifying text data, especially in small samples (Kowsari et al., 2019). However, considering that many research projects will have larger samples available, which might also be more thematically focused, one can expect that the MNB approach will perform even better in these cases.

Finally, we present results of a combined weighted classifier (Kittler et al., 1998) in order to further improve the decision for one (combined) classification approach over another (Seliya et al., 2009; Liu et al., 2014). We use a combined weighted vote classifier (Dogan and Birant, 2019), aggregating the individual performance metrics of all automated classification methods into one metric, which then could serve as a further basis for choosing the best classification strategy to determining the gender of YouTube creators efficiently in large-scale data. More precisely, we assigned the final gender classification of the automated methods of each YouTube channel a vote, then weighted those votes with the overall accuracy of each method, and counted the votes in the end. The linear weighting assured

that we obtain results even in those cases where each method assigns a different gender classification. When we compared the combined classifier to the multi-platform research strategy, we obtained 141 correctly classified cases. Thus, the combination of all three automated methods provided the highest overall accuracy for male and female accuracy, as well as precision for multi-agent and NA classification.

Looking at details of the performance in each gender category, we want to point out some further key insights of the methods: (1) The true value of the survey method seems to lie in its precision, where it clearly outperforms the other methods in its classification of men, women, and multi-agent channels. For men, the precision reaches 98%, meaning that 98% of male channels were correctly classified as male. (2) Interestingly, the survey and dictionary method seem to perform poorly when dealing with the No Answer-category, where they give rather moderate results. While the accuracy in this category is average, its precision of 2–4.7% is quite low. Both methods tend to give more conservative gender-estimates, which refrain from classification when no information is found, therefore increasing the number of NAs. In comparison, machine learning approaches generally tend to use any information given and will more likely estimate cases to belong to the majority groups (in our case male). (3) The Gender API approach is not designed to identify multi-agent channels. This illustrates an advantage for more adaptable dictionary approaches, which can add multi-agent identifiers (such as "we," "us," "our") to already existing name-lists. (4) All methods show lower precision when predicting women vs. men. Especially with the dictionary method and MNB, only 20% of predicted female-led channels were actually led by women. Concerning the machine learning approach, this problem derives from an imbalance between the classes[8], meaning that women are represented by a smaller number of cases in our sample and training data (Note: The accuracy is higher since it also takes correctly predicted men and multi-agent channels into account). In contrast, the survey seems very apt at classifying women both in accuracy and precision. (5) The combined weighted classifier demonstrates the probabilities of what can be achieved when multiple methods are integrated into one model. It produced a higher precision in its prediction of minority classes than the single automated methods, and increased the overall number of correctly predicted cases. Depending on the research interest, this increased performance could prove to be a vital step toward a better classification of text based social media content.

Limits and benefits

The survey method is a cost intensive, but highly valid, and an adaptable approach to classify YouTube data. The classification questionnaire can be closely tailored to the researchers' interest, and easily allows for the inquiry into multiple variables at once (Hoßfeld et al., 2011; Biel and Gatica-Perez, 2013). Furthermore, one can implement multiple sources of information for the respondents to classify, such as pictures, text, or even audio or video material. Since human respondents can synthesize different kinds of information more easily, this permits precise categorizations. However, the researcher should be aware that this approach is time-consuming, taking into account the development and testing of the questionnaire, as well as its distribution among respondents. The median time for the classification of the channel type and potential gender of the YouTuber amounted to 28 s per case. Since this data set only consisted of 200 cases, the overall amount of time set aside for the actual classification was manageable. If, however, one was to apply the same method to a large set of data, or include further sources as stimulus for the respondents, more time would be needed. Additionally, this method relies on the availability of trustworthy respondents and meaningful names and descriptions provided by the YouTube channel. While other methods can be easily repeated in case of a mistake, this can be rather difficult for the survey method, requiring accurate survey construction and pretesting.

As mentioned, the expenditure of dictionary-based methods is highly dependent on the preexistence and availability of dictionaries, since their construction takes a lot of time and effort (González-Bailon and Patoglou, 2015; Rosenbusch et al., 2019). In our case, name-based gender identification proved a feasible strategy, since name lists are a relatively common open-source material. The World Gender Name Dictionary (Raffo, 2021) proves an extensive resource that is applicable to a wide range of countries. Therefore, its use on YouTube data can be recommended. However, as our detailed examples will show, researchers should put careful thought into the range of names, and the type of text this method is applied to. More "fuzzy" text always yields the potential to misclassify random words as names, thus adding errors into the gender score. The method is most resource effective when the likelihood of names (and only names) appearing in the text is high, as in the example of channel

8 Imbalanced datasets present a challenge to machine learning algorithms for which various strategies exist (Weiss, 2013). A common way is to resample the training dataset by either undersampling the majority class or oversampling the minority class (Chawla et al., 2002; Agrawal et al., 2015). In our case we refrained from this step since (1) our imbalance is only moderate with the highest proportional difference being between the male and the NA class. (2) This class imbalance in our data seems to mimic real life as it is very similar to previous findings in other studies (see Regueira et al., 2020; Wegener et al., 2020; Debove et al., 2021) and thus gives the machine learning algorithm further information about the natural occurrence of each class. (3) We used a modest data set of $N = 200$ (see chapter 3). Undersampling the male class would risk the loss of valuable information which is needed for a valid classification and oversampling the three minority classes risked problems of overfitting the model, since the relative number of cases was rather low.

names. Channel descriptions can also include names, which may not be present in channel names. However, they also introduce a lot of other words, and therefore an increased probability of misclassifications. This could extend the time needed for text cleaning, such as removing stopwords in order to reduce errors. Therefore, the fit of the text to the dictionary should be assessed carefully. Furthermore, the identification of multi-agent channels made the definition of identifying words necessary. As explained before, our list of identifiers included only 9 words (e.g., "us," "we," "our"), which were chosen at face value. More effort could be spent to empirically identify key words that are present in YouTube channels managed by multiple people, to create a more evidence-based dictionary. However, this would rely on a database of pre-classified channels, whereas our strategy could be applied without known cases.

The implementation of Gender API is simple and time efficient (Karimi et al., 2016). Gender API applies an already trained algorithm by comparing the YouTube channel names to an unknown online web data basis, and therefore does not require text preprocessing as long as only channel names are included in the analysis (Wais, 2016). The code is made available to implement the algorithm into common programming languages (including R and python) Alternatively, the website offers a service to simply upload text columns online (e.g., using Excel or csv files) and receiving finished classification results. For evaluating the time efficiency, the API provides the duration for assessing the gender for each record in seconds. For one record the Gender API required around 20 milliseconds to assess the gender. However, in order to process large data volumes, it would be necessary to make use of a fee-requiring premium account. At the time of writing this article, the API allows 500 names to be classified per month without charge (see https://gender-api.com/).

Finally, as with all supervised machine learning approaches, our MNB model relied on the availability of a reliable, labeled dataset to train the model (Agarwal and Sureka, 2015; Parvandeh et al., 2020). In our case, the training data consisted of a dataset constructed by the authors on the basis of a multi-platform research. This approach is time intensive and requires accurate assessment of multiple sources of information. More time efficient approaches than the multi-platform research could include a classification survey as we used in our first approach, a self-reporting survey amongst YouTubers or even using commercial providers for the human-based labeling of huge amounts of data, e.g., Amazon Mechanical Turk. Once the machine learning model is established, it can be applied to new and large datasets not feasible for manual coding. This approach is especially efficient when very large samples are available, or the number of channels that have to be coded is unclear (e.g., channels are added to the dataset over time). As our example shows, the setup of a MNB classifier is relatively simple and time efficient. Since the model assumes no relation between the features, and relies on simple word count, there are few hyperparameters that have to be tuned and monitored. However, the text preprocessing is an important step before training the model and requires careful attention. In our case, the treatment of stopwords and unicodes provided challenges, as will be further explained below. Furthermore, as shown for the three automated methods, the machine learning classification can be improved through its' combination with other methods.

Meaningfulness

The performance and cost-efficiency of methods must also be weighed against the meaningfulness and interpretability of the results, especially when sensitive subjects such as gender are involved (Wu et al., 2015; Hartmann et al., 2019). To evaluate our results, we provide an exemplarily illustration of seven YouTube channels, chosen in order to present differences and problems that occurred in our classification methods (see Table 3 for details).

First of all, name-based approaches risk the misclassification of common words as given-names. The channel "Jana's Welt" ["Jana's World"] is hosted by a woman but assessed as male by Gender API. This discrepancy is based on the fact, that Gender API uses "Welt" ["World"] as sole gender indicator (excluding "Jana's" as a reference word), thus assessing a male gender with a probability of 100%. The algorithm is only referring to four examples of "Welt" in the underlying (unknown) online database, whereas usually the algorithm classifies other records based on several thousand examples. Nevertheless, as the underlying algorithm is unknown to us, the actual decision making process of Gender API remains a sort of black box. Jana's Welt was also not recognized as a female name by our dictionary approach, likely due to the possessive "s" included in the name. It remained unclassified by the dictionary method, since no names were detected. The channel does not provide a channel description that can serve as the basis for further information. Only the survey managed to classify this case correctly.

Looking at the channel "Cookie" we know that this channel is hosted by a man. We obtained no result by the Gender API (non-classified), since no name was detected. Interestingly, in this case the survey method also failed to correctly classify this channel[9]. Even though the channel description mentions the name of the YouTuber (see Table 3), the respondents reported difficulties with deciding whether "Felipe" was a male or a female name. Similar problems might arise with names that are uncommon among the German population, or that are gendered differently in different cultures (i.e., Andrea being a female name

9 However, in other examples of fictious names the survey approach might be more powerful in classifying gender information. One example is the use of names, associated with a gender, such as "Legolas" or "Yoda" as prominent (science) fiction characters from Lord of the Rings and Star Wars.

TABLE 3 Exemplary results of different classification methods.

Channel name and description	Ref.	Survey	Dict.	API	MNB	Vote
Jana's Welt	Female	Female	NA	Male	Male	Male
Cookie Hi I am Felipe. I only do YouTube and Twitch as a hobby. On this channel is actually only gaming content such as Fortnite. Have fun on my channel	Male	NA	Female	NA	Male	Male
Gleichberechtigt - A self-portrait - Born in Baden in the 196x-er, studied technology at the University of Karlsruhe, graduated in 1993, employed for 4 years, then self-employed in EDP. Why not more precise?–Because state-subsidized terror executes again and again "progressive" politics of the left establishment CDU/CSU, SPD, Greens, Left and FDP and the 68'er justice finds pleasure in it–briefly because "DDR 2.0." (…)	Male	Male	Female	NA	Multi-agent	Multi-agent
Christelle Proudwatcher Christelle Proudwatcher Player level 20 Server 11 | Germany 53 horses (…)	NA (lgbtqia+)	Female	Female	Female	NA	Female
Tini and Uwe Mayer "The Mayers on Tour"–that's Tini and Uwe Mayer–formerly from Göppingen in Baden Württemberg. Our topics: Moving into the camper and "living on the road"–travel–photography–image editing–music–lifestyle. (…)	Multi-agent	Multi-agent	Multi-agent	Male	Multi-agent	Multi-agent
Faina Yunusova П р и в е т ! Я х у д о ж н и к . М о я ц е л ь - п о з н а к о м и т ь в а с с с о в р е м е н н ы м и с к у с с т в о м , х у д о ж н и к а м и и и н т е р е с н ы м и и д е я м и ! П р и с о е д и н я й т е с ь !	Female	NA	Female	Female	Female	Female

Source: own calculations, texts were translated from German to English by the authors.

in Germany and a male name in Italy). In this case, the MNB model was the only method successful in classifying the gender correctly based on the content of the channel description.

The channel "Gleichberechtigt" is an example of a YouTuber who reveals a lot information about himself in his channel description. Coding by hand or through the survey, we could identify gender, decade of birth, education and even occupational path. However, this case also illustrates the limitations of automatic classifications. Since the channel name "Gleichberechtigt" (meaning "having equal rights" in German) does not hold any name information, the dictionary method and Gender API could not derive any classification from this information. The dictionary method however, thought it identified five female names and four male names in the channel description, and misclassified the channel as female. One concern when using large dictionaries on social media text, is that random words can be misinterpreted as names. For example, the dictionary recognizes "mehr" as a Persian name present in German records. However, "mehr" is also the German word for "more," misinterpreted as a name in this case. Finally, the MNB misclassified the channel as a multi-agent channel, likely because the description talks about many

political issues, also present in other news or party channels in our sample.

The classification methods presented in this paper aim to capture the gender self-presentation of the owners of YouTube channels. They are dependent on YouTube channels to reveal gender-relevant information within the texts or pictures representing the channel. As we have seen within our sample, many channels include the given name (or a self-chosen given name) of the YouTuber within the channel name or description, which allows for gender inference. However, these methods also have limitations when more nuanced gender-identities are concerned. One such case in our dataset is "Christelle" who identifies themselves as a demiboy[10] in their introduction video and focuses their channel around a game called "starstable" and lqbtqia+ pride content. However, since this identity is not declared in the name or channel description, our methods mostly misclassified them as a female. Christelle is also the only apparent lgbtqia+ member in our small sample, making it difficult for machine learning approaches to consider these gender identities. Less common identities such as Christelle's will likely be underestimated in most automatic classification efforts, which should be taken into account in research design.

Tini and Uwe Mayer present an example of a multi-agent channel owned and lead by a couple. YouTube, like many other social media platforms, hosts a mix of private channels representing individuals, as well as a variety of multi-agent channels. Our sample includes channels by couples or groups of people (e.g., bands, siblings, married couples), as well as organizational channels (e.g., news outlets, TV-shows, political parties, co-operations). It can be important to distinguish between these types of channels, not only when gender is concerned, since the resources behind public or professional channels might differ significantly, therefore the number of videos, content, and views reached might also be significantly different. In terms of gender identification, these multi-agent channels provide some difficulties. In some cases, the gender of their members may be classifiable, such as with "Tini and Uwe Mayer" (see Table 3), which could be classified as a multi-agent channel by most methods, and could further be identified as consisting of a man and a woman. Using Gender API, this channel was wrongly assessed as male (with a high probability of 96%), because multi-agent channels were not included and therefore "Tini and Uwe Mayer" was read as one single male individual by the algorithm. This problem is not to be neglected, since according to the survey, out of 47 multi-agent channels in our sample, 15 were classified as multi-agent channels (groups or pairs) and 32 as non-agent channels (e.g., events, organizations).

Finally, the channel "Faina Yunusova" illustrates the problem of multi-lingual channels in our sample. Although our sample included only German YouTubers, several channels from female YouTubers used the Cyrillic alphabet. Since this YouTuber uses a given name, the dictionary method as well as the Gender API managed to classify this channel correctly as female. However, the respondents of our survey did not know the gender of the name "Faina," neither were they able to read the Cyrillic description, and therefore did not classify this channel. Interestingly, due to the small sample size and few opportunities to compare, the MNB model interprets Cyrillic letters as being more representative of female YouTubers, and therefore classifies Faina as a woman. Furthermore, a problem arises because Cyrillic letters are represented as Unicode in our dataset (e.g., the letter и is represented as <U+0438>). The machine learning approach interprets these unicodes as words instead of letters, giving each letter more weight in the final data. Such encrypting problems resulting from multiple language use are likely common in social media data. On the one hand, authors have to decide whether these transnational identities are important for their research or not, and if more rigorous data cleaning has to be applied beforehand to remove unicodes. On the other hand, unicodes such as emojis can also yield important information for the model. For example, in our sample heart emojis were more commonly used by female YouTubers. Such gender specific use of emojis can greatly aid when using a machine learning model. Several studies concur that emoji use is especially beneficial in determining the author's gender (Wolf, 2000; Chen et al., 2018; Beltran et al., 2021).

Ethical challenges

Based on our study, we want to contribute to existing research by highlighting some ethical challenges discussed in social sciences, which may arise from the inference of gender from YouTube data. One major pitfall of applying automatic classification methods involves the (re-)production of gender stereotypes (Dinan et al., 2020). The MNB machine learning approach is especially at risk of such behavior, since all words of the channel description are processed and assigned with a certain gendered probability, based on the information the model derives from the training data. However, if the training data finds men to be mainly dealing with politics, and women with beauty issues, the attributed words will then be associated with stereotypic gender categories. This reproduction of statistical differences is known as statistical discrimination (Arrow, 1974) in the social sciences, and is related to profound consequences, especially when looking at members of small or vulnerable groups (Leavy, 2018). At this point, it seems plausible that representatives of the lgbtqia+ community, for example, would have to face higher risks of stereotypical gender classification or even misclassification, since randomly selected training data presumably does not rely on valid information in this realm. With respect to our own study, we find an unwanted association between Cyrillic letters and women, as

10 The term demiboy describes a non-binary gender identity with predominantly male characteristics.

well as a higher association of men with video games (see Meaningfulness for examples). While the first observation is bound to the process of stereotypical classification and calls for more rigorous pre-processing of the text data, the second observation might represent both aspects at the same time: the result of biased training data and/or an interesting finding. This underlines the need for a thoughtful interpretation of results, a diligent evaluation of the field of application, sample selection criteria and the fit of research question to the selected design[11].

Discussion

The purpose of the present paper was to compare four text-based classification methods in order to assess the gender of German social media content creators. By using the example of a random sample of 200 YouTube channels, we compare a classification survey, a name dictionary method with the World Gender Name Dictionary as a reference list, an algorithmic approach using APIs of the website gender-api.com, and a Multinomial Naïve Bayes (MNB) machine learning technique. With the help of these different approaches, we identified gender attributes based on YouTube channel names or descriptions, and contrasted our results with a reference dataset to evaluate them. The reference dataset contained all information available on each channel using a multi-platform research strategy (Jordan, 2018; Van Bruwaene et al., 2020), including YouTube channels, Facebook and Instagram profiles, Twitter accounts, Google and Wikipedia data.

Our main conclusions concerning the pros and cons of each method are summarized in Table 4. They reveal that the MNB machine learning technique performs the best within our sample of single classifiers, since the model's accuracy, precision and recall all score highly (∼66%). However, the presence of a training sample is required, and one should be aware of stereotypical classification problems (see Ethical challenges). Second best is the online survey method, with accuracy, precision, and recall scores around 60%, especially when multiple information sources are combined. Here, one should take into account that this method is rather time consuming and possibly in need of a large number of respondents. Using gender-api.com underperforms the classification survey, with its overall accuracy, precision, and recall around 50%. Nevertheless, this method is simple, time efficient, and the use of resources is quite low when small data volumes are processed. The dictionary method based on the World Gender Name Dictionary performs the worst, with its overall accuracy, precision, and recall around 40%. Here the performance is especially low when the text includes a lot of non-name noise. Finally, with respect to the combined voted classifier (Kittler et al., 1998), we observe that the integration of all three automated classification techniques would yield even better results on gender classification outcomes than single classifiers (Khaled and Ali, 2020). These improved results are achieved because the weaknesses of each single classification method is compensated for in the combined metric, and should therefore be noted in future research.

We have shown that the inference of gender categories from YouTube channel names and descriptions is very well-possible, given some limitations. At best, about two thirds of channels will be correctly classified, depending on the methods used. In our case, the combination of automated classification techniques outperformed the other methods. The availability of a valid training data set is key to the quality of the outcome, and decisive for the level of detail achieved in this kind of research. Nevertheless, our study also shows that the final classifications do have their biases. They overestimate the presence of men on YouTube, for example due to false name-classification. Minority groups such as women, and more extensively non-binary gender identities, remain underrepresented or undetectable by the methods presented.

In light of our results, we want to offer some further thoughts on the use of (automated) classification methods for the social sciences. Overall, the classification of socio demographic characteristics is a key agenda for this field of study, because it allows scholars to explore the social contexts of online behavior. If we remain blind to the enhanced functionalities of gender, but also age, ethnicity, or education in online spaces, we risk overlooking the social structures and inequalities in contemporary digitized societies (Wagner et al., 2015; Karimi et al., 2016). Because the lack of information on vulnerable groups (e.g., women or non-binary individuals) and the hurdle to gather other crucial socio demographic characteristics (e.g., education or migrant background) opens a window of stereotypic digital narratives, preventing to tackle traditional patriarchal images associated with prestige, reason and power (Sobande, 2017; Fosch-Villaronga et al., 2021). This becomes even more relevant when we interlink social science theories and empirical findings to the emerging research field of machine learning. Against this background, we want to encourage scholars to further elaborate on text-based classification methods of social media data in future research:

11 In terms of sample selection criteria and field of application, our study points towards to some additional ethical challenges, which are not central to the present paper, yet interesting to discuss in future research. Our sample includes several YouTube channels which feature minors under the age of 10, and although all information is made publicly available (most likely by their parents or agents), these children are vulnerable subjects of research as their consent to the publication of the material cannot be taken for granted. This calls for a broader discussion on how to handle the passive participation of individuals portrayed in YouTube channels in social science research, although scholars do not technically require the direct consent of the subjects, nor is it (yet) necessary to inform them about the study.

TABLE 4 Overview of the results.

	Class survey	Name dictionary	Gender API	MNB	Weighted vote
Performance	High, especially with multiple sources combined	Low, especially when text includes a lot of non-name noise	Moderate, depending on the noise within the text	High, especially for large samples and majority groups	High, even for minority groups
Limits and benefits	Time consuming, though with little requirements	Very low when already present dictionary (e.g., WGND) are used; text preprocessing might be necessary	Very low when small data volumes are processed, large volumes require a fee	Presence of a training sample is required. Otherwise, low number of parameters	Low, but dependent on existing models that are included and their requirements
Meaningfulness	High, though dependent on the openness of answers available to respondents	Dependent on the noise within the text, and number of words misidentified as names; identifies names can be accessed	Dependent on the noise within the text, and number of words misidentified as names; high accessibility of feature probabilities	High accessibility of feature probabilities, chance of stereotypical classification	Dependent on previous models included in the vote and their meaningfulness
Ethical challenges	Reinforcement of stereotypes based on individual experiences of respondents	Reinforcement of stereotypes based on country-specific name lists	Reinforcement of stereotypes based on structure of unknown online reference data	Reinforcement of stereotypes based on bias in the training data	Reinforcement on stereotypes and misclassification of included models

Source: own illustration.

- To date, we see great potential in automated classification methods in social science matters, since the results achieved by these relatively simple approaches are impressive and especially eligible for processing great volumes of data. However, this paper focused on gender classification which is more easily detectable and assignable compared to ethnicity, educational background, or occupational affiliation for example. Therefore, we also see some credible challenges, which should be subject to future studies.
- In light of key empirical findings and existing challenges, we would strongly recommend the combination of the application of ML based text classification with other methods, such as self-reporting surveys or classification surveys in order to generate precise data that allows the investigation of (re-)producing social inequalities in platform-based societies (van Dijk, 2020).
- We encourage researchers to actively counter steer the invisibility or misrepresentation of information within automated classifications of social media data, especially when marginalized groups are involved. At this point, more research is needed to find ways to reduce the bias present in all methods discussed above. This again indicates a need for elaborating on existing classification methods, and might even point toward the requirement to integrating other methods, for example in-depth qualitative interviews, in order to tackle blind spots and achieve a solid interlinkage of theory production and empirical research.
- Based on our findings, the presence of Emojis, multiple languages (which might provide encoding issues), multi-agent channels, and "noisy" text in the YouTube channel descriptions present hurdles to automated classifications. We have outlined some strategies to mitigate these problems in the presented study. However, these topics also warrant more methodological inquiry.
- Finally, we are convinced that further research should be dedicated to the valuation of multiple information sources available on YouTube and other social media platforms such as Instagram or Tiktok. In the present study, we use the channel names and descriptions as the only data source. Nevertheless, video content, channel profile pictures, audio and video data are further valuable sources of information, which might still be in their early days of development, but already yield some promising and trendsetting approaches.

Data availability statement

The data used in this paper is available on GitLab (project id 37844399).

Author contributions

LS and RV revised the project, the main conceptual ideas, and proof outline. LS prepared the reference data set, performed

the dictionary categorizations, and designed the machine learning model. CB designed the SoSci survey and analyzed the survey data. N-SF performed the API categorizations. LS, RV, CB, and N-SF contributed to the initial submission of the manuscript. RV, CB, and N-SF produced the final version of this text, which was approved by all authors.

Funding

The work on this study was supported through a research grant of the German Research Foundation (DFG, Grant Number VE 375/10-1).

Acknowledgments

Jan Paul Möller helped preparing the reference data set. Ulrich Kohler commented on a first draft of this contribution. We also thank the reviewers and the editors of this Special Issue for a number of very helpful comments and their productive critique.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Agarwal, S., and Sureka, A. (2015). "Using KNN and SVM based one-class classifier for detecting online radicalization on twitter," in *Distributed Computing and Internet Technology, eds* R. Natarajan, G. Barua, and M. R. Patra (Cham: Springer).

Agrawal, A., Viktor, H. L., and Paquet, E. (2015). "SCUT: multi-class imbalanced data classification using SMOTE and cluster-based undersampling," in *7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. Lisbon: IEEE.

Arrow, K. J. (1974). *The Theory of Discrimination*. Princeton: Princeton University Press.

Balaban, S. (2015). Deep learning and face recogniction: the state of the art. *Paper Presented at the Biometric and Surveillance Technology for Human and Activity Identification XII*. Baltimore, MD: SPIE Defense + Security.

Beltran, J., Gallego, A., Huidobro, A., Romero, E., and Padró, L. (2021). Male and female politicians on Twitter: a machine learning approach. *Eur. J. Polit. Res*. 60, 239–251. doi: 10.1111/1475-6765.12392

Bermingham, A., and Smeaton, A. F. (2010). "Classifying sentiment in microblogs: is brevity an advantage?," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, eds J. Huang (Toronto: ACM).

Berryman, R., and Kavka, M. (2018). Crying on youtube: vlogs, self-exposure and the productivity of negative affect. *Convergence*. 24, 85–98. doi: 10.1177/1354856517736981

Biel, J.-I., and Gatica-Perez, D. (2013). The youtube lens: crowdsourced personality, impressions and audiovisual analysis of Vlogs. *IEEE Trans. Multimedia*. 15, 41–55. doi: 10.1109/TMM.2012.2225032

Bishop, S. (2019). Managing visibility on YouTube through algorithmic gossip. *New Media Soc*. 21, 2589–2606. doi: 10.1177/1461444819854731

Bishop, S. (2020). Algorithmic experts: selling algorithmic lore on Youtube. *Soc. Media Soc*. 6, 1–11. doi: 10.1177/2056305119897323

Boxman-Shabtai, L. (2018). The practice of parodying: YouTube as a hybrid field of cultural production. *Media Cult Soc*. 41, 3–20. doi: 10.1177/0163443718772180

Brew, A., Greene, D., and Cunningham, P. (2010). Using crowdsourcing and active learning to track sentiment in online media. *Paper Presented at the Proceedings of the 2010 Conference on ECAI 2010: 19th European Conference on Artificial Intelligence*. Lisbon: IOS Press.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Mon Weather Rev*. 78, 1–2.

Bryant, L. V. (2020). The youtube algorithm and the alt-right filter bubble. *Open Inform Sci*. 4, 85–90. doi: 10.1515/opis-2020-0007

Burgess, J., and Green, J. (2018). *YouTube: Online Video and Participatory Culture*. Cambridge: Polity Press.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res*. 16, 321–357. doi: 10.1613/jair.953

Chen, Z., Lu, X., Ai, W., Li, H., Mei, Q., and Liu, X. (2018). "Through a gender lens: learning usage patterns of emojis from large-scale android users," in *Proceedings of the 2018 World Wide Web Conference*. Lyon.

Choi, G. Y., and Behm-Morawitz, E. (2017). Giving a new makeover to STEAM: establishing YouTube beauty gurus as digital literacy educators through messages and effects on viewers. *Comput. Human Behav*. 73, 80–91. doi: 10.1016/j.chb.2017.03.034

Dave, K., Lawrence, S., and Pennock, D. M. (2003). "Mining the peanut gallery: opinion extraction and semantic classification of product reviews," in *Proceedings of the 12th International Conference on World Wide Web*, eds G. Hencsey and B. White (Budapest: ACM), 519–528.

Debove, S., Füchslin, T., Louis, T., and Masselot, P. (2021). French science communication on youtube: a survey of individual and institutional communicators and their channel characteristics. *Front. Commun*. 6, 612667. doi: 10.3389/fcomm.2021.612667

Devika, M. D., Sunitha, C., and Ganesh, A. (2016). Sentiment analysis: a comparative study on different approaches. *Procedia Comput. Sci*. 87, 44–49. doi: 10.1016/j.procs.2016.05.124

Dinan, E., Fan, A., Wu, L., Weston, J., Kiela, D., and Williams, A. (2020). "Multi-dimensional gender bias classification," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg: Association for Computational Linguistics (ACL).

Dinkov, Y., Ali, A., Koychev, I., and Nakov, P. (2019). Predicting the leading political ideology of youtube channels usind acoustic, textual, and metadata information. *Proceed. Interspeech*. Graz: ISCA.

Dogan, A., and Birant, D. (2019). "A weighted majority voting ensemble approach for classification," in *International Conference on Computer Science and Engineering*. Samsun.

Duffy, B. E. (2020). Algorithmic precarity in cultural work. *Commun. Public*. 5, 103–107. doi: 10.1177/2057047320959855

Duguay, S. (2019). Running the numbers: modes of microcelebrity labor in queer women's self-representation on instagram and vine. *Soc. Media Soc.* 5, 1–11. doi: 10.1177/2056305119894002

Ekman, M. (2014). The dark side of online activism: Swedish right-wing extremist video activism on YouTube. *MedieKultur* 30, 79–99. doi: 10.7146/mediekultur.v30i56.8967

Fägersten, K. B. (2017). The role of swearing in creating an online persona: the case of YouTuber PewDiePie. *Discourse Context Media*. 18, 1–10. doi: 10.1016/j.dcm.2017.04.002

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Commun. ACM*. 56, 82–89. doi: 10.1145/2436256.2436274

Filho, J., Pasti, R., and de Castro, L. N. (2016). "Gender classification of twitter data based on textual meta-attributes extraction," in *New Advances in Information Systems and Technologies*, *eds* Á. Rocha, A. M. Correia, H. Adeli, L. P. Reis, and M. M. Teixeira (Cham: Springer), 1025–1034.

Fosch-Villaronga, E., Poulsen, A., Søraa, R. A., and Custers, B. H. M. (2021). A little bird told me your gender: gender inferences in social media. *Inf. Process Manag.* 58, 102541. doi: 10.1016/j.ipm.2021.102541

Fox, C., Burns, S., Muncy, A., and Meyer, J. (2016). Gender differences in patterns of authorship do not affect peer review outcomes at an ecology journal. *Funct. Ecol.* 30, 126–139. doi: 10.1111/1365-2435.12587

García-Rapp, F. (2017). Popularity markers on youtube's attention economy: the case of Bubzbeauty. *Celebr. Stud.* 8, 228–245. doi: 10.1080/19392397.2016.1242430

Giannakopoulos, O., Kalatzis, N., Roussaki, I., and Papavassiliou, S. (2018). *Gender Recognition Based on Social Networks for Multimedia Production. 13th Image, Video, and Multidimensional Signal Processing Workshop*. Aristi Village: IEEE.

González-Bailon, S., and Patoglou, G. (2015). Signals of public opinion in online communication: a comparison of methods and data sources. *Ann. Am. Acad. Pol. Soc. Sci.* 659, 95–107. doi: 10.1177/0002716215569192

Grimmer, J., Roberts, M. E., and Stewart, B. M. (2021). Machine learning for social science: an agnostic approach. *Ann. Rev. Polit. Sci.* 24, 395–419. doi: 10.1146/annurev-polisci-053119-015921

Haraway, D. (2006). "A cyborg manifesto: Science, technology, and socialist-feminism in the Late 20th Century," in *The International Handbook of Virtual Learning Environments, eds* J. Weiss, J. Nolan, J. Hunsinger, and P. Trifonas (Netherlands: Springer), 117–158.

Hartmann, J., Huppertz, J., Schamp, C., and Heitmann, M. (2019). Comparing automated text classification methods. *Int. J. Res. Mark.* 36, 20–38. doi: 10.1016/j.ijresmar.2018.09.009

Hassan, K. R., and Ali, I. H. (2020). "Age and gender classification using multiple convolutional neural network," in *IOP Conf. Series: Materials Science and Engineering (928)*. Thi-Qar (Iraq): IOP.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-R., Jaitly, N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE*. 29, 82–97. doi: 10.1109/MSP.2012.2205597

Hoßfeld, T., Seufert, M., Hirth, M., Zinner, T., Tran-Gia, P., and Schatz, R. (2011). Quantification of YouTube QoE via crowdsourcing. *IEEE International Symposium on Multimedia.* Dana Point, CA: IEEE.

Jaidka, K., Giorgi, S., Schwartz, H. A., Kern, M. L., Ungar, L. H., and Eichstaedt, J. C. (2020). Estimating geographic subjective well-being from Twitter: a comparison of dictionary and data-driven language methods. *Proc. Nat. Acad. Sci.* 117, 10165–10171. doi: 10.1073/pnas.1906364117

Jerslev, A. (2016). In the time of the microcelebrity: celebrification and the youtuber zoella. *Int. J. Commun.* 10, 5233–5251.

Jindal, R., Malhotra, R., and Jain, A. (2015). Techniques for text classification: literature review and current trends. *Webology*. 12, a139.

Jordan, K. (2018). Validity, reliability, and the case for participant-centered research: reflections on a multi-platform social media study. *Int. J. Hum-Comput. Int.* 34, 913–921. doi: 10.1080/10447318.2018.1471570

Kalra, G. S., Kathuria, R. S., and Kumar, A. (2019). "Youtube video classification based on title and description text," in *Proceedins of the 2019 International Conference on Computing, Communication, and Intelligent Systems*. Greater Noida: ICCCIS.

Karimi, F., Wagner, C., Lemmerich, F., Jadidi, M., and Strohmaier, M. (2016). "Inferring gender from names on the web: a comparative evaluation of gender detection methods," in *Proceedings of the 25th International Conference Companion on World Wide Web*. Montreal: WWW '16.

Kasar, M., Bhattacharyya, D., and Kim, T. H. (2016). Face recognition using neural network: a review. *Int. J. Secur. Appl.* 10, 81–100. doi: 10.14257/ijsia.2016.10.3.08

Kittler, J., Hatef, M., Duin, R., and Matas, J. (1998). On combining classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 226–239. doi: 10.1109/34.667881

Konijn, E. A., Veldhuis, J., and Plaisier, X. S. (2013). YouTube as a research tool: three approaches. *Cyberpsychol. Behav. Soc. Network*. 16, 695–701. doi: 10.1089/cyber.2012.0357

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., and Brown, D. (2019). Text classification algorithms: a survey. *Information* 10, 1–68. doi: 10.3390/info10040150

Ladhari, R., Massa, E., and Skandrani, H. (2020). YouTube vloggers' popularity and influence: the roles of homophily, emotional attachment, and expertise. *J. Retail. Consum. Serv.* 54, 102027. doi: 10.1016/j.jretconser.2019.102027

Larivière, V., Ni, C., Gringras, Y., Cornin, B., and Sugimoto, C. (2013). Bibliometrics: global gender disparities in science. *Nature*. 504, 211–213. doi: 10.1038/504211a

Leavy, S. (2018). "Gender bias in artificial intelligence: the need for diversity and gender theory in machine learning," in *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering*. New York, NY: ACM.

Leiner, D. J. (2019). SoSci Survey (version 3.1.06).

Lewis, D. D. (1998). Naive (Bayes) at forty: the independence assumption in information retrieval. *European Conference on Machine Learning.* Berlin: Springer. doi: 10.1007/BFb0026666

Lewis, R., Marwick, A. E., and Partin, W. C. (2021). We dissect stupidity and respond to it: response videos and networked harassment on YouTube. *Am. Behav. Sci.* 65, 735–756. doi: 10.1177/0002764221989781

Litvinenko, A. (2021). YouTube as alternative television in Russia: political videos during the presidential election campaign 2018. *Soc. Media Soc.* 7, 1–9. doi: 10.1177/2056305120984455

Liu, Y., Zhou, Y., Wen, S., and Tang, C. (2014). A strategy on selecting performance metrics for classifier evaluation. *Int. J. Mobile Comput. Multimedia Commun.* 6, 20–35. doi: 10.4018/IJMCMC.2014100102

Mardon, R., Molesworth, M., and Grigore, G. (2018). youtube beauty gurus and the emotional labour of tribal entrepreneurship. *J. Bus. Res.* 92, 443–454. doi: 10.1016/j.jbusres.2018.04.017

Mitchell, A., Simmons, K., Matsa, K. E., and Silver, L. (2018). *Publics Globally Want Unbiased News Coverage, but Are Divided on Whether Their News Media Deliver*. Washington, DC: Pew Research Center.

Molyneaux, H., O'Donnell, S., Gibson, K., and Singer, J. (2008). Exploring the gender divide on youtube: an analysis of the creation and reception of Vlogs. *Am. Commun. J.* 10, 1–14.

Montes-Vozmediano, M., García-Jiménez, A., and Menor-Sendra, J. (2018). Teen videos on YouTube: features and digital vulnerabilities. *Comunicar. Media Educ. Res. J.* 54, 61–69. doi: 10.3916/C54-2018-06

Moor, P. J., Heuvelman, A., and Verleur, R. (2010). Flaming on youtube. *Comput. Human Behav.* 26, 1536–1546. doi: 10.1016/j.chb.2010.05.023

Munger, K., and Phillips, J. (2022). Right-wing youtube: a supply and demand perspective. *Int. J. Press/Politics*. 27, 186–219. doi: 10.1177/1940161220964767

Muñoz Morcillo, J., Czurda, K., Geipel, A., and Robertson-von Trotha, C. Y. (2019). "Producers of Popular Science Web Videos – Between New Professionalism and Old Gender Issues," in *Proceedings Public Communication of Science and Technology Conference*. Available online at: https://arxiv.org/abs/1908.05572

Murphy, K. P. (2012). *Machine Learning - A Probabilistic Perspective*. Cambridge: The MIT Press.

Oakley, A. (2016). *Sex, Gender and Society*. London: Routledge.

Obadimu, A., Mead, E., Hussain, M. N., and Agarwal, N. (2019). "Identifying toxicity within youtube video comment," in *Social, Cultural, and Behavioral Modeling*, eds R. Thomson, H. Bisgin, C. Dancy, and A. Hyder (Cham: Springer).

Park, S., and Woo, J. (2019). Gender classification using sentiment analysis and deep learning in a health web forum. *Appl. Sci.* 9, 1–12. doi: 10.3390/app9061249

Parvandeh, S., Yeh, H.-W., Paulus, M. P., and McKinney, B. A. (2020). Consensus features nested cross-validation. *Bioinformatics* 36, 3093–3098. doi: 10.1093/bioinformatics/btaa046

Postigo, H. (2016). The socio-technical architecture of digital labor: converting play into YouTube money. *New Media Soc.* 18, 332–349. doi: 10.1177/1461444814541527

Pratama, B. Y., and Sarno, R. (2015). "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," in *2015 International Conference on Data and Software Engineering (ICoDSE)*. Yogyakarta: IEEE.

Raffo, J. (2021). *World Gender Name Dictionary 2.0 - Harvard Dataverse*. Available online at: https://dataverse.harvard.edu/dataset.xhtml?persistentId (accessed March 02, 2022).

Raun, T. (2018). Capitalizing intimacy: new subcultural forms of micro-celebrity strategies and affective labour on youtube. *Convergence* 24, 99–113. doi: 10.1177/1354856517736983

Regueira, U., Ferreiro, A. A., and Da-Vila, S. (2020). Women on youtube: representation and participation. *Comunicar. Media Educ. Res. J.* 63, 31–40. doi: 10.3916/C63-2020-03

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., and Meira, W. (2020). "Auditing radicalization pathways on YouTube," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Barcelona: ACM.

Rieder, B., Matamoros-Fernández, A., and Coromina, Ò. (2018). From ranking algorithms to 'ranking cultures' Investigating the modulation of visibility in YouTube search results. *Convergence* 24, 50–68. doi: 10.1177/1354856517736982

Rosenbusch, H., Evans, A. M., and Zeelenberg, M. (2019). Multilevel emotion transfer on YouTube: Disentangling the effects of emotional contagion and homophily on video audiences. *Soc. Psychol. Personal. Sci.* 10, 1028–1035. doi: 10.1177/1948550618820309

Schwemmer, C., and Ziewiecki, S. (2018). Social media sellout: the increasing role of product promotion on youtube. *Social Media Soci.* 4, 1–20. doi: 10.1177/2056305118786720

Scolari, C. A., and Fraticelli, D. (2018). The case of the top Spanish youtubers: emerging media subjects and discourse practices in the new media. *Ecology* 25, 496–515. doi: 10.1177/1354856517721807

Sebo, P. (2021). Using genderize.io to infer the gender of first names: how to improve the accuracy of the inference. *J. Med. Libr. Assoc.* 109, 609–612. doi: 10.5195/jmla.2021.1252

Seliya, N., Khoshgoftaar, T., and Van Hulse, J. (2009). "Aggregating performance metrics for classifier evaluation," in *IEEE International Conference on Information Reuse and Integration*. Las Vegas.

Sobande, F. (2017). Watching me watching you: black women in Britain on youtube. *Eur. J. Cult. Stud.* 20, 655–671. doi: 10.1177/1367549417733001

Soha, M., and McDowell, Z. J. (2016). Monetizing a meme: youtube, content ID, and the Harlem Shake. *Soc. Media Soc.* 2, 1–12. doi: 10.1177/2056305115623801

Sreberny, A. (2005). Gender, empowerment, and communication: looking backwards and forwards. *Int. Soc. Sci. J.* 57, 285–300. doi: 10.1111/j.1468-2451.2005.00551.x

Tang, Q., Gu, B., and Whinston, A. (2012). "Content contribution in social media: the case of YouTube," in *45th Hawaii International Conference on System Sciences*. Maui: IEEE.

Van Bruwaene, D., Huang, Q., and Inkpen, D. (2020). A multi-platform dataset for detecting cyberbullying in social media. *Lang. Resour. Eval.* 54, 851–874. doi: 10.1007/s10579-020-09488-3

van Dijk, J. (2020). *The Digital Divide*. Cambridge: Polity Press.

Wagner, C., Garcia, D., Jadidi, M., and Strohmaier, M. (2015). "It's a man's wikipedia? assessing gender inequality in an online encyclopedia," in *Proceedings of the Ninth International AAAI Conference on Web and Social Media*. Oxford: AAAI.

Wais, K. (2016). Gender prediction methods based on first names with genderizeR. *R. J.* 8, 17–37. doi: 10.32614/RJ-2016-002

Wegener, C., Prommer, E., and Linke, C. (2020). Gender representations on youtube. the exclusion of female diversity. *M/C J.* 23, 27–28. doi: 10.5204/mcj.2728

Weiss, G. M. (2013). "Foundations of Imbalanced Learning," in *Imbalanced Learning: Foundations, Algorithms, and Applications*, eds H. He and Y. Ma (Hoboken: John Wiley and Sons), 13–42.

Weissman, G. E., Ungar, L. H., Harhay, M. O., Courtright, K. R., and Halpern, S. D. (2019). Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical illness. *J. Biomed. Inform.* 89, 114–121. doi: 10.1016/j.jbi.2018.12.001

West, J., Jacquet, J., King, M., Correll, S., and Bergstrom, C. (2013). The role of gender in scholarly authorship. *PLoS ONE.* 8, e66212. doi: 10.1371/journal.pone.0066212

Wolf, A. (2000). Emotional expression online: gender differences in emoticon use. *Cyberpsychol. Behav.* 3, 827–833. doi: 10.1089/1094931005019 1809

Wolny, W. (2016). *Emotion Analysis of Twitter Data That Use Emoticons and Emoji Ideograms*. Available online at: https://aisel.aisnet.org/isd2014/proceedings2016/CreativitySupport/5/ (accessed February 02, 2022).

Wu, Y., Zhuang, Y., Long, X., Lin, F., and Xu, W. (2015). *Human Gender Classification: A Review*. Available online at: https://arxiv.org/pdf/1507.05122v1.pdf (accessed March 03, 2022).

Yan, X., and Yan, L. (2006). *Gender Classification of Weblog Authors. AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*. Available online at: www.aaai.org/Papers/Symposia/Spring/2006/SS-06-03/SS06-03-046.pdf (accessed February 02, 2022).

Zad, S., Heidari, M., Jones, J. H., and Uzuner, O. (2021). "A survey on concept-level sentiment analysis techniques of textual data," in *2021 IEEE World AI IoT Congress (AIIoT)*. Vancouver: IEEE.

Zeni, M., Miorandi, D., and Pellegrini, F. D. (2013). "YOUStatAnalyzer: a tool for analysing the dynamics of YouTube content popularity," in *Proceedings of the 7th International Conference on Performance Evaluation Methodologies and Tools*. Torino: ICST.

Zhou, R., Khemmarat, S., Gao, L., Wan, J., and Zhang, J. (2016). How youtube videos are discovered and its impact on video views. *Multimed. Tools Appl.* 75, 6035–6058. doi: 10.1007/s11042-015-3206-0

frontiers | Frontiers in Sociology

TYPE Original Research
PUBLISHED 10 October 2022
DOI 10.3389/fsoc.2022.883999

OPEN ACCESS

EDITED BY
Heinz Leitgöb,
Catholic University of Eichstätt-Ingolstadt, Germany

REVIEWED BY
Dimitri Prandner,
Johannes Kepler University of Linz, Austria
Stephan Poppe,
Leipzig University, Germany

*CORRESPONDENCE
Matthias Kuppler
matthias.kuppler@uni-siegen.de
Christoph Kern
c.kern@uni-mannheim.de

SPECIALTY SECTION
This article was submitted to Sociological Theory, a section of the journal Frontiers in Sociology

RECEIVED 25 February 2022
ACCEPTED 20 September 2022
PUBLISHED 10 October 2022

CITATION
Kuppler M, Kern C, Bach RL and Kreuter F (2022) From fair predictions to just decisions? Conceptualizing algorithmic fairness and distributive justice in the context of data-driven decision-making.
Front. Sociol. 7:883999.
doi: 10.3389/fsoc.2022.883999

From fair predictions to just decisions? Conceptualizing algorithmic fairness and distributive justice in the context of data-driven decision-making

Matthias Kuppler[1]*, Christoph Kern[2,3]*, Ruben L. Bach[2] and Frauke Kreuter[3,4]

[1]Department of Social Sciences, University of Siegen, Siegen, Germany, [2]School of Social Sciences, University of Mannheim, Mannheim, Germany, [3]Joint Program in Survey Methodology, University of Maryland, College Park, MD, United States, [4]Department of Statistics, LMU Munich, Munich, Germany

Prediction algorithms are regularly used to support and automate high-stakes policy decisions about the allocation of scarce public resources. However, data-driven decision-making raises problems of algorithmic fairness and justice. So far, fairness and justice are frequently conflated, with the consequence that distributive justice concerns are not addressed explicitly. In this paper, we approach this issue by distinguishing (a) fairness as a property of the algorithm used for the prediction task from (b) justice as a property of the allocation principle used for the decision task in data-driven decision-making. The distinction highlights the different logic underlying concerns about fairness and justice and permits a more systematic investigation of the interrelations between the two concepts. We propose a new notion of algorithmic fairness called error fairness which requires prediction errors to not differ systematically across individuals. Drawing on sociological and philosophical discourse on local justice, we present a principled way to include distributive justice concerns into data-driven decision-making. We propose that allocation principles are just if they adhere to well-justified distributive justice principles. Moving beyond the one-sided focus on algorithmic fairness, we thereby make a first step toward the explicit implementation of distributive justice into data-driven decision-making.

KEYWORDS

automation, prediction, algorithm, fairness, distributive justice

1. Introduction

In 2019, the United States signed into law the Foundations of Evidence-based Policy Act (Hart and Yohannes, 2019) which requires government agencies to exploit available evidence and data when making policy decisions. Similar initiatives are under way in other countries. The German government, for example, pledged over two hundred million euro to build data labs in every ministry to improve decision-making and

bring data-driven evidence into everyday policy-making (Engler, 2022b). Prediction algorithms play an increasingly important role in meeting evidence-based policy-making goals in settings where policies affect the allocation of social benefits and interventions to individuals. In these settings, algorithms are used to predict the likelihood of a risk in order to target an intervention or help.

Carton et al. (2016), for instance, predicted the risk of adverse behavior among police officers and used the predicted risk scores to prioritize preventive training and counseling. The New Zealand government used prediction algorithms and historic data about families to identify new-born children who are at high risk for maltreatment and, hence, are prioritized for preventive services (New Zealand Ministry of Social Development, 2014). More recently, prediction models were used to support COVID-19 prevention and treatment decisions in Israel (Barda et al., 2020).

The adoption of prediction algorithms reflects a long-standing trend toward less discretionary, data-driven decision procedures (Elster, 1992). Data-driven approaches promise to render decision-making processes more accurate and evidence-based and, by limiting decision-maker discretion, less susceptible to human biases and manipulation (Lepri et al., 2018). In domains with profound impacts on life chances, including decisions regarding policing (Alikhademi et al., 2021), welfare benefits (Desiere et al., 2019), and criminal justice (Angwin et al., 2016), concerns are raised that prediction algorithms, despite the gained efficiencies, can inherit human biases and perpetuate unfair discrimination against vulnerable and historically disadvantaged groups (Barocas and Selbst, 2016). Such perpetuation is particularly likely when prediction algorithms are based on data where (a) key groups are misrepresented or missing, (b) outcomes are systematically mislabeled, and (c) past discriminatory behavior is recorded and creates historical bias (Rodolfa et al., 2021). These concerns are fundamental to the discussion of an AI Act for the European Union (Engler, 2022a).

To address these challenges and to guide the design of non-discriminatory prediction algorithms, the research community developed formal fairness definitions—called *fairness metrics*—that quantify the extent to which model predictions satisfy various notions of fairness (Makhlouf et al., 2020; Mitchell et al., 2021). Independence, for instance, states that predictions are fair if they are statistically independent from a pre-defined set of protected attributes like sex or disability. Disagreement exists over which metric captures the underlying concern about fairness best. The debate is exacerbated by the fact that some fairness definitions are incompatible, such that a prediction model cannot satisfy all definitions simultaneously (Chouldechova, 2016; Kleinberg et al., 2016). Recent research attempted to resolve this conundrum by identifying the moral assumptions underlying the different fairness definitions and delineating the situations in which certain assumptions are (not) justified (Heidari et al., 2019; Friedler et al., 2021).

In this paper, we propose an alternative approach to fairness and justice in data-driven decision-making. Existing fairness approaches tend to mix technical concerns about the statistical properties of algorithmic predictions with moral concerns about the justice of decisions that are based on these predictions. To highlight the distinction between technical and moral concerns, we define fairness as a property of the prediction algorithm and justice as a property of the decision rule. From this perspective, there is little room for moral debate at the prediction step. Predictions should represent the true underlying values of the prediction target as accurately as possible for all candidates to which the prediction algorithm is applied. No candidate should have a disproportionate risk of an erroneous prediction that systematically depends on her characteristics. We call this notion *error fairness* and define it as the requirement that prediction errors are not systematically related to observed and unobserved features of the candidates. While this perspective is immanent in the (multi-group) fairness notions of Kim et al. (2019) and Hebert-Johnson et al. (2018), we highlight that the common group-based approach to algorithmic fairness is unable to guarantee error fairness. Suggestions for metrics that capture error fairness are made.

We define the decision step as a problem of local justice (Elster, 1992). Local justice focuses on the principles that organizations use to allocate benefits and burdens—a focus that aligns well with the scope of data-driven decision-making. The selection of allocation principles is informed by middle-range distributive justice principles. We consider four justice principles: equality, desert, need, and efficiency (Deutsch, 1975; Konow, 2003; Törnblom and Kazemi, 2015). Each justice principle defines a class of criteria that should guide the allocation of benefits and burdens.

We make two contributions to the literature on algorithmic fairness and justice. First, we clarify the relation between fairness and justice in data-driven decision-making and provide a clear definition of both concepts. Second, we provide an overview of distributive justice principles and a recipe for implementing the principles into the decision-making pipeline. Taken together, our approach guides the design of data-driven decision procedures that go from fair predictions to just decisions.

The argument proceeds as follows: Section 2 defines the class of decision problems that we deal with in this paper and introduces our definitions of justice and fairness. Section 3 elaborates on the problem of justice in data-driven decision-making from within the framework of local justice. Section 4 discusses the problem of fairness in data-driven decision-making and introduces the notion of error fairness. Section 5 provides a broader picture of the problem of bias in data-driven decision-making that is inspired by the distinction between

justice and fairness. Section 6 concludes with a discussion of the practical implications and limitations of our approach.

2. Problem statement: Data-driven decisions, justice, and fairness

2.1. Decision problem

Our entire argument deals with the following decision problem: Consider an institution with a fixed amount $X \in \mathbb{N}_+$ of a good that it can allocate among a fixed number of $i = 1, \ldots, n$ candidates. The institution must decide which candidates should receive a unit the good. Goods are material and immaterial things that can be attached to or owned by the candidates. Goods can be valued positively (as something one would like to have) or negatively (as something one would like to avoid). Positively valued goods are benefits, negatively valued goods are burdens. Because of the symmetry between benefits and burdens (exemption from a burden is a benefit and vice versa), we use the general term *good* in the following[1].

Candidates are the actors who are eligible for the good. Candidates can be individual (e.g., humans or animals) and corporate (e.g., organizations or sub-units of organizations) actors. The pool of eligible candidates is usually specific to the allocating institution. For instance, not every citizen is eligible for participation in the labor market activation programs allocated by a public employment agency. Similarly, only sub-units of a firm are eligible for the allocation of resources by the central governance unit of the firm.

The decision problem is further characterized by scarcity, indivisibility, homogeneity, and rivalry. *Scarcity* means that the number of candidates (demand for the good) exceeds the number of units of the good that can be allocated (supply of the good). Scarcity may be natural (there is no way to increase supply) or artificial (supply could be increased at the cost of decreasing supply of another good). Paintings by Pablo Picasso are a naturally scarce good. Prison sentencing is an artificially scarce good. Courts could exempt every defendant from the burden of a prison sentence at the cost of reducing the overall safety of society. *Indivisibility* means that the good comes in fixed units that cannot be sub-divided any further—at least not without losing value or getting destroyed. Kidney transplants, for instance, are indivisible. One cannot (at least currently) transplant one kidney into two patients. *Homogeneity* means that only one version of the good exists and that any two units of the good are indistinguishable. *Rivalry* means that ownership of the good by one candidate A precludes ownership of the good by any other candidate B, C, … now and in the future unless the good is re-allocated.

Finally, we focus on *binary decisions*. For each candidate, the institution decides between two options: allocate one unit of the good to the candidate (positive decision) or allocate no unit of the good to the candidate (negative decision). The decision problem amounts to selecting the subset $n^* \subset n$ of candidates who receive the good. In this paper, we do not consider decisions about the number of units of the good allocated to each candidate. In principle, however, our approach could be extended to such decisions.

The task of the institution is to formulate an *allocation principle*. An allocation principle is a rule that defines how goods are allocated to candidates. The principle defines a set of decision-relevant criteria and specifies the relationship between the criteria and the allocation of goods. In most cases, the decision-relevant criteria are attributes of the candidates. Allocation principles differ in the amount of discretion awarded to human decision-makers, varying from informal open-ended (high discretion) to formal rule-based (low discretion) principles.

In this paper, we focus on formal rule-based allocation principles because the impetus for implementing data-driven decision-making is usually a desire to reduce human discretion. For the most part, we rely on *ranking-based allocation principles*. Candidates are brought into a rank order based on the value they have on the decision-relevant criterion. At the top of the rank order are the candidates who, according to their value on the decision criterion, have the strongest claim to the good. If there are X units of the good, each of the X top-ranked candidates receives one unit of the good[2]. A bank, for instance, might allocate loans (the good) based on candidates' history of loan repayment (decision criterion). If the bank can allocate $X = 10$ loans among $n = 100$ applicants, it will allocate the loan to the ten candidates with the best repayment history.

The proposed definition captures a large class of decision-problems that have been subjected to data-driven approaches. Examples include decisions by banks to grant or deny a loan (Kozodoi et al., 2021), decisions by courts to grant or deny probation (Metz and Satariano, 2020; Završnik, 2021), decisions by public employment agencies to grant or deny participation in active labor market programs (Desiere et al., 2019), and decisions by hospitals to grant or deny certain types of medical

1 We make the simplifying assumption that the good is valued uniformly across all candidates. All candidates define the good either as a benefit or as a burden and candidates do not differ in the degree to which they value the good as either positive or negative. We believe that these simplifications are justified because real-world institutions typically lack information about inter-individual differences in the valuation of goods. Even if information on candidate preferences is available, its relevance as a decision criterion for the allocation is usually low.

2 We could also use an admission procedure (Elster, 1992) whereby each candidates whose value on the decision criterion surpasses a threshold value receives a unit of the good. Such a procedure, however, makes it difficult for an institution to plan because the number of candidates who surpass the threshold can fluctuate strongly over time.

treatment (Obermeyer et al., 2019). In each case, an institution has to formulate an allocation principle that regulates which candidates receive a unit of the good and which candidates do not.

2.2. Data-driven decision pipeline: Prediction and decision step

How could a solution to the decision problem look like? We already sketched half of the answer: Institutions formulate an allocation principle that regulates how goods are allocated across the candidates. This is the *decision step* in the decision pipeline: The institution uses the decision criterion to select the candidates who receive the good. We implicitly assumed that the decision criterion—the input to the allocation principle—is observed by the institution at the time point of the decision. This is not always the case, however. The decision criterion might be unobserved at decision time because it materializes only in the future or because it is too costly for the institution to measure it for each candidate.

Indeed, many instance of data-driven decision-making are motivated by the fact that the decision criterion is unobserved (or even unobservable) at decision time. Courts would like to base their decision to grant or deny probation on the knowledge about the future criminal behavior of the defendant. Banks would like to base their decisions to grant or deny loans on the knowledge about the future repayment behavior of the loan applicant. Public employment agencies would like to base decisions to grant or deny access to support programs on the knowledge about whether the job-seeker would find re-employment without further support. In all these cases, the decision criterion (criminal behavior, repayment behavior, re-employment) lies in the future and, hence, is unobservable at decision time.

If (and only if) the decision criterion is unobserved at decision time, the decision pipeline is extended by a *prediction step*. In the prediction step, the institution uses observed attributes of the candidate to predict the unobserved decision criterion. Banks would, for instance, use the past repayment history of the candidate (observed attribute) to predict the probability that the candidate will repay the next loan (decision criterion).

The resulting two-step decision pipeline is shown in Figure 1. First, the decision criterion is predicted from the observed candidate attributes (prediction step). Then, an allocation decision is made based on the predicted criterion (decision step; Loi et al., 2021; Mitchell et al., 2021).

2.2.1. Decision step

The task of the institution is to find an allocation principle that defines how to select the subset of candidates who receive the good. Let y_i denote the value of the decision-relevant criterion for the i-th candidate. The criterion can be categorical or continuous[3]. Let $d_i \in \{0, 1\}$ be the allocation decision that records whether the i-th candidate receives the good ($d_i = 1$) or not ($d_i = 0$). Let Y and D be random variables of the values for a candidate randomly drawn from the population of candidates. The allocation principle is a function $\theta : Y \rightarrow D$ that maps the decision-relevant criterion Y onto the allocation decision D. Applying the allocation principle $\theta(y_i) = d_i$ gives the decision for the i-th individual with value y_i on the decision criterion. In words, the allocation principle states: Allocate the good to the i-th candidate if and only if the candidate's value on the decision criterion qualifies her for the good. If the decision criterion is not observed at decision time, $\theta(\hat{y}_i) = \hat{d}_i$ gives the decision for the i-th candidate given their predicted value $\hat{y}_i$ on the decision criterion. $\hat{Y}$ and $\hat{D}$ are the predicted criterion and the prediction-based allocation decision for a candidate randomly drawn from the population.

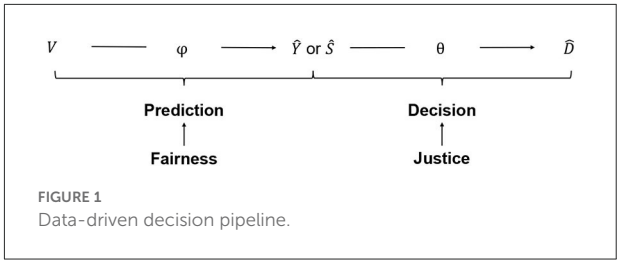


FIGURE 1
Data-driven decision pipeline.

2.2.2. Prediction step

The prediction step is a classification problem for categorical and a regression problem for continuous decision criteria. The prediction task makes use of a training set of $j = 1, \ldots, m$ candidates for which the criterion is observed. Let $v_{j.}$ denote the values of the observed features for the j-th candidate. We denote additional features that are unobserved, but potentially relevant, as $u_{j.}$. Let V and U be random variables for the features of a candidate. The prediction task is: Given a training data set of candidates of the form $\{(v_{1.}, y_1), \ldots, (v_{m.}, y_m)\}$, find a function $\phi : V \rightarrow Y$ that maps the observed features onto the criterion. The function ϕ that is estimated in the training data set is then used to obtain predictions of the criterion value for the candidates of interest at decision time. For a continuous decision criterion, $\phi(v_{i.}) = \hat{y}_i$ returns the predicted value of the i-th candidate's criterion value. Candidates can be ranked according to their predicted value of the decision criterion. For a categorical decision criterion, $\phi(v_{i.}) = \hat{s}_i = \hat{P}(y_i = 1)$ returns the score $\hat{s}_i$, the predicted probability that the i-th candidate possesses the decision criterion. The scores can be used to rank

3 In the categorical case, we are focusing on binary (or binarized) decision criteria.

candidates according to their predicted probability of possessing the criterion. Let $\hat{S}$ be the score for a random candidate.

2.3. Algorithmic fairness and justice

The remainder of the paper explicates the implications of a consequent distinction between prediction and decision for the design and evaluation of data-driven decision procedures. *Our main argument is that the distinction between prediction and decision implies a corresponding distinction between the concepts of fairness and justice*. We propose that fairness is a property of the prediction algorithm and is only relevant at the prediction step. Justice is a property of the allocation principle and only relevant at the decision step. We propose the following definitions of justice and fairness in the context of data-driven decision-making.

Definition 1 (algorithmic justice). *An allocation principle is called just iff it approximates a well-justified distributive justice principle.*

Definition 2 (algorithmic fairness). *A prediction algorithm is called fair iff its predictions satisfy a well-justified substantive fairness definition.*

The definition of algorithmic justice is elaborated in Section 3. Working within the framework of local justice (Elster, 1992), we show how the design of allocation principles can be guided by middle-range justice principles. The definition of algorithmic fairness is further discussed in Section 4. At this point, we note that our definition of algorithmic fairness is kept at a very general level and, by design, can accommodate a large range of substantive fairness definitions that have been proposed in the literature (Mitchell et al., 2021). Substantive fairness definitions formally describe the concrete properties that algorithmic predictions must satisfy in order to be considered fair. Independence (also called Statistical Parity), for instance, requires that predictions are statistically independent from protected attributes like gender and ethnicity.

Note that justice and fairness are indeed separate concepts. A just allocation principle does not guarantee a fair prediction algorithm and vice versa. For instance, the final outcome of the data-driven decision process—the decision to allocate the good to a candidate or not—can be just but unfair. The decision that a candidate does not receive a loan might be just because the bank's allocation principle to choose the candidates with the highest predicted probability to repay approximates the well-justified desert-based justice principle. At the same time, the (prediction-based) decision might be unfair because the algorithm that predicts the repayment probability systematically under-predicts the repayment probability of women compared to men. In the same vein, the outcome of the data-driven decision process might be fair but unjust. It becomes obvious that we need both: Data-driven decision-making should be fair and just. Importantly, there is no conflict between fairness and justice. We actually can have both and do not need to trade off one against the other.

3. Just decisions

3.1. Local justice

Local justice is concerned with the allocation of goods to individuals by relatively autonomous meso-level institutions (Elster, 1992; Schmidt, 1992b). Institutions are formal organizations that, in fulfilling their respective function, make decisions about the allocation of goods (Schmidt, 1992a). Local justice problems are local in the double sense that (a) they are solved de-centrally by relatively autonomous institutions and (b) their solutions are highly context-dependent and vary across sectors or "localities" within one society. Global justice, in contrast, is concerned with the overall design of the basic structure of society (Rawls, 1971), the "constitutional ground rules of a social, political, and economic order" (Schmidt, 1994, 322). The class of decision problems discussed in this paper clearly falls within the scope of local justice.

The building blocks of local justice are (a) the *good* that is allocated, (b) the individuals (*candidates*) to whom the good can potentially be allocated, (c) some functional rule (*allocation principle*) that specifies how goods are allocated to candidates, and (d) a normative standard (*distributive justice principle*) against which the resulting allocation is evaluated (Cohen, 1987). Goods, candidates, and allocation principles were already introduced as part of the decision problem in Section 2.1. The following discussion, therefore, focuses on the distributive justice principles and how they can guide the selection of allocation principles.

3.1.1. Distributive justice principles

Distributive justice principles are well-justified accounts of how goods should be allocated. The justice principles define an ideal standard against which non-ideal allocation principles—that have to operate under non-ideal real-world conditions—are evaluated. Generally, we wish to select the allocation principle or combination of allocation principles that best approximates our preferred distributive justice principle. Our focus lies on what we call *middle-range distributive justice principles*. Middle-range principles are general enough to apply across multiple empirical cases. At the same time, they are not as general as global justice theories that aim to regulate the basic structure of society but give little guidance for the resolution of concrete allocation problems. Examples of global justice theory include *A Theory of Justice* (Rawls, 1971), *Anarchy, State, and Utopia*

(Nozick, 1974), and the hypothetical insurance scheme laid out by *Luck Egalitarianism* (Dworkin, 1981).

In the next section, we discuss four middle-range distributive justice principles: equality, desert, need, and efficiency (Deutsch, 1975; Konow, 2003; Törnblom and Kazemi, 2015). The principles draw inspiration from broader distributive justice theories, namely egalitarianism (Arneson, 2013), desert-based justice (Feldman and Skow, 2020), sufficiency (Brock, 2018) and prioritarianism (Parfit, 1997; Adler and Holtug, 2019), and consequentialism (Sinnott-Armstrong, 2021), respectively.

Distributive justice principles are context-dependent, pluralistic, and contested (Schmidt, 1994; Konow, 2003). *Context dependency* means that the selection of justice principles is guided and justified by the empirical facts that characterize the concrete allocation problem. No distributive justice principle satisfies justified moral expectations in every empirical case. *Pluralism* emphasizes that there are allocation problems for which multiple (potentially conflicting) justice principles are equally well-justified such that a compromise between principles is necessary. The *contestedness* of justice principles highlights that the allocating institution is often subject to demands other than justice, such as profitability or public preferences, that preclude the implementation of the preferred justice principle.

3.1.2. From justice principle to allocation principle

The process of formulating an allocation principle is akin to the operationalization of a latent construct for empirical research. Each middle-range distributive justice principle identifies a distinct latent construct—equality, desert, need, or efficiency—that should guide the allocation of goods. Formulating an allocation principle amounts to finding a manifest indicator for this latent construct. The indicator is a context-fitting interpretation of the justice principle in the sense that it transports the general intention of the justice principle into the specific allocation context. For instance, life expectancy might be a manifest indicator for the latent concept *need* in the context of allocating kidney transplants. Repayment probability might be a manifest indicator for the latent concept *desert* in the context of allocating loans.

3.1.3. Choosing allocation principles

Local justice is a descriptive (and partly explanatory) rather than normative approach (Elster, 1992). Local justice has three broad goals: (1) Cataloging the allocation principles implemented by existing institutions. (2) Identifying the mechanisms that lead to the implementation of certain types of principles in certain types of allocation problems. (3) Describing the typical distributive consequences associated with each principle. The distributive consequences encompass the direct results (how gets what?) and also the indirect (unintended) incentive effects of an allocation principle[4]. Local justice does not provide a normative argument for why a certain principle should be chosen. It does not formulate a moral justification—in the sense of a rational defense of the principle to all candidates who are eventually affected by it—for why a certain principle should be chosen.

In this realist (rather than idealist or normative) perspective, local justice shows that the selection of allocation principles results from complex negotiation and bargaining between the allocating institution, political actors, the candidate population, and the overall public represented by the media (Elster, 1992; Schmidt, 1992a). The actors are (at least partially) aware of the distributive consequences of different allocation principles and formulate their preferences accordingly. Which allocation principle is selected depends on the relative bargaining power of the actors. Classically, bargaining power is a function of actors' relative dependence on each other. The less dependent an actor is on the others for realizing her preferences, the higher her bargaining power. Bargaining is also a discourse, however, in which the best argument may win irrespective of the nominal bargaining power of the actor who formulates the argument. The actors can, therefore, be expected to leverage moral arguments and justifications that support their preferred allocation principle. These arguments and justifications are drawn from the middle-range distributive justice principles. An exact explanation of the negotiation process underlying the selection of allocation principles is not the purpose of the paper. Suggestions for the organization of such negotiation processes are formulated in the literature on impact assessment frameworks for data-driven decision systems (Selbst, 2018; Mantelero, 2018; Metcalf et al., 2021).

We also adopt the realist approach. That is, we do not provide a universal argument for why a certain type of allocation principle should always be selected for a certain type of allocation problem. It might turn out that such a universal argument does not exist. We, at least, are not aware of such an argument. Instead, we sketch the likely distributive consequences of each principle and present the main arguments

4 Allocation principles produce (unintended) incentive effects when the principle is public knowledge and the decision-relevant criteria can be modified by the candidates. Positive incentive effects arise when attempts to acquire the decision-relevant criteria induce socially valuable behavior from the candidates (e.g., if kidney transplants are assigned preferably to non-smokers). Negative incentive effects occur when the induced behavior is socially harmful (e.g., if the military policy to draft only young men who can fire a gun induces young men to cut off their index fingers). Public knowledge of the decision-relevant criteria invites candidates to game the system by misrepresenting their private attributes.

that justify the implementation of the principle. The ultimate selection of an allocation principle is the task of the actors who are embedded in the allocation context. Our hope is that a better understanding of the allocation principles helps these actors to make better decisions.

3.2. Distributive justice principles

Four middle-range distributive justice principles are discussed: equality (E), desert (D), need (N), and efficiency (EFF) (Deutsch, 1975; Konow, 2003; Törnblom and Kazemi, 2015). Table 1 provides a short definition of each principle. Note that sub-forms of the main justice principles exist (Törnblom and Kazemi, 2015). While we try to present a comprehensive list of justice principles and their sub-forms, we do not claim that our list is exhaustive. The principles defined below are our interpretations of the underlying middle-range distributive justice theories. Other interpretations are certainly possible and may prove to better reflect the central concerns of the underlying theories. For now, however, our definitions should provide a useful starting point.

3.2.1. Equality

The equality (E) principle requires either equal treatment or equal outcomes across candidates. Equal treatment (Et) demands that all candidates receive the same amount of the good. When the good is scarce and indivisible, it is impossible to implement this principle. Lotteries are a second-best approximation in this case: Each candidate has the same probability $p = X/n$ to receive the good (Elster, 1992). The decision criterion Y is the selection result of the lottery. The corresponding allocation principle θ is: Allocate the good to the i-th candidate if and only if the candidate is selected by the lottery.

Equal outcomes (Eo) demands that candidates have the same post-allocation outcome, i.e., the same outcome after the allocation decision is implemented. This raises the thorny problems of (a) defining the relevant outcome, (b) estimating the—potentially inter-individually varying—effect of the good on the outcome, and (c) defining a metric that measures inequality in outcomes. The decision criterion Y is the post-allocation outcome. The corresponding allocation principle θ is: Allocate the good to the i-th candidate if and only if this allocation is part of an overall allocation scheme that minimizes inequality (as measured by the metric) in the distribution of post-allocation outcomes across candidates[5]. If multiple allocation schemes minimize inequality to the same extent, a rule must be defined to select one of the schemes (e.g., a random draw).

Figure 2 illustrates the equal outcome (Eo) principle. There are three candidates (A, B, and C), shown on the X-axis. The Y-axis shows the outcomes of the candidates. The height of the blue bar indicates the outcome of the candidate before the allocation decision is made. The height of the orange bar indicates the increase in the outcome that the candidate would experience if the good is allocated to her. The impact of the good on the outcome differs across the candidates. The outcome of candidate A would increase by the smallest amount, the outcome of candidate C by the largest amount. The combined height of the orange bar and the blue bar indicates the level of the outcome after the allocation—under the condition that the candidate receives the good. Imagine that there are $X = 2$ units of the good that we can allocate to the three candidates. Because we do not want to be wasteful (and to avoid the leveling-down objection), we allocate both units of the good, even if a more equal state could be reached if we allocate fewer units of the good[6]. There are three possibilities to allocate the two goods to the three candidates: (1) Allocate to A and B, (2) allocate to A and C, and (3) allocate to B and C. Following the equal outcome (Eo) principle, we choose the allocation scheme that minimizes inequality in the post-allocation outcomes. We measure inequality with the Gini-coefficient. The Gini-coefficient varies between 0 (perfect equality: every candidate has the same outcome) and 1 (perfect inequality: only one candidate has a positive outcome, the outcome of the other candidates is zero). The Gini-coefficients for the allocation schemes are 0.33, 0.05, and 0.24, respectively. Accordingly, the equal opportunity (Eo) principle recommends allocating the good to candidate A and candidate C.

The equality principle is justified by the egalitarian ideal that all candidates are moral equals—at least with respect to

5 Example: The outcome is the probability of re-employment among unemployed job-seekers. The good is participation in an active labor market training program that positively affects re-employment probability. The inequality metric is the Gini-coefficient applied to the distribution of post-allocation re-employment probability. The goal is to find the allocation of program participation to job-seekers that minimizes the Gini-coefficient. The allocation principle states that all candidates who receive program participation in this allocation scheme should actually receive program participation.

6 The leveling-down objection is best understood *via* an example. Let (A, B, C) be the outcomes for candidates A, B, and C that are generated by an allocation scheme. The choice is between two such allocation schemes: (1, 1, 1) and (10, 20, 20). The equal outcome principle, in its strictest sense, would force us to choose the first allocation scheme because it is clearly more equal than the second. The leveling-down objection is that we could make everyone better-off by choosing the second allocation scheme, even the candidate who receives the lowest outcome. The equal outcome principle forces us to make everyone worse-off than they could be even if this reduces inequality only by a small amount.

10.3389/fsoc.2022.883999

TABLE 1 Middle-range distributive justice principles.

Justice principle	Sub-form	Decision criterion	Allocation principle
Equality (E)	Equality of treatment (Et)	Selection by lottery	Allocate the good to the candidate if and only if the candidate is selected by an unbiased lottery.
	Equality of outcome (Eo)	Post-allocation outcome	Allocate the good to the candidate if and only if this allocation is part of an overall allocation scheme that minimizes inequality in the distribution of post-allocation outcomes.
Desert(D)	Productive contribution (Dp)	Past or future contribution to cooperative production of the good	Rank candidates according to their desert in descending order. Allocate the good to the candidate if and only if the candidate is among the top-ranked candidates.
	Effort (De)	Effort expended in the cooperative production of the good	
	Costs and Sacrifice (Dc)	Costs incurred in the cooperative production of the good	
Need (N)	Biological (Nbi)	Need for goods essential for survival	Rank candidates according to their need in descending order. Allocate the good to the candidate if and only if the candidate is among the top-ranked candidates.
	Basic (Nba)	Need for goods essential for recognizably human life	
	Functional (Nf)	Need for goods essential for fulfilling social roles	
Efficiency (EFF)	–	Outcome-increment realized by allocating good to candidate	Rank candidates according to their outcome-increment in descending order. Allocate the good to the candidate if and only if the candidate is among the top-ranked candidates.

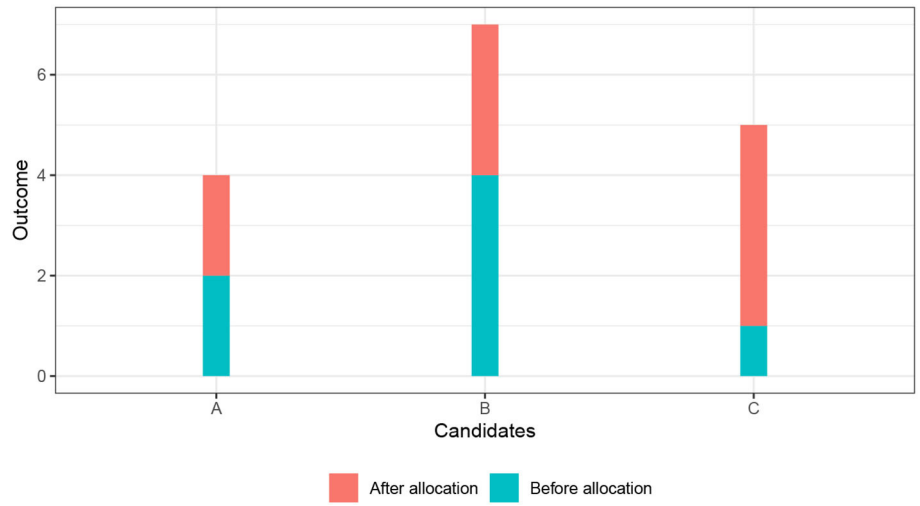


FIGURE 2
Illustration of equal outcomes principle.

the factors that are morally relevant to the allocation problem and should therefore affect its outcome (Gosepath, 2011; Arneson, 2013). Equality is the baseline principle whenever no candidate can make an inter-personally comprehensible and acceptable claim to more than an equal share. Such claims might refer to personal need and desert or to gains in efficiency

realized by allocating the good to a specific candidate. The equality principle likely produces negative incentive effects: Candidates are not held responsible for their actions and—especially in the case of equality of outcomes—can count on compensation for socially harmful actions that lower their pre-allocation outcomes.

3.2.2. Desert

The desert (D) principle ties the allocation of goods to so-called desert-bases (Moriarty, 2018; Feldman and Skow, 2020). Desert-bases are properties of an individual by virtue of which the individual can make a claim to the good. The decision criterion Y is the desert-base deemed relevant in the allocation context. The allocation principle θ is: Rank the candidates according to their desert Y in descending order. Allocate the good to the i-th candidate if and only if the candidate is among the X top-ranked candidates. If desert is unobserved at decision time, candidates are ranked based on either the predicted value $\hat{Y}$ for continuous desert-bases or the score $\hat{S}$ for categorical desert-bases.

Not all properties qualify as desert-bases (Feldman and Skow, 2020). Desert-bases generally reflect socially beneficial properties and actions. The bases should be morally-relevant, i.e., they should stand in some relation to the good. It should be possible to evaluate desert-bases as good or bad. Only then can we say that a candidate has a stronger claim to the benefit (burden) because she has a property or performed an action that is deemed good (bad). Desert-bases can be limited to properties and actions for which the candidate can be reasonably held responsible (Arneson, 2015). Candidates cannot be held responsible for things that are not under their control (Lippert-Rasmussen, 2018) or that result from brute luck (Dworkin, 1981), i.e., from outcomes of gambles that candidates could not anticipate or could not decline because they lacked a reasonable alternative.

In contexts concerned with the allocation of goods produced *via* cooperation, three desert-bases are often proposed: past or future contribution of the individual to the production of the good (Dp), effort expended in the production process (De), and costs or sacrifices incurred due to the production activity (Dc) (Lamont and Favor, 2017).

Figure 3 illustrates the desert principle. Consider a university department that wants to allocate job-interviews for open tenure-track positions (the good) based on desert (latent decision criterion). Here, desert is operationalized as the future h-index (manifest decision criterion) of a candidate. The h-index quantifies the scientific impact of a researcher based on her number publications and the number of citations that her publications received (Hirsch, 2005). An h-index of k indicates that the k most cited papers of a researcher received at least k citations. The university department wants to invite 10% of the candidates to job-interviews. We consider two scenarios: (1) The department invites candidates whose predicted h-index is in the top-10% of the candidate distribution. (2) The department invites candidates whose predicted probability to become a high-performer is among the top-10% of the candidate distribution. High-performers are candidates whose predicted h-index is above the 75% percentile of the candidate distribution. The first scenario describes a regression problem and the allocation decision is based on the predicted value $\hat{Y}$ of the h-index. The second scenario describes a classification problem (candidate is either a high-performer or not) and the allocation decision is based on the predicted score $\hat{S}$ of becoming a high-performer. The left panel of Figure 3 shows the density plot for the h-index prediction, the right panel shows the density plot for the high-performer prediction[7]. In both cases, the candidates whose predicted value (either $\hat{Y}$ or $\hat{S}$) falls in the red area to the right of the dashed line are invited to the job-interview. According to the chosen indicator of desert, these are the candidates with the strongest claim to the good.

The desert principle is justified whenever reasonable desert-bases exist and are not overridden by other concerns like need or efficiency. Then, candidates can make an interpersonally comprehensible and acceptable claim to more than an equal share of the good that is based on their personal desert. The egalitarian ideal of the candidates' moral equality does not prescribe equality *per se* (Gosepath, 2011). Treating candidates as moral equals can also mean to take their actions and responsibility serious and to allocate goods accordingly (Moriarty, 2018). The desert principle can produce positive incentive effects: Candidates are rewarded for socially productive behavior and punished for harmful behavior. It can be difficult, however, to identify desert-bases for which candidates can be truly held responsible.

3.2.3. Need

The need (N) principle ties the allocation of goods to need claims (Brock, 2018). Need claims have the following structure: The candidate requires the good in order to realize a certain end-state. Following prioritarianism (Parfit, 1997; Holtug, 2007), the strength of a need claim to the good increases the worse-off the candidate is, i.e., the farther away the candidate is from achieving the end-state. The decision criterion Y is the strength of the need claim. Need claims grow in strength the farther away the candidate is from the end-state prior to the allocation. The allocation principle θ is: Rank the candidates according to the strength of their need claim Y. Allocate the good to the i-th candidate if and only if the candidate is among the X top-ranked

7 The estimates are based on a Gradient Boosting Machine trained on a sample of Computer Science researchers active from 1993 to 2016. For more information on the data, see Weihs and Etzioni (2017). More information on the prediction model and an evaluation of the prediction fairness can be found in Kuppler (2022).

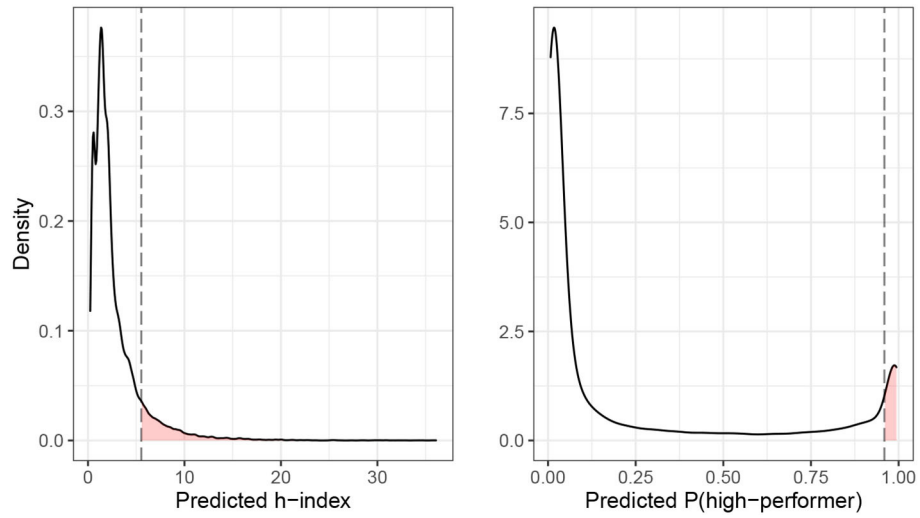


FIGURE 3
Illustration of desert principle.

candidates. If need is unobserved at decision time, candidates are ranked based on either predicted value $\hat{Y}$ for continuous end-states or the score $\hat{S}$ for categorical end-states.

Three classes of end-states generally qualify as bases for need claims (Törnblom and Kazemi, 2015; Brock, 2018): Biological needs (Nbi) are states that are essential to survival. Basic needs (Nba) are states that are essential to lead a recognizably human life, according to the standards of one's society. Functional needs (Nf) are states that enable the candidates to fulfill their social roles. Need justifies claims to the good irrespective of whether the candidate is responsible or not for failing to achieve the end-state.

Going back to the university department example in Figure 3, we could imagine that the department wants to allocate a career support program (the good) to its current employees. The department decides to allocate the program based on need, operationalized as the predicted h-index of the employees. The department could allocate the program to the 10% of employees with the lowest predicted h-index ($\hat{Y}$) or to the 10% of employees with the lowest predicted probability of becoming a high-performer ($\hat{S}$). The idea is that a high h-index (continuous end-state) and being a high-performer (categorical end-state) are valuable end-states for researchers and that the support program helps researchers to realize these end-state. Employees with a low h-index are farther away from realizing the end-state and, therefore, have a stronger need claim to the support program.

The need principle is justified whenever reasonable need claims exist and are not overridden by other concerns like desert or efficiency. Then, candidates can make an inter-personally comprehensible and acceptable claim to more than an equal share of the good that is based on their personal need. Treating all candidates as moral equals (Gosepath, 2011) does not necessarily mean to equalize outcomes but can also mean to work toward a situation in which all candidates can at least fulfill their biological, basic, and functional needs (Brock, 2018). Meeting needs is socially beneficial as it enables the candidates to function as productive members of society. The need principle can produce negative incentive effects: Candidates are not held responsible for their need and, hence, are not punished if they squander the allocated good because they expect additional transfers in the future.

3.2.4. Efficiency

The efficiency (EFF) principle allocates goods to promote a valued outcome (Elster, 1992). Goods are allocated across candidates in a way that maximizes the degree to which the outcome is attained in the aggregate[8]. The decision criterion Y is the increment in outcome-attainment that is realized by allocating the good to a specific candidate. In other words: The criterion is the candidate-specific effect of the good on the outcome. The allocation principle θ is: Rank, in descending order, the candidates according to the increment in outcome-attainment Y that is achieved by allocating the good to the candidate. Allocate the good to the i-th candidate if and only if the candidate is among the X-top ranked candidates[9].

8 Maximization applies if attainment of the outcome is beneficial. Minimization applies if the outcome is harmful. Example: With a given supply of kidney transplants, we wish to *maximize* the years of life saved. With a fixed supply of food, we wish to *minimize* hunger.

9 Continued example: We allocate the kidney transplants to the candidates in whom the transplant will produce the largest increment in

Figure 2 that we used to illustrate the equal outcome principle can also illustrate the efficiency principle. Remember that the effect of the good on the outcome (the height of the orange bar) differed across the three candidates A, B, and C. The effect was strongest for candidate C (four outcome units), followed by candidate B (three outcomes units), and then candidate A (two outcome units). The efficiency principle recommends to allocate the two units of the good to candidate B and candidate C. This happens to be a much more unequal allocation (Gini-coefficient of 0.24) then the one recommended by the equal outcome principle (Gini-coefficient of 0.05). It is, however, not generally true that the efficiency principle necessarily favors unequal allocations.

Concerns for efficiency arise whenever the effect of the good on the outcome differs across candidates. Then, candidates can make an inter-personally comprehensive and acceptable claim to more than an equal share of the good that is based on the gain in efficiency that is realized by allocating the good to them. Concerns for equality, desert, and need can override efficiency claims, however—especially because efficient allocations can be very unequal and might not benefit the candidates with the highest desert or need. The efficiency principle, as formulated here, is a local version of consequentialism (Sinnott-Armstrong, 2021)[10]. It is justified by a concern for maximizing the aggregate well-being (in terms of outcome attainment) of the candidate pool, the allocating institution, or society as a whole. The principle likely produces no incentive effects because candidates cannot actively influence the size of the outcome increment that is gained by allocating the good to them.

3.3. Combining principles

Due to the pluralism of distributive justice principles, there are frequently allocation problems for which multiple (potentially conflicting) principles are equally well-justified. Strategies for building compromises between principles include: (a) Combining decision criteria *via* a *weighting* scheme (Konow, 2003). Each of the C different decision criteria Y_c is assigned a weight w_c and the allocation decision is based on the weighted sum $Y^* = \sum_{c=1}^{C} w_c \cdot Y_c$ of the criteria. (b) Establishing a hierarchical ordering of the principles, where higher-ordered principles take precedence and lower-ordered principles break ties (Törnblom and Kazemi, 2015). (c) Conjunctive (disjunctive) procedures, where candidates are ranked according to the decision criterion on which they score lowest (highest) (Elster, 1992). The allocation decision is then based on the combined rank order.

Note that a given decision criterion can also be over-determined, i.e., supported by multiple justice principles. For instance, in a medical context, need and efficiency suggest allocating a kidney transplant to the candidate with the lowest pre-allocation life expectancy because (a) this expresses a concern for the worst-off and (b) the gain in additional life expectancy is highest for this candidate. Over-determination facilitates the selection of an allocation principle because it is easier to build a winning coalition of actors who support the principle (Elster, 1992).

4. Fair predictions

In Section 2.3, we introduced a general definition of algorithmic fairness and noted that it is compatible with a large range of substantive fairness metrics. It is beyond the scope of this paper to comprehensively review existing fairness metrics. Interested readers are referred to the summary article of Mitchell et al. (2021). Let us highlight, however, that many of the most popular metrics share a focus on equalizing predictions (independence, counterfactual fairness) or prediction errors (equal accuracy, sufficiency, separation) across groups that are defined by so-called protected attributes drawn from anti-discrimination law. Protected attributes include, amongst others, sex, gender, sexuality, ethnicity and race, and disability (Barocas and Selbst, 2016). Inequalities in predictions or prediction errors that are not systematically associated with these protected attributes are not considered as relevant instances of unfairness. It has been shown that these metrics are motivated by moral arguments derived from equality of opportunity theories (Heidari et al., 2019; Castro et al., 2021; Loi et al., 2021). Equality of opportunity, mostly applied in the allocation of social positions, states that access to goods (e.g., a job position) should only depend on candidates' qualification for the good (e.g., their educational credentials) and not on any other (morally arbitrary) attributes of the candidates (Arneson, 2015).

We argue that the exclusive focus on protected attributes is too narrow for data-driven decision procedures, an argument to which we return in Section 5. Our main point is the following: The exclusive focus on protected attributes is not justified because any systematic tendency of the prediction algorithm to assign more prediction error to a group—protected or unprotected—is unfair. Each systematically biased prediction algorithm creates a new algorithm-specific group of candidates who are systematically disadvantaged and have a reasonable claim to protection (Fazelpour and Lipton, 2020). Further, equality of opportunity is intended to regulate the allocation of goods, not the allocation of prediction errors. It cannot, therefore, be used to justify a certain allocation of prediction

life expectancy. We allocate the food to the candidates in whom the food will produce the largest increment in hunger reduction.

10 It is a generalization of the *Individual increments of welfare* criterion (Elster, 1992, p. 85) to outcomes that are not necessarily related to candidates' individual welfare.

errors. It might be justified to account for protected attributes at the decision step. However, the same is not true at the prediction step because all candidates, irrespective of their membership in protected groups, have the same claim to receiving equally good predictions.

In an attempt to go beyond the narrow focus on protected attributes, we now provide a formal representation of our notion of algorithmic fairness, which we call *error fairness*.

Definition 3 (error fairness). *Let V be the observed features and U be the unobserved features of the candidates. Let ε measure the deviation between the predicted criterion value (either $\hat{Y}$ or $\hat{S}$) and the observed criterion value Y. A prediction algorithm is error-fair iff $\varepsilon \perp (V, U)$.*

For continuous decision criteria, the residuals $(\hat{y}_i - y_i)$ may be used as an error measure. For binary criteria, pseudo-residuals $(\hat{s}_i - y_i)$ may be used to measure deviations. However, our notion of error fairness is not tied to a specific error type and other measures could be considered dependent on the application context.

Error fairness is satisfied if prediction errors are not systematically related to (i.e., statistically independent of) observed or unobserved candidate features. An error-fair prediction algorithm accomplishes the prediction task equally well for all candidates without systematic error. We acknowledge two caveats of error fairness: (a) It is impossible to check the statistical independence between unobserved candidate features and prediction errors. (b) It is very difficult or even impossible to satisfy error fairness perfectly. Nevertheless, we maintain that error fairness is valuable as an aspirational goal. It motivates us to check, within our capabilities, whether prediction errors systematically befall certain segments of the candidate population. We return to this point in Section 5, where error fairness is embedded into a broader discussion of bias in data-driven decision-making.

With these limitations in mind, an approach for measuring the degree to which a prediction algorithm satisfies error fairness could proceed as follows. Error fairness is assessed in an independent test sample of candidates who were not used for model training. Prediction errors are computed using a task-specific error measure (e.g., pseudo-residuals for classification tasks). A linear regression of the prediction errors on the observed candidate features V (including interactions between candidate features and non-linear terms) is performed. More flexible types of regression methods could be considered to capture complex relationships between V and ε. The $R^2 \in [0, 1]$ statistic of the regression model—the share of variance in the errors explained by the observed features—is the fairness metric. Large R^2 indicate that there are systematic relationships between observed candidate features and the prediction errors. The larger the R^2, the more the prediction algorithm violates error fairness. The underlying idea of the R^2-metric is similar in spirit to the first step of multi-accuracy boosting (Kim et al., 2019), which aims to identify subgroups of candidates for which a prediction algorithm produces large prediction error. In addition, mutual information (Cover and Thomas, 2006, chapter 2), which is not limited to linear dependence, could be used as an alternative metric to check independence of prediction errors from observed candidate features[11].

These approaches, however, can only check the independence of prediction errors from observed candidate features V and not from unobserved candidate features U. Even if the algorithm has a small R^2, it may violate error fairness due to dependence of deviations on unobserved features—a small R^2 is a necessary but not sufficient condition for error fairness.

5. Pitfalls and biases of data-driven decision-making

Existing research on data-driven decision-making identified a series of biases that can affect data-driven decisions and lead to systematic discrimination against segments of the candidate population (Mehrabi et al., 2019; Suresh and Guttag, 2020). Here, we extend the framework of Friedler et al. (2021) to explicitly account for the distinction of prediction and decision step. We thereby also illustrate why the notion of error fairness is a valuable aspirational goal, even if it will be difficult or even impossible to achieve it perfectly in real-world applications.

Building on Friedler et al. (2021), we distinguish four spaces: The construct space (CS) contains the latent decision criteria identified by the middle-range distributive justice principles—namely equality, desert, need, and efficiency. The indicator space (IS) contains the manifest decision criteria that are chosen to operationalize the latent criteria. Construct space and indicator space are connected by the operation of operationalization. The measurement space (MS) contains the measured values of the chosen manifest decision criterion. These are the Ys in our notation. Indicator space and measurement space are connected by the operation of measurement. Finally, the prediction space (PS) contains the predicted values of the decision criterion. These are the $\hat{Y}$ (for continuous decision criteria) and $\hat{S}$ (for categorical decision criteria) in our notation. Measurement space and prediction space are connected by the operation of prediction. The prediction space is only needed if the decision criterion is unobserved at decision time. In this case, the measured decision criterion is only available in the training data, not for the candidates for which an allocation decision must be made[12].

11 We are indebted to an anonymous reviewer for suggesting mutual information.

12 Hertweck et al. (2021) present a similar four-space framework to identify the conditions under which egalitarian arguments support the application of statistical parity, a fairness metric that requires that predictions are statistically independent from protected attributes like

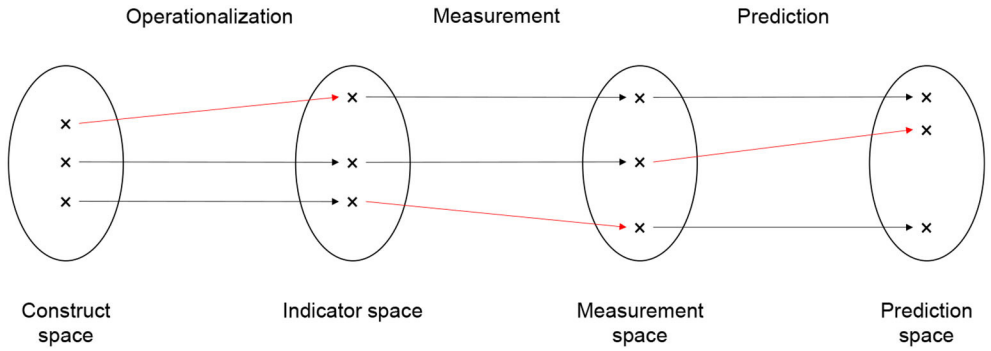


FIGURE 4
Spaces and biases.

Figure 4 depicts the four spaces (construct, indicator, measurement, and prediction space) as circles and the three operations that connect the spaces (operationalization, measurement, and prediction) as arrows. There are three candidates whose relative positions in the spaces are indicated by the black crosses. The crosses that are connected by arrows all belong to the same candidate. The relative position of a candidate indicates her value on the latent decision criterion (in the construct space), the manifest decision criterion (in the indicator space), the measured decision criterion (in the measurement space), and the predicted decision criterion (in the measurement space) in relation to the values of the other candidates.

Bias can arise at the transition from one space to the next, that is, in the operations we called operationalization, measurement, and prediction. Bias is present if the operations change the relative distance between candidates in the spaces. Accordingly, there are three types of bias: operationalization bias, measurement bias, and prediction bias. In Figure 4, the biases are indicated by red arrows. The red arrows are not strictly horizontal, indicating that the relative distance between the candidate changes. Starting from the left, the first red arrow indicates operationalization bias, the second red arrow indicates measurement bias, and the third red arrow indicates prediction bias.

Operationalization bias is present if the relative distance between candidates on the manifest decision criterion differs from their distance on the latent construct. Consider, for instance, a hospital that wants to allocate access to treatment options (the good) based on patients medical need (the latent decision criterion; Obermeyer et al., 2019). Patients with higher needs should receive more treatment. Medical need is operationalized as a patient's past spending on treatment (manifest decision criterion) under the assumption that patients with higher needs will have spend more money on treatment. Bias is introduced because poor patients cannot spend as much on treatment as wealthier patients, even if they have the same level of medical need. The operationalization under-represents the medical need of poor patients but not the medical need of wealthy patients. Accordingly, the distance between poor and wealthy patients in the indicator space will be larger than their distance in the construct space.

Measurement bias is present if the relative distance between candidates on the measured decision criterion differs from their distance on the manifest decision criterion. Consider a probation panel that wants to allocate probation (the good) based on a convicted person's desert (the latent decision criterion). Desert is operationalized as the number of re-offenses of the convicted person in the past (the manifest decision criterion). The number of re-offenses is measured as the number of re-arrests (the measured decision criterion) of that person, as recorded in police documents. In the US (and probably also in other countries), the number of re-arrests is a biased measure of the number re-offenses (Lum and Isaac, 2016). Black persons, for instance, are more likely to be arrested than White persons even if they re-offend to the same level. The measurement under-represents the number of re-offenses among White persons but not among Black persons. Accordingly, the distance between Black persons and White persons in the measurement space is larger than their distance in the indicator space.

Prediction bias is present if the relative distance between candidates on the predicted decision criterion differs from their distance on the measured decision criterion. Prediction bias occurs for a number of reasons (Mehrabi et al., 2019; Suresh and Guttag, 2020). It arises, for instance, when the data

gender and ethnicity. The framework presented in this paper has a different focus. It is not limited to egalitarian arguments and it is not geared at justifying a specific fairness metric.

generating process changes such that the training data on which the prediction model is estimated were generated by a different process than the data for the candidates to which the estimated model is eventually applied. Consider a public employment agency that wants to allocate access to support programs (the good) based on need (the latent decision criterion), which is operationalized by re-employment after a maximum of 6 months of job-search (manifest decision criterion) and measured without bias. Consider further that the training data were generated by a process that features discrimination against female job-seekers. That is, female job-seekers in the training data are less likely than male job-seekers to find re-employment. The prediction model learns the association between gender and re-employment and, therefore, predicts a lower re-employment probability for female job-seekers than for male job-seekers. Imagine (somewhat unrealistically) that gender discrimination would suddenly disappear from one day to the next. Female job-seekers in the candidate pool would no longer be less likely to find re-employment than male job-seekers. Bias is introduced because the prediction model continues to predict lower re-employment probabilities for female than for male job-seekers (unless it is retrained on newer observations). The predictions under-represent the re-employment probability among female job-seekers but not among male job-seekers. Accordingly, the distance between female job-seekers and male job-seekers in the prediction space is larger than their distance in the measurement space.

Bias that is introduced at one transition tends to be carried forward to later transitions, unless there is a purposeful de-biasing or different biases happen to cancel each other out. Consider again the case of the hospital that wants to allocate medical treatment based on medical need. Due to the operationalization bias, the distance between poor and wealthy patients on medical spending (manifest decision criterion) is larger than the difference on medical need (latent decision criterion). Even if medical spending is measured without bias, the difference on measured medical spending (measured decision criterion) between poor and wealthy patients is larger than the difference on medical need (latent decision criterion). The lesson is: It is necessary to think about all three types of bias. The absence of prediction bias, for instance, does not guarantee that there is no measurement bias or operationalization bias. The major problem is that we usually do not know the relative distance between candidates' latent decision criteria in the construct space. If we did, we would not need to go through the entire process of operationalizing, measuring, and (sometimes) predicting. Similarly, we usually do not know the relative distance between candidates' manifest decision criteria in the indicator space. The distances become only visible after we applied the measurement operation. Substantive background knowledge and critical thinking appear to be the most effective weapons to detect operationalization bias and measurement bias. The same is true for prediction bias, with the addition that we can rely on fairness metrics (Mitchell et al., 2021) to detect bias.

In the absence of these three biases, it is reasonable to assume that the measured decision criterion (in cases where it is observed at decision time) or the predicted decision criterion (in cases where the manifest criterion is unobserved at decision time) is a good representation of the latent decision criterion. In the absence of bias, we could, for instance, assume that differences in measured medical spending between candidates correspond to equal differences between candidates in their medical need. In the absence of bias, it makes sense to allocate goods based on the measured (or the predicted) decision criterion.

A final point of discussion is the question whether distances between candidates should be defined on the group-level or the individual-level. Friedler et al. (2021) and the majority of fairness metrics (Mitchell et al., 2021) chose the group-level. This includes fairness metrics that measure differences in prediction errors between (a set of pre-defined) groups, such as overall accuracy equality (Berk et al., 2021) or equalized odds (Hardt et al., 2016). Bias is present if the transition between spaces (operationalization, measurement, or prediction) changes the distance between members of pre-defined social groups. The groups are defined based on so-called protected attributes, including amongst others gender, sexual orientation, disability, and ethnicity (Barocas and Selbst, 2016). Under most anti-discrimination laws, these protected attributes should not affect allocation decisions. Put differently, changes in the relative distance between individual candidates are only considered as bias when these changes align with protected groups. Consider again the probation panel that wants to allocate probation (the good) based on the number of re-offenses (manifest decision criterion), measured as the number of re-arrests (measured decision criterion). We noted that measurement bias increases the distance between Black persons and White persons in the measurement space compared to the indicator space. The group-level perspective would indeed recognize this measurement bias as a relevant form of bias. Now imagine that the measurement operation introduces a similar bias between left-handed and right-handed persons. For whatever reason, left-handed persons are re-arrested at higher rates than right-handed persons even if they have the same level of re-offending. Accordingly, the measurement would increase the distance between left-handed persons and right-handed persons in the measurement space compared to the indicator space. The group-level perspective would not recognize this as a relevant instance of bias because handedness is not a protected attribute.

In our opinion, this is a problematic implication of the group-level perspective. In general, we agree that some attributes are especially important because we need to redress historical injustices and because unequal treatment based

on these attributes exists across a large range of allocation decisions (Loi et al., 2021). We argue, however, that bias that systematically affects groups defined by seemingly innocent attributes like handedness becomes problematic in the context of data-driven decision-making. Once data-driven decision-making is implemented by an institution, it is applied to a large number of candidates and regulates the access to the goods that the institution allocates. The institution frequently has the monopoly on the good, such that candidates have no other choice than to subject themselves to the data-driven decision process if they want the good. Job-seekers who seek access to support programs can only turn to the public employment agency, convicted persons can only turn to the probation panel if they want probation. In this situation, we believe, it would also be wrong if the decision process were systematically biased against left-handed persons (or any other group defined by non-protected attributes). More succinctly: If a data-driven decision process is applied to fully regulate the allocation of a good, it has the potential to create new groups that are systematically disadvantaged in the access to that good (Fazelpour and Lipton, 2020). These groups may be defined by protected attributes but can also be defined by any other non-protected attribute.

The individual-level perspective helps to address this point. The individual candidate is seen as the collection of all her attributes (protected and non-protected). Any change in the relative distance between candidates that aligns with one (or more) attribute of the candidates is a relevant instance of bias—irrespective of whether the attribute is protected or not. This is the spirit in which we formulated error fairness (Section 4). In the language of this section, error fairness states that the prediction operation is unbiased (or fair) if and only if the prediction does not change the relative distance between candidates in the prediction space compared to the measurement space. We acknowledge that error fairness is an aspirational target. To proof that a prediction operation is error fair requires showing that changes in the relative distance between candidates that occur in the transition from measurement space to prediction space are independent of all candidate attributes. The major problem is that we never observe all attributes of the candidates. Therefore, it is impossible to show that none of the candidate attributes is related to changes in relative distance. Again, substantive background knowledge and critical thinking are the best weapons to fight bias. We should strive to test (within the limits of privacy rights) associations between distance changes and the attributes that background knowledge and critical thinking suggest as the most important. We should not, however, limit ourselves to a pre-defined set of protected attributes. Protected attributes might be relevant in the case under study but they might just as well be irrelevant.

6. Discussion

The advent of data-driven decision-making in more and more areas of life (e.g., automated job advertisements, employee management, college admission, credit scoring, or more general access to public services) raises the dual problem of fairness in predictions and justice in decisions. Fairness and justice are conflated in the existent literature on data-driven decision-making systems, with the consequence that there exists a multitude of mutually incompatible fairness definitions—each motivated by a distinct set of moral concerns. To advance the literature, we propose an alternative approach that builds on a clean distinction between fairness and justice. Fairness regulates the distribution of prediction errors, whereas justice regulates the allocation of goods. The approach has practical implications for the design of data-driven decision systems but should also be viewed in light of its limitations.

6.1. Implications for practice

The approach suggests the following four-step process to designing fair and just decision systems. (a) Make a well-justified choice of the distributive justice principle. The principle is well-justified if a convincing rational defense of the principle can be provided to all candidates who are eventually affected by it. The design of an allocation system, therefore, requires stake-holder involvement—a requirement shared by impact assessment frameworks developed for data-driven decision systems (Selbst, 2018; Mantelero, 2018; Metcalf et al., 2021). (b) Make a context-fitting translation of the chosen justice principle(s) into an allocation principle. The allocation principle consists of a set of decision criteria and a rule that specifies how the criteria are related to allocation decisions. The translation is context-fitting if the chosen criterion and rule transport the general intention of the justice principle into the specific allocation context. (c) If the decision criterion is unobserved at decision time, use a fair prediction algorithm to predict its value. (d) Investigate whether the decision procedure is affected by operationalization bias, measurement bias, or prediction bias.

The approach highlights that fairness in predictions is one among multiple concerns. The selection of the distributive justice principle, its translation into an allocation principle, and the instrument that measures the decision criterion require equally close scrutiny. Note that the approach is modular: It is possible to reject our fairness definition and still accept our justice definition (and vice versa). A researcher who is not convinced by error fairness can apply her favored alternative fairness definition. The resulting predictions are then translated into an allocation decision *via* a well-justified allocation principle.

6.2. Limitations

While we believe that *error fairness* formalizes an intuitive and useful definition of fairness, its translation into a fairness metric proved rather difficult. We proposed the R^2 from a linear regression of prediction errors on candidate features as a possible metric. The R^2-metric is a necessary but not sufficient condition for error fairness: A prediction algorithm that satisfies error fairness achieves good results on the R^2-metric. But the algorithm can violate error fairness and still achieve good results if the violation is due to systematic relationships between prediction errors and unobserved features. Future research should identify metrics with a stronger connection to error fairness.

Error fairness is not sensitive to historical bias (Suresh and Guttag, 2020). Historical bias is present if the data on which the prediction algorithm is trained reflect past discrimination against certain groups of candidates. Discrimination creates differences in bases rates: Members of the disadvantaged groups have less favorable values on the decision criterion. The prediction algorithm learns the historical bias and assigns less favorable predictions to members of the disadvantaged groups. Error fairness is not violated in the presence of historical bias as long as the predictions accurately reflect the true values of the decision criterion for all candidates. The predictions should track differences in base rates. Other fairness definitions (independence, counterfactual fairness) are sensitive to historical bias. We adopt the position of Corbett-Davies and Goel (2018) on this point: The fact that differences in base rates are a product of past discrimination does not mean that current predictions are inaccurate or that better societal outcomes could be achieved by altering predictions. More succinctly: "It would be misleading to characterize an algorithm or its training data as unfair for accurately identifying existing statistical patterns" (Corbett-Davies and Goel, 2018, p. 13). Importantly, we do not reject the need to correct unwanted discrimination or historical bias. Corrections should be applied at the decision step and not the prediction step, however. If there exists a justice principle that justifies such corrections in a given allocation problem (and we believe that there often is such a principle), it is permissible to define an allocation principle that implements the necessary corrections.

Finally, the list of *middle-range distributive justice principles* is not exhaustive. We invite researchers and practitioners to add to the list. To be admissible to the list, justice principles must define and justify (a) a decision criterion and (b) a rule that relates the criterion to the allocation of goods. The justice principle should not regulate the allocation of prediction errors. Equality of opportunity (Arneson, 2015) is a promising candidate. Equality of opportunity restricts the set of permissible decision criteria to criteria that are not related to protected features. Or else, it recommends allocation rules that compensate members of historically disadvantaged groups for discrimination that prevented them from developing the decision criterion.

7. Conclusion

Prior work on data-driven decision-making systems extensively explored the moral foundations of prominent algorithmic fairness definitions. This paper contributes a cleaner distinction between fairness and justice in data-driven decision-making. This distinction is instrumental for ethical self-assessment when building data-driven decision systems and can also guide regulations such as the EU AI Act. We clarify the relation between fairness and justice and provide clear definitions of both concepts. The paper provides an overview of distributive justice theories and a recipe for implementing the theories into the decision-making pipeline. Taken together, we contribute the outline of a principled local justice approach to the design of fair and just data-driven decision procedures—an approach that is urgently needed as data-driven decision-making increasingly enters all walks of life.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Author contributions

All authors contributed equally to the conception and design of the study. MK and CK conducted the main conceptual analysis. MK wrote the first draft of the manuscript. CK wrote sections of the manuscript. All authors contributed significantly to manuscript revision and read and approved the submitted version.

Funding

The work of MK was supported by the University of Mannheim's Graduate School of Economic and Social Sciences. We acknowledge funding from the Baden-Württemberg Stiftung (grant FairADM—Fairness in Algorithmic Decision Making) and the Volkswagen Stiftung (grant Consequences of Artificial Intelligence for Urban Societies). Part of this work was done while CK was visiting the Simons Institute for the Theory of Computing, UC Berkeley. The publication of this article was funded by the University of Mannheim.

Acknowledgments

We thank Patrick Schenk (LMU Munich) for his thorough feedback on an earlier version of this article. The article was presented at the joint conference of the German (DGS) and Austrian (ÖGS) Sociological Associations in 2021.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Adler, M. D., and Holtug, N. (2019). Prioritarianism: a response to critics. *Polit. Philos. Econ*. 18, 101–144. doi: 10.1177/1470594X19828022

Alikhademi, K., Drobina, E., Prioleau, D., Richardson, B., Purves, D., and Gilbert, J. E. (2021). A review of predictive policing from the perspective of fairness. *Artif. Intell. Law* 30, 1–17. doi: 10.1007/s10506-021-09286-4

Angwin, J., Mattu, S., and Kirchner, L. (2016). *Machine Bias*. Technical report, ProPublica.

Arneson, R. (2013). "Egalitarianism," in *The Stanford Encyclopedia of Philosophy*, ed E. N. Zalta (Stanford: Metaphysics Research Lab; Stanford University).

Arneson, R. (2015). "Equality of opportunity," in *The Stanford Encyclopedia of Philosophy*, ed E. N. Zalta (Stanford: Metaphysics Research Lab; Stanford University).

Barda, N., Riesel, D., Akriv, A., Levy, J., Finkel, U., Yona, G., et al. (2020). Developing a COVID-19 mortality risk prediction model when individual-level data are not available. *Nat. Commun.* 11:4439. doi: 10.1038/s41467-020-18297-9

Barocas, S., and Selbst, A. D. (2016). Big data's disparate impact. *Calif. Law Rev*. 104, 671–732. doi: 10.2139/ssrn.2477899

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2021). Fairness in criminal justice risk assessments: the state of the art. *Sociol. Methods Res*. 50, 3–44. doi: 10.1177/0049124118782533

Brock, G. (2018). *Sufficiency and Needs-Based Approaches, Vol. 1*. New York, NY: Oxford University Press. doi: 10.1093/oxfordhb/9780199645121.013.6

Carton, S., Helsby, J., Joseph, K., Mahmud, A., Park, Y., Walsh, J., et al. (2016). "Identifying police officers at risk of adverse events," in *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery), 67–76. doi: 10.1145/2939672.2939698

Castro, C., O'Brien, D., and Schwan, B. (2021). "Fairness and machine fairness," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, AIES '21* (New York, NY: Association for Computing Machinery), 446. doi: 10.1145/3461702.3462577

Chouldechova, A. (2016). Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *arXiv:1610.07524 [cs, stat]*. doi: 10.1089/big.2016.0047

Cohen, R. L. (1987). Distributive justice: theory and research. *Soc. Just. Res*. 1, 19–40. doi: 10.1007/BF01049382

Corbett-Davies, S., and Goel, S. (2018). The measure and mismeasure of fairness: a critical review of fair machine learning. *arXiv: 1808.00023*. doi: 10.48550/arXiv.1808.00023

Cover, T. M., and Thomas, J. A. (2006). *Elements of Information Theory, 2nd Edn*. Hoboken, NJ: Wiley-Interscience.

Desiere, S., Langenbucher, K., and Struyven, L. (2019). "Statistical profiling in public employment services: an international comparison," in *OECD Social, Employment and Migration Working Papers* (Paris), 224.

Deutsch, M. (1975). Equity, equality, and need: what determines which value will be used as the basis of distributive justice? *J. Soc. Issues* 31, 137–149. doi: 10.1111/j.1540-4560.1975.tb01000.x

Dworkin, R. (1981). What is equality? Part 1: equality of welfare. *Philos. Publ. Affairs* 10, 185–246.

Elster, J. (1992). *Local Justice: How Institutions Allocate Scarce Goods and Necessary Burdens*. New York, NY: Russell Sage Foundation.

Engler, A. (2022a). *The EU AI Act Will Have Global Impact, But a Limited Brussels Effect*. Technical report. Brookings Institute.

Engler, A. (2022b). *Institutionalizing Data Analysis in German Federal Governance*. Technical report. Brookings Institute.

Fazelpour, S., and Lipton, Z. C. (2020). "Algorithmic fairness from a non-ideal perspective," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY: ACM), 57–63. doi: 10.1145/3375627.3375828

Feldman, F., and Skow, B. (2020). "Desert," in *The Stanford Encyclopedia of Philosophy*, ed E. Zalta (Stanford: Metaphysics Research Lab; Stanford University).

Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. (2021). The (im)possibility of fairness: different value systems require different mechanisms for fair decision making. *Commun. ACM* 64, 136–143. doi: 10.1145/3433949

Gosepath, S. (2011). "Equality," in *The Stanford Encyclopedia of Philosophy*, ed E. N. Zalta (Stanford: Metaphysics Research Lab; Stanford University).

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *arXiv:1610.02413 [cs]*. doi: 10.48550/arXiv.1610.02413

Hart, N., and Yohannes, M. (2019). *Evidence Works: Cases Where Evidence Meaningfully Informed Policy*. Bipartisan Policy Center.

Hebert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. (2018). "Multicalibration: calibration for the (computationally-identifiable) masses," in *Proceedings of the 35th International Conference on Machine Learning, Vol. 80 of Proceedings of Machine Learning Research*, eds J. Dy and A. Krause (Stockholm), 1939–1948.

Heidari, H., Loi, M., Gummadi, K. P., and Krause, A. (2019). "A moral framework for understanding fair ML through economic models of equality of opportunity," in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA), 181–190. doi: 10.1145/3287560.3287584

Hertweck, C., Heitz, C., and Loi, M. (2021). "On the moral justification of statistical parity," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (ACM), 747–757. doi: 10.1145/3442188.3445936

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. U.S.A.* 102, 16569–16572. doi: 10.1073/pnas.0507655102

Holtug, N. (2007). "Prioritarianism," in *Egalitarianism: New Essays on the Nature and Value of Equality*, eds N. Holtug and K. Lippert-Rasmussen (Oxford; New York, NY: Clarendon Press), 125–156.

Kim, M. P., Ghorbani, A., and Zou, J. (2019). "Multiaccuracy: black-box post-processing for fairness in classification," in *Proceedings of the 2019 AAAI/ACM*

Conference on AI, Ethics, and Society, AIES '19 (New York, NY: Association for Computing Machinery), 247–254. doi: 10.1145/3306618.3314287

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv: 1609.05807*. doi: 10.48550/arXiv.1609.05807

Konow, J. (2003). which is the fairest one of all? A positive analysis of justice theories. *J. Econ. Liter*. 41, 1188–1239. doi: 10.1257/00220510377 1800013

Kozodoi, N., Jacob, J., and Lessmann, S. (2021). Fairness in credit scoring: assessment, implementation and profit implications. *Eur. J. Oper. Res*. 297, 1083–1094. doi: 10.1016/j.ejor.2021. 06.023

Kuppler, M. (2022). Predicting the future impact of Computer Science researchers: is there a gender bias? *Scientometrics*. doi: 10.1007/s11192-022-04337-2. [Epub ahead of print].

Lamont, J., and Favor, C. (2017). "Distributive justice," in *The Stanford Encyclopedia of Philosophy*, ed E. N. Zalta (Stanford: Metaphysics Research Lab; Stanford University). doi: 10.4324/9781315257563

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., and Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes: the premise, the proposed solutions, and the open challenges. *Philos. Technol*. 31, 611–627. doi: 10.1007/s13347-017-0279-x

Lippert-Rasmussen, K. (2018). "Justice and bad luck," in *The Stanford Encyclopedia of Philosophy*, ed E. N. Zalta (Stanford: Metaphysics Research Lab; Stanford University).

Loi, M., Herlitz, A., and Heidari, H. (2021). "Fair equality of chances for prediction-based decisions," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, AIES '21* (New York, NY: Association for Computing Machinery), 756. doi: 10.1145/3461702.3462613

Lum, K., and Isaac, W. (2016). To predict and serve? *Significance* 13, 14–19. doi: 10.1111/j.1740-9713.2016.00960.x

Makhlouf, K., Zhioua, S., and Palamidessi, C. (2020). On the applicability of ML fairness notions. *arXiv:2006.16745 [cs, stat]*. doi: 10.48550/arXiv.2006.16745

Mantelero, A. (2018). AI and big data: a blueprint for a human rights, social and ethical impact assessment. *Comput. Law Secur. Rev*. 34, 754–772. doi: 10.1016/j.clsr.2018.05.017

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv:1908.09635 [cs]*. doi: 10.48550/arXiv.1908.09635

Metcalf, J., Moss, E., Watkins, E. A., Singh, R., and Elish, M. C. (2021). "Algorithmic impact assessments and accountability: the co-construction of impacts," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (ACM), 735–746. doi: 10.1145/3442188.3445935

Metz, C., and Satariano, A. (2020). *An Algorithm That Grants Freedom, or Takes It Away*. New York Times.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. (2021). Algorithmic fairness: choices, assumptions, and definitions. *Annu. Rev. Stat. Appl*. 8, 141–163. doi: 10.1146/annurev-statistics-042720-125902

Moriarty, J. (2018). *Desert-Based Justice, Vol. 1*. New York, NY: Oxford University Press. doi: 10.1093/oxfordhb/9780199645121.013.7

New Zealand Ministry of Social Development (2014). *The Feasibility of Using Predictive Risk Modelling to Identify New-Born Children Who Are High Priority for Prevention Services*. Ministry of Social Development, Wellington.

Nozick, R. (1974). *Anarchy, State, and Utopia. Basic Books, a Member of the Perseus Books Group*, New York, NY.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 447–453. doi: 10.1126/science.aax2342

Parfit, D. (1997). Equality and priority. *Ratio* 10, 202–221. doi: 10.1111/1467-9329.00041

Rawls, J. (1971). *A Theory of Justice*. Cambridge: Harvard University Press. doi: 10.4159/9780674042605

Rodolfa, K. T., Saleiro, P., and Ghani, R. (2021). "Bias and fairness," in *Big Data and Social Science: Data Science Methods and Tools for Research and Practice, 2nd Edn*, eds I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter, and J. Lane (Boca Raton, FL: CRC Press), 281–312.

Schmidt, V. H. (1992a). Adaptive justice: local distributive justice in sociological perspective. *Theory Soc*. 21, 789–816. doi: 10.1007/BF00992812

Schmidt, V. H. (1992b). Lokale gerechtigkeit: perspektiven soziologischer gerechtigkeitsanalyse. *Zeitsch. Soziol*. 21, 3–15. doi: 10.1515/zfsoz-1992-0101

Schmidt, V. H. (1994). Bounded justice. *Soc. Sci. Inform*. 33, 305–333. doi: 10.1177/053901894033002009

Selbst, A. D. (2018). Disparate impact in big data policing. *Georgia Law Rev*. 52. doi: 10.2139/ssrn.2819182

Sinnott-Armstrong, W. (2021). "Consequentialism," in *The Stanford Encyclopedia of Philosophy*, ed E. N. Zalta (Stanford: Metaphysics Research Lab; Stanford University).

Suresh, H., and Guttag, J. V. (2020). A framework for understanding unintended consequences of machine learning. *arXiv:1901.10002*. doi: 10.48550/arXiv.1901. 10002

Törnblom, K., and Kazemi, A. (2015). "Distributive justice," in *The Oxford Handbook of Justice in the Workplace*, eds R. S. Cropanzano and M. L. Ambrose (New York, NY: Oxford University Press), 15–50.

Weihs, L., and Etzioni, O. (2017). "Learning to predict citation-based impact measures," in *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries, JCDL '17* (IEEE Press), 49–58. doi: 10.1109/JCDL.2017. 7991559

Završnik, A. (2021). Algorithmic justice: algorithms and big data in criminal justice settings. *Eur. J. Criminol*. 18, 623–642. doi: 10.1177/1477370819876762

frontiers | Frontiers in Sociology

TYPE Original Research
PUBLISHED 17 November 2022
DOI 10.3389/fsoc.2022.909458

OPEN ACCESS

EDITED BY
Tobias Wolbring,
University of Erlangen
Nuremberg, Germany

REVIEWED BY
Heinz Leitgöb,
Catholic University of
Eichstätt-Ingolstadt, Germany
Christopher Barrie,
University of Edinburgh,
United Kingdom
Kokil Jaidka,
National University of
Singapore, Singapore

*CORRESPONDENCE
Nicole Schwitter
nicole.schwitter.1@warwick.ac.uk
Ulf Liebe
ulf.liebe@warwick.ac.uk

SPECIALTY SECTION
This article was submitted to
Sociological Theory,
a section of the journal
Frontiers in Sociology

RECEIVED 31 March 2022
ACCEPTED 24 October 2022
PUBLISHED 17 November 2022

CITATION
Schwitter N, Pretari A, Marwa W,
Lombardini S and Liebe U (2022) Big
data and development sociology: An
overview and application on
governance and accountability
through digitalization in Tanzania.
Front. Sociol. 7:909458.
doi: 10.3389/fsoc.2022.909458

Big data and development sociology: An overview and application on governance and accountability through digitalization in Tanzania

Nicole Schwitter[1]*, Alexia Pretari[2], William Marwa[3], Simone Lombardini[2] and Ulf Liebe[1]*

[1]Department of Sociology, University of Warwick, Coventry, United Kingdom, [2]Oxfam GB, Oxford, United Kingdom, [3]Oxfam International, Dar es Salaam, Tanzania

The digital revolution and the widespread use of the internet have changed many realms of empirical social science research. In this paper, we discuss the use of big data in the context of development sociology and highlight its potential as a new source of data. We provide a brief overview of big data and development research, discuss different data types, and review example studies, before introducing our case study on active citizenship in Tanzania which expands on an Oxfam-led impact evaluation. The project aimed at improving community-driven governance and accountability through the use of digital technology. Twitter and other social media platforms were introduced to community animators as a tool to hold national and regional key stakeholders accountable. We retrieve the complete Twitter timelines up to October 2021 from all ∼200 community animators and influencers involved in the project (over 1.5 million tweets). We find that animators have started to use Twitter as part of the project, but most have stopped tweeting in the long term. Employing a dynamic difference-in-differences design, we also do not find effects of Oxfam-led training workshops on different aspects of animators' tweeting behavior. While most animators have stopped using Twitter in the long run, a few have continued to use social media to raise local issues and to be part of conversations to this day. Our case study showcases how (big) social media data can be part of an intervention, and we end with recommendations on how to use digital data in development sociology.

KEYWORDS

accountability, big data, development sociology, difference-in-difference, digital data, Tanzania, Twitter

Introduction

The digital revolution and the widespread use of the internet have influenced and changed many realms of empirical social science research. The usages of (big) digital data are flourishing in the growing field of computational social sciences, and novel digital sources of data are becoming popular to gain new insights into old and new questions of the social sciences (Lazer et al., 2009, 2020; Keuschnigg et al., 2017; Salganik, 2018; Edelmann et al., 2020).

Many studies have made use of digital technologies to access rich data sources. Particularly, digital trace data—records of activity undertaken through an online information system such as websites, social media platforms, smartphone apps, or other digital trackers and sensors (Howison et al., 2011; Stier et al., 2019)—are increasingly used as a substitute of or complement to more traditional data sources as their availability tends to allow the time- and cost-effective real-time collection of large amounts of data. While big data come with their traps and biases (Lazer et al., 2014), particularly around representativeness due to access to the internet and/or devices, which are not neutral to race, class, gender, and geography, they can provide a uniquely unobtrusive way to access information from people who are in positions of marginalization and may be reluctant to engage with institutions and institutional players, such as researchers.

An increasing interest in development sociology has emerged throughout the last decades and more and more sociologists work (again) on sociological issues in the context of so-called "developing countries" to reflect on the "development sector" (Viterna and Robertson, 2015). Next to theoretical accounts, empirical research is an integral part of development research. It is therefore natural to consider big data analysis a promising tool in research and intervention studies. In fact, the development sector already explores the possibilities of big data, discussing both the advantages and disadvantages of using various data sources such as satellite images, social media data, or large text corpora for research and evaluation (e.g., Abreu Lopes et al., 2018; Data2X, 2019; York and Bamberger, 2020). Against this background, we aim to shed more light on the nexus between big data and development sociology based on a transdisciplinary collaboration between sociologists and Oxfam[1]. We want to highlight opportunities for analysis with digitally available data in the context of development sociology. To support our argument, we present the collaborative project as a case study and particularly focus on the long-term sustainability analyzable with (big) digital trace data.

This paper is structured as follows: In the next section, we provide a brief overview of the use of big data in development research and discuss different data types. We review example studies and highlight the sociological potential of big data in this context. After this overview, we discuss our own case study on active citizenship in Tanzania where we have used digital trace data from Twitter. We will introduce the study, contextualize it, and discuss important concepts employed. We will present our data and methods, as well as the findings regarding the Twitter activity. We analyze short- and long-term effects of the intervention and assess to what extent the project allowed citizens in rural areas of Tanzania to express themselves, reach key stakeholders, and to hold them accountable. In the last section, we will give concluding remarks and recommendations on using big data in development sociology.

Big data in development research: A brief overview and its sociological potential

Big data provides new opportunities for international development research and evaluation and is receiving increasing attention (Abreu Lopes et al., 2018; Data2X, 2019; York and Bamberger, 2020). Big data are characterized by their remarkably large volume, variety, and velocity—big data is enormous, comes from different sources and in both structured and unstructured formats, and the data flows at a fast pace and is often generated continuously (Salganik, 2018, Chapter 2). Following York and Bamberger (2020, p. 10) three categories of big data can be differentiated: (1) "human-generated (centered) data" including social media data, internet searches, and text data; (2) "administrative (transactional) data" including migration reports, employment data, and combinations of different governmental and non-governmental data sources; (3) "geospatial data" including "satellites, drones, and remote sensing". Such data is created and collected by humans, companies, and governments for purposes other than research and require repurposing (Salganik, 2018, Chapter 2). Our own example study presented below uses Twitter data, thus falling into the first category.

Big data can be relevant for development research and especially impact evaluation—assessing the difference a specific intervention makes in people's lives—in at least two ways. First, in addition to other sources of data gathered through surveys or focus groups for example, big data can be used to evaluate the effects of social interventions. Individual interviews (face-to-face or on the phone), which can be both semi-or fully structured, are currently the standard for evaluation in international development. However, they are prone to many forms of biases, including social desirability bias (Krumpal, 2013), and the frequent use of interviews can lead to respondent fatigue. Table 1 provides a comparison of big data and standard survey data regarding selected characteristics (see York and Bamberger, 2020 for a more comprehensive overview). It exemplifies that both big data and survey data have advantages and disadvantages. For example, while big data can easily cover a whole population for which data is available and can (repeatedly) be collected in a relatively short time, surveys are less prone to sample bias, and they are typically tailored to the specific research question at hand. Using big data comes with limitations placed by the platform accessed (and its application programming interface and terms of service), thus restricting a researcher's autonomy. While social desirability bias toward the researcher is often present in survey data, users are generally not tailoring their online activity toward a potential researcher.

1 Oxfam is an international confederation of charitable organizations focusing on the alleviation of global poverty, founded in Oxford in 1942.

TABLE 1 Comparison of big data and survey data used in development research and evaluation.

Category	Big data	Survey data
Coverage	Often whole population for which data are available but can be limited due to platform constraints and data deletion. [+/−]	Sample size requirements, e.g., due to costs, limit coverage. [−]
Sample bias	Data can be selective (e.g., only social media users included). [−]	Selection bias can be controlled as part of sampling. [+]
Relevance for a specific research question/evaluation	Data often created for different purposes. [−]	Data created for specific research. [+]
Social desirability bias (toward the researcher)	Often not present. [+]	Can be present. [−]
Time for data collection	Short(er) time needed for data collection. [+]	Long(er) time needed for data collection. [−]
Longitudinal data	Easy to collect panel data. (+)	Difficult to collect panel data. [−]

Adapted from York and Bamberger (2020, p. 9).

Still, it needs to be considered that individuals craft an online (public) brand or profile and/or that undesirable content is not even allowed on and thus moderated out by a platform. Overall, and given these different advantages and disadvantages, big data analysis is not expected to replace but to complement existing approaches in development research.

Second, big data can be part of intervention programs: interventions conducted on or using social media, with their effectiveness subsequently evaluated. This way, (big) digital trace data can be more easily combined with causal analysis which is a crucial part of estimating the effectiveness of interventions. Our example study presented below falls in this second category.

To highlight the potential of big data for development research and sociology, in the following we briefly present five example studies employing different types of big data and focusing on sociologically relevant topics (see also Table 2). The first example is the study of Jean et al. (2016) in which they predict poverty in developing countries using satellite imagery and machine learning. Data can guide research and (political) decision-making to combat and prevent poverty. However, valid and specific data can sometimes be missing depending on the country's context: In particular, data may be incomplete and not capture all aspects of the multidimensionality of poverty. It is unrealistic that large-scale surveys can be used to compensate for this lack of data (e.g., due to high costs). Jean et al. (2016) use publicly available high-resolution daytime satellite imagery in combination with machine learning to obtain poverty and wealth estimates at the "village" level. To this end, they train a deep learning model based on a "noisy" but easily accessible measure of poverty: night-time lightning. As part of this process, their "model learns to identify some livelihood-relevant characteristics of the landscape" (Jean et al., 2016, p. 791). This approach was validated for five African countries (Nigeria, Malawi, Tanzania, Uganda, Rwanda).

The second example is also concerned with existing data gaps; the study of Fatehkia et al. (2018) tracks the global digital gender gap. Such gender gaps are difficult to measure, especially in low-income countries. Fatehkia et al. (2018) use Facebook advertisement data on users by age and gender to predict digital gender gaps for over 150 countries. Facebook data is shown to be highly correlated with official data on digital gender gaps and their study is thus another important example showing how web data can expand coverage of development indicators (see also follow-up study Kashyap et al., 2020). In this line of research, several studies have discussed approaches to monitoring Sustainable Development Goals using big data (see for a review Allen et al., 2021).

The third example focuses on gender-based educational inequalities. Using mobile phone data from a large provider in Pakistan, Khan (2019) analyzes anonymized call detail record data comprising over one billion voice and text messages from approximately six million individual users. These data also include information on individuals' gender. Khan (2019) calculated district level averages for social network characteristics such as number of calls, network size, or friendship clusters. With a focus on gender differences, he then predicted gender-based educational inequalities in terms of primary school enrolment based on social network characteristics. He found that three network characteristics can explain almost 50 percent of the educational inequalities at the district level. These characteristics are "gender diversity of male calling networks", "clustering of friend groups across all networks", and "geographical reach across networks". Data2X (2019) presents many more of this type of big data studies in development research.

The fourth example study refers to combating HIV among men who have sex with men in Ghana where same-sex sexual acts between men are criminalized and gay men face stigma. As part of this pilot intervention study by Green et al. (2014), three "communication liaison officers" were employed who reached out to the target group on social media platforms including Facebook, WhatsApp, and Badoo. The overall project also had a face-to-face component based on 110 peer educators. The project team managed to reach over 15,000 men of the target

TABLE 2 Five examples of Big Data used in sociologically relevant development research.

Topic (and study)	Country context	Type of big data	Role of big data
Poverty reduction (Jean et al., 2016)	Nigeria, Malawi, Tanzania, Uganda, Rwanda	Satellite imagery	Estimating poverty and wealth indicators
Digital gender gaps (Fatehkia et al., 2018)	Global (over 150 countries)	Facebook advertisement data	Predicting digital gender gaps
Educational inequalities (Khan, 2019)	Pakistan	Mobile phone data	Explanation of gender-based educational inequalities at the district level
Combating HIV in contexts of social stigma (Green et al., 2014)	Ghana	Facebook, WhatsApp and Badoo data	Part of an intervention to promote HIV testing and counseling
Social protest / revolutions (Koehler-Derrick, 2013)	Egypt	Google Search Data	Monitoring public opinion and attention where polls are biased

group *via* the social media approach and over 12,000 *via* the face-to-face approach. Both approaches seemed to increase HIV testing and counseling uptake with a 99 percent increase *via* social media and a 64 percent increase *via* the offline intervention. While this pilot study has several limitations, for example regarding recording actual contacts with HIV testing and counseling, it demonstrates the potential of social media to get in contact with hard-to-reach populations in contexts of strong social stigma. Also, in this case study, social media approaches are shown to be much more cost-effective than face-to-face approaches.

The fifth example employs Google Search Data to examine political developments in Egypt in 2011/2012 (Koehler-Derrick, 2013). The Google data indicates a sustained interest in revolutionary figures and actions which contrasts with reports by the Supreme Council of the Armed Forces. This can be seen as an example of how big data can help to uncover "true preferences" and public opinion when other data sources such as "official polls" provide biased results. Yet, Koehler-Derrick (2013) also points to the disadvantages of Google Search Data which ideally needs to be combined with other forms of data collection to validate findings. Furthermore, the use of such data is only meaningful in contexts with sufficient internet penetration.

Using computational tools, we argue that, for at least three reasons, sociology can make substantial contributions to better understand and explain development issues. First, as also indicated in Table 2, it is obvious that many development issues refer to core explananda of sociological analysis. Such issues include for example poverty reduction, tackling social and structural inequality, strengthening civil society, and promoting norm and cultural change. Big data related research on these issues can benefit from sociological insights on these substantive topics. Thus, it might be especially beneficial in inter- and transdisciplinary contexts, which is most often the case in development research. Second, sociologists make important contributions to computational social science research in general which can benefit research on development issues. Important areas include the study of social networks (e.g., group formation), collective action (e.g., social protest movements), sociology of knowledge (e.g., consensus in science), cultural sociology (e.g., processes of cultural change), economic sociology (e.g., the role of culture for economic transactions), and population studies (e.g., estimating migration patterns) (see Edelmann et al., 2020 for an overview). Here, analytical approaches in sociology might help to move from prediction to explanation (i.e., uncovering behavioral determinants and mechanisms) in development-related big data research. Third, the field of sociology of development is particularly strong in mapping and reflecting on the "development sector" including governmental and non-governmental actors, how their decision-making affects communities and individuals' lives (Viterna and Robertson, 2015), and the power dynamics at play. Regarding big data research, important topics include the link between knowledge and power, whose knowledge is valued, and how structural inequalities can be reproduced in and through (computational) research. Combined with a "digital sociology" perspective (e.g., Marres, 2017), sociology can help to shed more light on the interplay between the development sector, big data approaches/analysis, and community/individual material living conditions in so-called developing countries.

In the following, we present a case study on active citizenship and governance, an inherently sociological topic. As part of a larger development project, this case study also employs a social media intervention. It is therefore a more in-depth example of how (big) social media data can be integrated into a development intervention. With our case study, we want to highlight the potential use of social media data in action, highlight different avenues of analysis, and discuss its limitations.

Case study: Active citizenship in Tanzania

In this section, we will introduce and present the findings of our study on active citizenship in Tanzania (Pretari et al., 2019). We will follow several different approaches to work with and analyze the (big) digital trace data collected to highlight the value-added and the limitations of this data in the context of development research.

This Oxfam-led project was implemented from February 2017 until March 2019 in four rural areas in Tanzania. The project aimed at improving community-driven governance and accountability through the use of digital technology. We will first introduce the project and its broader (theoretical) context in more detail before describing the methods and data used in this analysis. In this case study, the question we aim to answer is whether the intervention focusing on digital technology was effective at increasing greater online engagement. We analyze the online activity levels of those involved in this intervention and take advantage of the unique opportunity of assessing potential long-term effects. Particularly, we focus on the following key research questions: How have the animators and influencers involved in the project used Twitter over time (extent and content) and how was their content received by both key stakeholders and the general public? The following section will try to answer these questions. Across all analyses, the focus of our study lies on the sustainability of the intervention and changes across time. We want to make it explicit that we do not evaluate the overall project in this paper (see for this Pretari et al., 2019) but that we focus on the online Twitter component only.

Project background

Oxfam in Tanzania launched the "Governance and Accountability through Digitalization" project in 2017. The project built on the traditional animation approach developed through a former project "Chukua Hatua" ("Take Action" in Swahili), namely community animators, village-level organizers, or facilitators who mobilize or *animate* communities around a common advocacy agenda. The former project was launched in 2010 and was implemented in five regions of Tanzania. By encouraging active citizenship, particularly for women, it aimed to achieve increased accountability and responsiveness of the government. According to the "Effectiveness Review" of the program—a series of impact evaluations conducted on a random sample of mature projects and commissioned by Oxfam—, it has made crucial contributions to its selected outcomes (it contributed toward making councilors more aware and responsive, toward citizens mobilization by animators, and toward gaining support for community forest ownership; see Smith and Kishekya, 2013). The "Governance and Accountability through Digitalization" project then enhanced the traditional animation approach by integrating digital tools. This project was developed and implemented in collaboration with three Tanzanian civil society organizations.

Oxfam itself has been working in Tanzania since the 1960s and has been aiming to ensure enhanced governance and transparency, women's empowerment, and to tackle rural poverty. This project is unique in integrating digital tools into this context of governance and accountability in Tanzania. Other development projects in Tanzania, which made use of digital technologies, have tackled issues regarding the job search costs in rural areas by introducing an SMS-based messaging application to connect agricultural workers and employers on wages and evaluating it using randomized trials (Jeong, 2021) or regarding violence against women by using (in an ongoing project) mass media campaigns to shift attitudes and behaviors (Green et al., 2018; the project is building on an earlier study in Uganda, see Green et al., 2020). Next to these studies making use of digital technology, other recent projects in Tanzania, for example, tested the impact of gender training interventions on intimate partner violence (Lees et al., 2021), of financial incentives for testing negatively for sexually transmitted infections to prevent HIV and other infections (De Walque et al., 2012), of handwashing and sanitation on child health (Briceño et al., 2017), of increased school resources and teacher incentives on student learning (Mbiti et al., 2019), or of financial incentives on female land ownership (Ali et al., 2016).

The "Governance and Accountability through Digitalization" project presented here took place within the setting of Tanzania's Cybercrime Act of 2015, which criminalized and penalized different cyber activities. This act has been criticized from the very beginning by civil society as a threat to freedom of expression and as a means to control online spaces. The project was implemented between February 2017 and March 2019, and these years have seen a shrinking of the civic spaces in East Africa and a change in the political climate in Tanzania. The Human Rights Watch World Report 2019 highlights that "since the election of President John Magufuli in December 2015, Tanzania has witnessed a marked decline in respect for free expression, association, and assembly". In particular, the report highlights cases of criminalization of the sharing of information on WhatsApp, Facebook, or other online platforms by citizens and activists following the Cybercrime Act of 2015.

In this setting, the project built on traditional village-level animation approaches and enhanced them through the use of digital media. In our following analysis, we focus on Twitter. Internet and social media penetration in Tanzania has been increasing in recent years. Tanzania is undergoing a digital transformation with a growing number of people

connected to communications and internet services (Okeleke, 2019). As of 2022, the (DataReportal., 2022) reports that 25 percent of Tanzania's 62 million inhabitants use the internet, while in 2017 when this project started, the number was 14 percent (DataReportal, 2017). 10 percent of the population is reported to use social media, and Twitter is used by 1 percent (DataReportal., 2022). No information on Twitter penetration is available for the past, but across all social media, 9 percent are reported to have used it in 2017 (DataReportal, 2017). In the case of the villages part of this project, 5 percent of women and 10 percent of men citizens reported owning a smartphone (Pretari et al., 2019, p. 51). We focus on Twitter as it is the platform which is most popular amongst the leaders, elites, and influential business leaders in Tanzania.

As outlined in the report of the former "Chukua Hatua" project (see Green, 2015), one of the main targets of the project has been to overcome the prevalent sense of powerlessness and futility in which citizens see no point in protesting or taking action as they expect no impact from it. The model/theory of change underlying the intervention holds that disempowered, marginalized people must feel a *power within*: people realizing they have rights and that those elected should serve them. This allows them then to build *power with*—the coming together of various forms of association around common issues—to achieve *power to*—asserting their rights, campaigning, and mobilizing. This exercise in active citizenship allows people to exercise *power over* key stakeholders. By promoting *power within, with*, and *to*, the project sought to enable people to raise their issues with those in authority and holding power, in whichever way they choose, including digital ones (see Pansardi and Bindi, 2021 for a critical account of the different concepts of power; in the present project it is conceptualized in relation to empowerment). Increased pressure from citizens for better delivery of public services is then expected to lead to local institutions being increasingly compelled to respond.

Using digital technologies to enhance animation approaches is also theoretically grounded in governance and social network approaches. The concept of *governance* includes more than the national government at the country level but includes the operation of formal power at national, regional, and local levels, as well as the way that informal powerholders influence those in power, and civil society engages with, and influences, formal powerholders (see Bevir, 2012 for a general overview). Good governance institutions are transparent and accountable to citizens, ensure that citizens' views and experiences are considered, and work to ensure that their needs are met (Smith, 2007; Rowlands, 2014). This project aimed to increase this self-awareness and power through online channels (see Criado et al., 2013 for a general discussion of the role of social media in governance). The internet brought new ways of socializing and instead of relying on closely-knit, location-based social ties, people moved into more fluid social environments (Wellman, 2001). This can enable new digital relationships with others that were previously unreachable: Digital platforms can thus be used to create new social ties.

Against this background, social media can become a way to raise local issues and join conversations, as well as mobilize other people and create online social networks. It builds people's capacity and skills so they can become active digital citizens. A further function of social media is that it allows obtaining (new) information (e.g., about one's own neighborhood). The project under consideration worked with animators and influencers to establish communication channels that facilitate the creation of and transition from power within to power with and power to. This can be further theoretically conceptualized as a form of *network governance* (Keast, 2022) and *social capital* creation (Lin, 2001). In this regard the animators and especially influencers function as brokers in a social (online) network (Kadushin, 2002) creating bridging social capital if authorities respond to citizens' demands. This is well in line with Putnam's (2000, p. 411) notion of bridging social capital in offline communities: "To build bridging social capital requires that we transcend our social and political and professional identities to connect with people unlike ourselves." Further, as animators (more details below) are well embedded at the village level, they facilitate both bridging (weak ties) and bonding social capital (strong ties) at the village level. While network governance structures are more fragile than other forms of governance, they can be more effective, for example in the transmission of new information (Granovetter, 1973; Park et al., 2018), a key aspect of the "Governance and Accountability through Digitalization" project.

Study design

Against the background of the Cybercrime Act and as a continuation of the former project, the "Governance and Accountability through Digitalization" project was launched in 2017. The project mobilized different actors, online and offline. The primary mechanism to achieve the project's aims relied on placing the power and information of the internet in the hands of roughly 200 community animators from four districts (in the regions of Arusha, Mtwara, Kigoma, and Geita) through the provision of smartphones and training workshops on the use of available associated technology such as search engines, WhatsApp, Facebook, Twitter, and other social media platforms, etc. These online mechanisms came in addition to offline interactions between animators and key stakeholders like government officials.

The selection process of animators was implemented by partners and supported by Oxfam. Villages were selected where at least a 2G connection was available with 3G being preferred, and the focus was on villages that had taken part in the previous

"Chukua Hatua" project[2]. 62 villages were identified in addition to the Nduta refugee camp in the regions of Mtwara, Kigoma, Arusha, and Geita. In these villages, animators were selected using the following criteria (see also Pretari et al., 2019, p. 14):

- has taken part in animation activities (for Oxfam or other organizations),
- can read and write (this criterion may not have been met in very rare cases if the animator was very active and influential in the community),
- is not a political party leader, or involved in politics, nor a leader of the village/ward government,
- is a resident of the village/locality,
- is confident, can explain issues clearly, is concerned about issues and bringing about change in their locality.

While animators thus generally had prior experience in activism, only about 20 percent had used smartphones/social media platforms before. The project strategy relied on working with both women and men animators to consider gender dynamics and the fact that women citizens may feel more comfortable talking to other women, particularly on issues related to violence or discrimination, and ultimately ensure representation of women and men citizens' voices. A total of 50 animators per region were involved in the project. Partners settled on different strategies to determine the number of animators per village, and the number of villages involved. Ten villages are part of the project in Mtwara and five in the host communities in Kigoma, each with five animators. In Arusha, 25 villages are part of the project, with between one and four animators per village; in Geita, 21 villages are part of the project, with between one and six animators per village. It is important to note that these villages and regions have specific dynamics, are embedded in specific contexts, face specific governance issues, and feature a particular setup of animators. For example, in Kigoma, half of the animators were refugees fleeing Burundi who live in the Nduta camp, and the other half were members of host communities. In Arusha, animators lived in the Ngorongoro district, a district with long-lasting land disputes between the Maasai people, the government, and companies.

The project also sought to strengthen the link between local activism enhanced by digitalization through animators, and national influencing, through the mobilization of influential bloggers and social media users (later on referred to as *influencers*). These influencers were online users who had reasonable followership on social media platforms (amount of people who followed them; followed by leaders, high profile individuals, etc.) and whose social media posts were more likely to attract engagement from diverse audiences. Substantially, they are users who were posting mostly about issues/topics that are core to the human rights agenda and are using social media platforms for social good. Influencers thus had prior experience with social media use, but not all of them had prior experience in activism.

Animators and influencers were not paid to participate, but Oxfam provided the animators with mobile handsets and a monthly airtime allowance of 30,000 Tanzanian shillings (equivalent to 12 US dollars). They were also provided with solar chargers to charge their phones since most of them came from rural areas with limited or no electricity. In addition, Oxfam ensured there was ongoing technical support from local partners should animators need support using their digital devices. The most important incentive was the expectation and experience of receiving immediate responses and solutions from key policy makers and duty bearers on issues they had raised.

Participants used different social media platforms to engage with the community and duty bearers and there were further mechanisms employed to supplement the use of online platforms and offline activities. The different digital platforms were used differently. Twitter proved to be the most popular amongst the leaders, elites, and influential business leaders in Tanzania, making it useful to reach these key stakeholders. On the other hand, WhatsApp was effective for organizational tasks: WhatsApp groups were used by animators to coordinate, share information, chat on issues, and agree on topics and strategies before going public. WhatsApp also proved to be effective to reach duty bearers at the regional and district level in Arusha.

Radio programs facilitated debate between citizens and duty bearers, raising awareness on various issues related to human rights and social services. Weekly Twitter debates and regular YouTube live streaming sessions were held. Participants were required to take part in these weekly debates and use the hashtag *#ChukuaHatua* to highlight various community challenges and demand responses and actions from policy makers and duty bearers. These live streaming sessions provided an alternative to mainstream media, as well as a link for the community animators from rural and urban areas to share their experiences. The social media influencers played a key role in capturing the attention of the public during these events.

Throughout the project timeline, animators were provided with introductory and refresher trainings (and certificates of attendance) on animation and on how to effectively use digital tools to raise issues that are important to their communities. In the training workshop on digital tools, animators received smartphones and were shown how to use them and their technical features like the camera to take photos and videos, as well as how to make use of existing social media platforms like Facebook, WhatsApp, and Twitter. Participants were also required to establish a social media strategy—they selected issues that were relevant to their communities and developed a work plan and timeline to address those issues. They created an issue-based calendar stating in which month they planned to focus

2 For this previous project, villages were selected by partner organizations based on the relationships they had held already and the trust they had built in the past (see on the former project Smith and Kishekya, 2013; Green, 2015).

on which topic. This was also the case for influencers who tweeted about the selected topic during the weekly debates. Participants were further trained on the relevant laws governing digital platforms. They were taught to understand key contents of the Cybercrime Act and were encouraged to post within the guidelines of the act. The influencers received this training from officers of the Tanzania Communications Regulatory Authority who are the key implementers/overseers of the act. Additionally, a sensitization workshop took place with civil society partner organizations and leaders/officials (at village, ward, district, and regional levels).

Methods and data

In the following section, we will provide details on the Twitter data collected and the statistical approaches used in our case study. The analyses presented here are building on the previous Oxfam impact evaluation (see Pretari et al., 2019) where we retrieved and analyzed over 130,000 tweets at the end of the official project timeline (March 2019). We extend this now and particularly focus on the long-term sustainability of this developmental intervention. Having this possibility is a unique advantage of online data in comparison to other data sources in developmental research which we want to highlight and explore in the following analyses.

Data collected

For our analysis, we collected Twitter data from all ∼200 animators and influencers involved in the project to analyze both, the animators' and influencers' behavior during the implementation of the project, and the potential long-term effects.

In the past, research efforts on Twitter were severely limited due to restrictions imposed by the application programming interface (API). In 2021, a new academic research track was launched by Twitter, allowing an expansion and improvement of data collection. This allowed us to now collect the complete Twitter timeline of all animators and influencers involved in the project since the creation of their accounts. We thus follow an elite-centered approach when collecting data, focusing on specific user accounts. Data was collected using the R-package *academicTwitteR* (Barrie and Ho, 2021).

It was attempted to collect the complete Twitter activity of all of the 194 animators and influencers (see Table 3 for descriptions of the users and tweets collected). However, some Twitter handles referred to profiles that did not exist, so user data of only 181 profiles was retrieved. While this only affected a small number of Twitter accounts in most regions, almost one fifth of accounts are missing for Kigoma, which might influence the data if these are systematic losses (for example, it might be that primarily non-active Twitter users have misremembered their Twitter handle). Past tweets were retrieved from a total of 168 Twitter users. Most tweets in the complete dataset come from the 24 influencers (98 percent). The 13 users for which no tweet data could be obtained either had a private profile, which could not be accessed, or have never tweeted. The oldest tweets for the animators and influencers date back to May and June 2009, respectively. The most recent tweets are from October 2021 which was set as the limit during data collection.

We also collected Twitter data on relevant officials in Tanzania. These relevant officials include national level leaders [such as (prime) ministers, the vice president, and the president], local level leaders (on the level of the village, district, and region, as well as councilors and members of parliament), institutions relevant to the project (such as the public electricity company, the communications regulatory authority, surface and marine transport regulatory authority, the ports authority, or the national bank), non-governmental organizations and Oxfam partners. The list of officials included 56 user accounts and was created by a subject matter expert (the complete list can be found in the Appendix). We were able to retrieve profile information of 49 and tweets of 48 of those accounts.

Methods

We will employ several different analytical strategies to shed light on the (long-term) effectiveness of the intervention. To answer our research questions, we first describe the extent of activity across time focusing on the animators' and influencers' Twitter usage from the start of the intervention. We then describe the content of all their tweets and identify the topics covered by counting the most frequent words. The majority of tweets collected are written in Swahili which makes it difficult to use more advanced out-of-the-box solutions to address natural language processing tasks, as these solutions are most often based on English texts.

The focus of our descriptive analyses lies on changes throughout time. During the project, Oxfam conducted several workshops to train their participants on the use of digital media. Thus, after the description, we assess whether these workshops were impactful. We focus on a refresher workshop on digital tools which took place on 4 weekends in July and August 2018, each weekend taking place in a different region. This allows us to make use of a difference-in-differences-design (DiD) (Angrist and Pischke, 2008). DiD is one of the most common approaches for identifying and estimating the causal effect of experiencing a treatment on some outcome.

In the canonical DiD, two groups, a treatment group (T) and a control group (C), are compared across two points in time, before treatment (pre) and after treatment (post). In this setting, the simple DiD estimator is the difference between the differences in the treatment group and the differences in the control group:

$$\hat{\delta}_{DiD} = E\left(\Delta y_T\right) - E\left(\Delta y_C\right)$$
$$= \left(E\left(y_T^{post}\right) - E\left(y_T^{pre}\right)\right) - \left(E\left(y_C^{post}\right) - E\left(y_C^{pre}\right)\right) \quad (1)$$

TABLE 3 Description of user and tweet data.

		Mtwara	Geita	Arusha	Kigoma	Influencer	Total
User-level information	Users on list	50	48	42	26	28	194
	Users with retrievable profile information	49	45	41	22	24	181
	Users with retrievable tweets	45	42	39	18	24	168
	Retrieved users' average followers count	160.6	95.4	54.0	228.7	49,358	6,652
	Retrieved users' average following count	150.5	166.1	72.8	520.5	3,245.5	592.1
	Users who received replies from officials	7	1	0	5	14	27
Tweet-level information	Tweets posted (excluding retweets)	10,626 (2,761)	5,378 (3,606)	1,497 (1,171)	10,649 (4,539)	1,575,567 (925,568)	1,603,717 (937,645)
	Hashtag used	23.4%	21.3%	30.4%	36.3%	18.2%	18.3%
	User mentioned	90.3%	78.9%	91.8%	70.8%	66.4%	66.7%
	Mentioned *chukua hatua*	2.9%	4.2%	6.3%	4.0%	0.13%	0.19%
	Average engagement received (excluding retweets)	1.9 (1.5)	1.2 (0.79)	1.2 (0.77)	1.8 (1.3)	1.4 (0.71)	1.4 (0.72)
	Replies received	10	1	0	11	579	601

In this setup, the untreated group never participates in the treatment and the treated group receives the treatment in the second period. In cases of more than two time periods and different treatment times for different units, the leading approach to estimate the effect of the treatment is to use a two-way fixed effects linear regression. However, a number of recent methodological papers have raised concerns about using the two-way fixed effects model with multiple time periods. Particularly, the model is shown not to be robust to treatment effect heterogeneity (De Chaisemartin and d'Haultfoeuille, 2020; Goodman-Bacon, 2021; Sun and Abraham, 2021). To tackle this issue, Callaway and Sant'Anna (2021) have proposed the use of a flexible DiD. They generalize the 2 x 2 DiD in a multi-group and multi-timing setting by computing group-time average treatment effects. With this approach, individual treatment effects for each combination of treatment-timing-group and control group (either never-treated or not-yet-treated) are estimated. These different treatment effects are aggregated in a second step to a group- or time-averaged treatment effect. The model assumes staggered treatment adoption, parallel trends, and no treatment anticipation. Applied to our context, this means that the animators based in Mtwara form the first treatment group and they are then compared to those in the other regions (as they have not taken part in the workshop yet, in other words they are "not-yet-treated"). Each region is thus then compared to the other groups. The dependent variable in our first dynamic DiD is the number of tweets sent; a measure of activity and the DiD thus measures whether tweeting activity increased after the refresher workshop.

In the next step, we focus on engagement received instead of activity, providing insight into how the level of engagement changed over time. First, we analyze the general popularity of the animators' and influencers' self-written tweets. Retweets are excluded as they do not represent the animator's and influencer's content as clearly (they might be retweeting more popular tweets). Again, we assess how this has varied over time and whether it was causally affected by the refresher workshop employing a dynamic DiD. Engagement is then defined as the sum of the retweet and the like count of tweets. Next, we focus on a second type of engagement: that between the project participants and key stakeholders. As this is a very different context, we focus on replies. A reply on Twitter is a time-stamped response to another tweet. It is a way to join a conversation. While follower-followee relationships can be a

measure of popularity (using the follower relationship see e.g., Verweij, 2012; Hofer and Aubert, 2013), and retweets can act as a signal of endorsement (using retweets see e.g., Conover et al., 2011a,b), replies are a measure of active interaction (Sousa et al., 2010; using replies see e.g., Bliss et al., 2012; Gaisbauer et al., 2021). A follower-followee approach is less meaningful in our setup as official accounts (such as accounts of institutions) do not tend to follow (many) others. There were 702 instances in which officials replied to animators or influencers. Most of these replies (676) have been toward one of 14 influencers. Only a few animators per region have been in conversation with officials (see Table 3; please note that while we find 702 instances in which officials replied, we only find 601 undeleted, accessible, and unique tweets which have received a reply)[3].

We use this information on key stakeholders to create a social network between officials and animators. We create an undirected, two-mode network with ties from reply-sending key stakeholders to reply-receiving animators/influencers. Again, we are making use of the longitudinal nature and compare social networks at different points in time (before the workshop, after the workshop, after the end of the intervention. We plot the networks and describe basic characteristics, making use of the R-package *igraph* (Csardi and Nepusz, 2006).

As a last part of our quantitative analysis, we go beyond description and ask which features of tweets are important in generating, on one hand, engagement with the general Twitter public, and, on the other hand, a reply from key stakeholders. We employ logit models to investigate these questions. Our data source to explain engagement are all tweets of animators and influencers which are not retweets (n = 937,645) while we work with the complete set of tweets for the latter analysis (n = 1,603,717). The level of engagement tweets receive varies greatly (mean 8.93, SD 108.94, minimum 0, maximum 70,452) and the majority of tweets receive no engagement at all (58.9 percent of tweets). As it is not our goal to focus on explaining what goes viral (see for such analyses in other contexts for example Zadeh and Sharda, 2022 or Pressgrove et al., 2018) we simplify our analysis by asking which tweets receive any engagement at all and thus create a binary measure. Receiving replies by officials is a rare occurrence; n = 601 (retrievable) tweets have received at least one reply (see Table 3).

In our models, we ask whether tweets referencing the project are particularly successful; to do this, we focus on the key term *chukua hatua*—the term is used as a hashtag to unite those active on Twitter and its usage has also been promoted through the Oxfam-led workshops. In line with our other analyses, we are further differentiating three project phases to assess to what extent engagement has varied throughout time and in the long term. We additionally include an interaction effect between the project phase and the project term *chukua hatua*. This setup will allow us to assess whether such strongly topic-related related tweets receive engagement from the public and key stakeholders and whether a possible effect of this project-relatedness has varied throughout time. We also test to what extent key stakeholders were more likely to reply to posts that were important to the public (by including logged engagement as an independent variable). To check the robustness of the keyword effect, we control for user- (whether they were an animator or an influencer, their geographic region, their popularity and activity measured as logged numbers of followers, followees and previous tweets) and tweet-level features (whether specific technical features were used, i.e., hashtags and mentions). Given that one participant has generally made multiple tweets, we employ cluster robust standard errors. We estimate four different models for both dependent variables. We run a set with and without control variables. In the first one, we include the total dataset; in the second model, we focus on the tweets posted after the beginning of the project (reducing the dataset to n = 430,866 for engagement, and n = 941,764 for replying behavior with n = 242 replies).

Results

We present the results of our analysis in the following subsections. We first focus on the tweets themselves over time—their quantity and their content. In a second step, we focus on engagement with tweets.

Activity on Twitter over time

Figure 1 shows the relative frequency of tweets of the animators (number of tweets per day divided by the number of registered Twitter users) in all of the four regions since May 2017; Figure 2 focuses on influencers only. Areas shaded in red refer to days of training, workshops, or summits organized within the setting of the project. All animators took part in a training on animation, a training on digital tools, and a refresher workshop. The area shaded in orange highlights the time without airtime support (after the end of the project on 31 March 2019).

In the case of the animators, tweeting behavior is clearly spurred by an Oxfam workshop. During the workshop on digital tools, smartphones were handed out to the animators,

3 We also want to note that some discrepancies in numbers reported here and in the previous impact evaluation (Pretari et al., 2019) stem from the fact that only publicly available users and tweets are being collected via the API. If user accounts as a whole or specific tweets are temporarily or permanently deactivated/deleted, this information will not be retrieved. While information collected at different points in time could be merged to achieve a dataset of better quality, we want to respect users' right to be forgotten and their ability to delete, deactivate, or privatize their information previously shared. We thus only use data that is available at the time of data collection (November 2021).

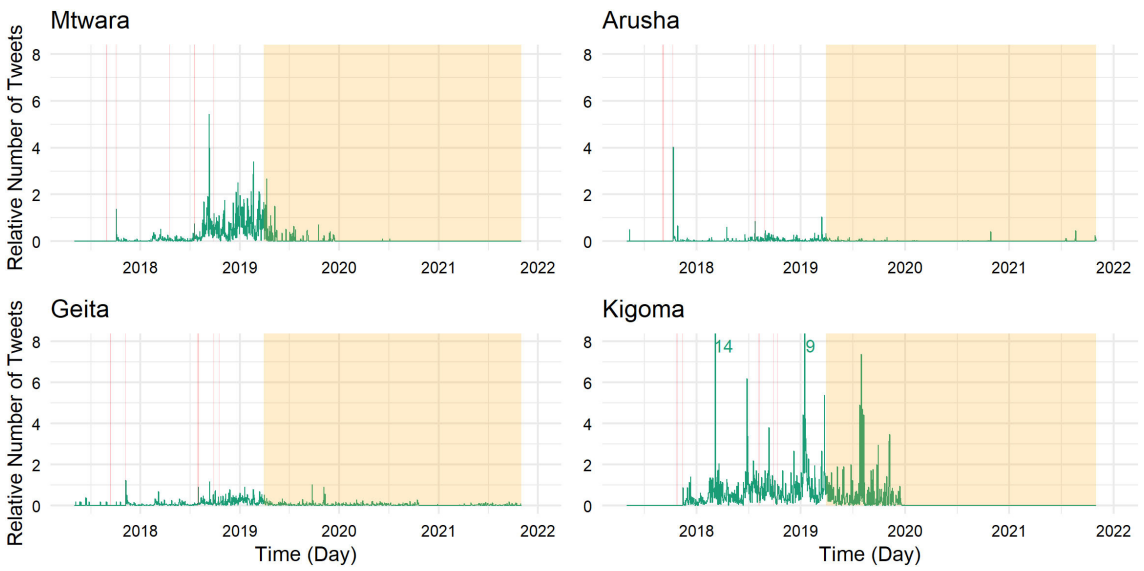


FIGURE 1
Twitter activity of animators across time per region.

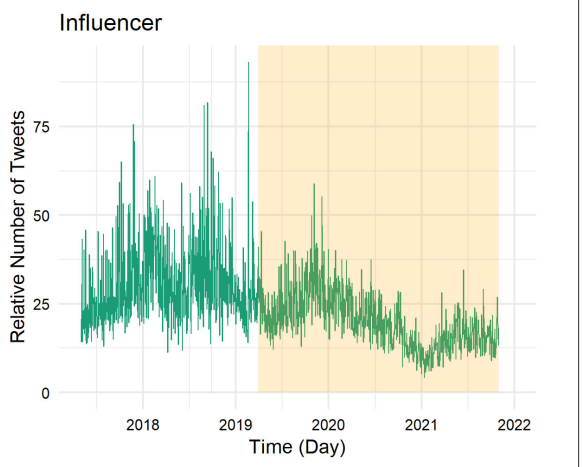


FIGURE 2
Twitter activity of influencers across time.

and they were instructed on how to use them and social media, which had a clear effect on Twitter activity: Before this workshop in October/November 2017, we observe next to no tweets.

During the time frame of the project, we observe levels of tweeting activity that are rather stable, but comparably low in Geita and Arusha. More activity is observed in Kigoma and Mtwara. These regions also show large variations in the frequency of tweets and some remarkable peaks in activity. Airtime support ceased at the end of March 2019, but tweeting continued. Both in Kigoma and Mtwara, animators were still active up until the end of 2019 (more so in Kigoma). In Arusha, only a few scattered tweets are observable since the end of the project. Geita is an exception to the other regions where continuing tweeting activity is observable, however to a lesser extent than before.

Focusing on the influencers, the pattern of activity looks very different (see Figure 2). The influencers are generally much more active, sending on some days almost 100 tweets (per influencer). They have joined the Oxfam project as active tweeters and have thus already been registered and tweeted before the project started. Their tweeting activity is thus expected to cover much more than just the project's time length. We observe a reduction in their tweeting activity in the recent past, especially starting the second half of 2020. Even though they reduced their activity, there is still no day in which they do not post any tweets. However, from the descriptive image, we do not observe any project-related changes.

While the introductory workshop on digital tools sparked the animators' online behavior, what was the effect of the refresher workshop they received around 10 months later? We employ a dynamic DiD to assess its effect (see Figure 3). The Oxfam-led workshop took place over the course of a weekend (three days) and the first day of the workshop is considered the first day *post* treatment (time 0). We find no significant effects of the workshop on tweeting activity. On the third day after the treatment (i.e., the last day of the workshop), tweeting activity tends to increase on average, while on the fourth day after the treatment (i.e., the first day after the workshop) tweeting activity tends to decrease,

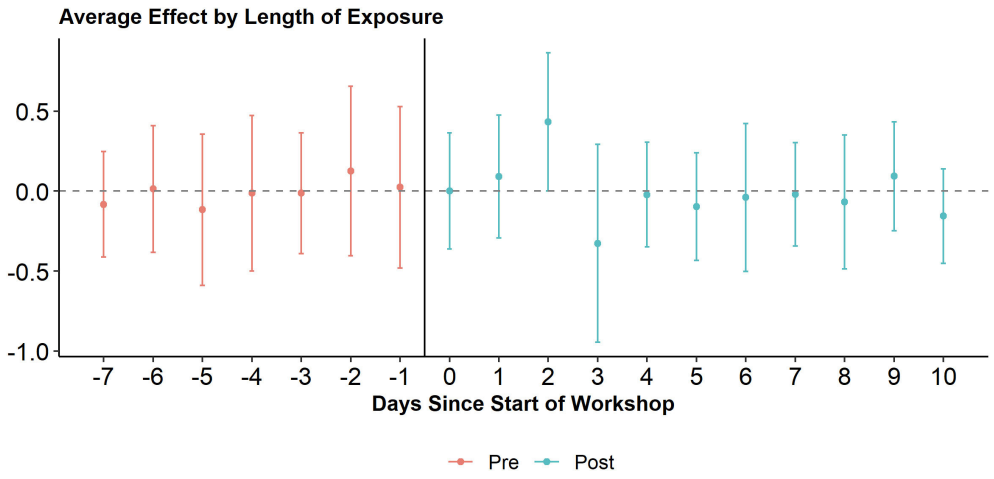


FIGURE 3
Average effect of Oxfam workshop on tweeting activity of animators. Including 95 percent confidence interval.

TABLE 4 Most frequent words per group and project phase.

Group	Project phase			
	Before project	**First phase**	**Second phase**	**After project**
Animators	No words occurred more than 5 times.	Kibondo	Chukuahatua (take action)	Mbogwe (vegetables)
		Kijiji (the village)	Serikali (government)	Kazi (work)
		Serikali (government)	Kijiji (village)	Serikali (government)
		Kazi (work)	Maji (water)	Jamii (society)
		Kigoma	Mtwara	Wilaya (district)
		Wananchi (citizens)	Wananchi (citizens)	Maendeleo (development)
		Mtwara	Kibondo	Wananchi (citizens)
		Shule (school)	Waraghabishi (animators/community	Chukuahatua (take action)
		Tunaomba (we pray)	activists)	Maji (water)
		Waraghabishi	Jamii (society)	Kijiji (village)
		(animators/community activists)	Kazi (work)	
Influencers	Tanzania	Elimikawikiendi	Tanzania	Tanzania
	Elimikawikiendi	Tanzania	Chukuahatua (take action)	Kazi (work)
	Rais (president)	Twittergulio (Twitter)	Maji (water)	Vijana (young people)
	Kazi (work)	Kazi (work)	Kazi (work)	Elimikawikiendi
	Mkuu (principal)	Serikali (government)	Elimikawikiendi	Mzee (old man)
	People	Watoto (children)	Serikali (government)	Aatoto (children)
	Jmaa (relatives)	Rais (president)	Vijana (young people)	Maana (meaning)
	Maana (meaning)	Mtoto (child)	Watoto (children)	People
	Mama (mother)	Mkuu (principal)	Mwaka (year)	Mwaka (year)
	Time	Nchi (country)	Jamii (society)	Mtoto (child)

but these changes are not significant on a five percent level. Overall, tweeting levels seem rather unaffected by the workshop. However, it is important to note that we cannot sufficiently take into account regional differences. The four regions in this project did exhibit very different dynamics, making comparison difficult.

Going beyond the quantity of tweets made by animators and influencers, we also analyzed their content. The 10 most frequent words per project phase (differentiating four phases: before the project started, before the refresher workshop, after the refresher workshop, after the end of the official project) and per participant type are shown in Table 4. As shown before,

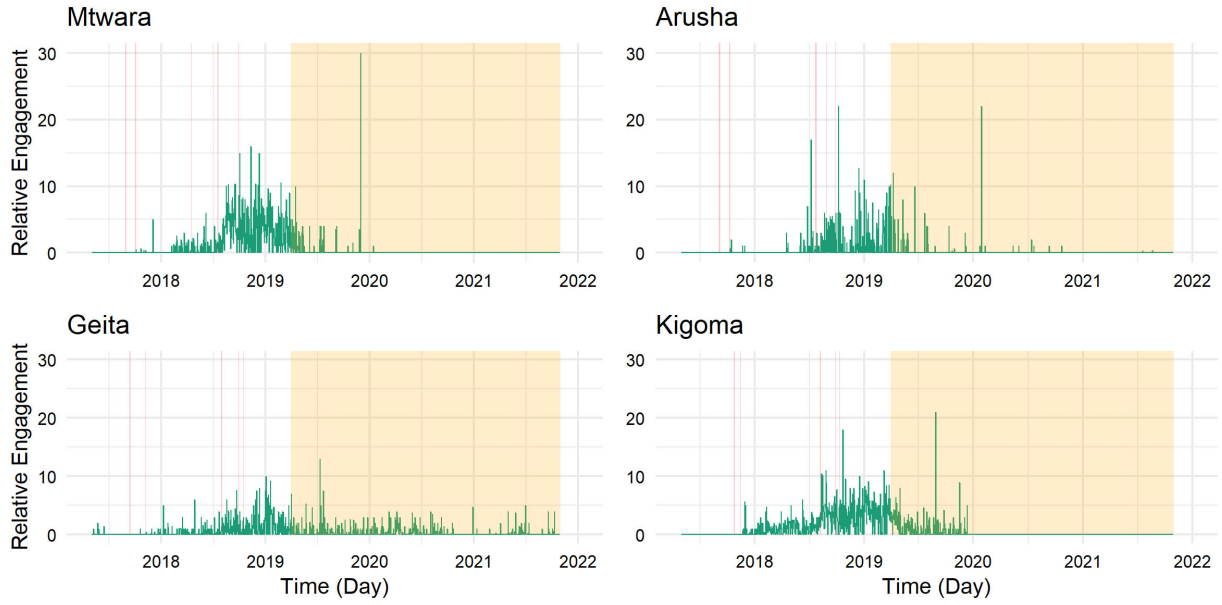


FIGURE 4
Twitter engagement received by animators across time per region.

animators have generally not been active on Twitter before the project started and no frequent words are retrieved. Since then, both animators and influencers have used Twitter. Both groups mostly tweet in Swahili and to a lesser extent in English (influencers use English more frequently than animators). During the project, animators often tweet about and explicitly mention their region or district (Kibondo, Mtwara, Kigoma) and often talk about the situation in their *village*. Influencers, on the other hand, more broadly mention the country context of *Tanzania*. In the first phase, animators discuss issues around the *government*, *work*, *citizens,* and *school* most often. In the second phase, both animators and influencers start to more frequently link and refer to the project by using the term (and hashtag) *chukua hatua.* Both groups also more frequently discuss issues around *water*. After the project, *chukua hatua* is not the most frequent word mentioned, but still belongs to the 10 most frequent words used in the group of animators; influencers, however, do not refer to the project that often anymore. For animators, references to the regions also seem to have become less, while the terms *vegetables* and *development* have increased in frequency. Over the timeline covered, influencers discuss a variety of topics: Across all time frames, they also often raise issues around *work*, the *government*, *society,* and parts thereof like *children* or *young people*. *Elimikawikiendi*, a term often occurring as a hashtag, was used during the popular, weekly Twitter session (organized by a company[4]); the hashtag united various Twitter users. Participating influencers and animators used this to share issues of concern from their localities.

Engagement with the public and key stakeholders on Twitter over time

In a digitalized world, using social media has become a way to raise local issues to the public and it allows joining conversations with duty bearers at the national, regional, or district level. In this section, we will describe the changes in the public's and the key stakeholders' engagement with the tweets over time and address the question of which kind of tweets are most likely to receive engagement and replies.

Using the same labeling as in Figures 1 and 2, Figures 4 and 5 display the relative levels of engagement for the tweets (excluding retweets) of the animators (count of retweets and likes per day divided by the number of tweets posted on that day) in the four regions (Figure 4) and focusing on the influencers (in Figure 5) since May 2017. This provides insight into how the popularity of the content put out by the animators and influencers changed over time and extends the previous section which focused on the volume of posts.

In the case of the animators, the level of engagement per tweet was highest in all regions in the later phase of the project (after the refresher workshop). This might suggest that their tweets might have become more effective in generating the public's interest after the workshop. While fewer tweets were made when the project ended, those few tweets still received the

4 It is organized by the Elimika Wikiendi Company LTD (https://www.elimika.co.tz/).

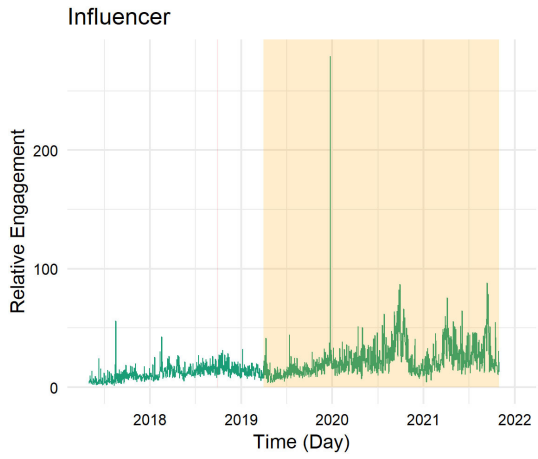


FIGURE 5
Twitter engagement received by influencers across time.

highest levels of engagement. The animators in Geita are the only ones still regularly tweeting, and their tweets still receive average levels of engagement from the public.

In terms of posts made by influencers, we observe a relatively stable level of engagement during the project's timeline, and an increase with some notable peaks since (see Figure 5). Influencers' tweets have received more engagement since the project ended. It might be the case that the topics they have tweeted about in the more recent past are more engaging and more popular, but it might also be that, as the quantity of tweets has decreased, only the more successful influencers are still active on the platform.

Figure 4 suggests that tweets by animators generate increasing levels of engagement in the long term after the refresher workshop. To test a potential causal effect of the workshop more explicitly and directly, we again make use of a dynamic DiD (see Figure 6). We use the same approach as when analyzing the workshop's effect on tweeting activity. We find no significant effects of the workshop on the average amount of engagement tweets receive: They seem unaffected by the workshop.

While the previous analyses have focused on the endorsement of tweets from the public *via* liking and retweeting, we will now focus on users' ability to connect with key stakeholders such as government officials, public service providers, non-governmental organizations (NGOs), and civil society organizations (CSOs) on Twitter. We are again making use of the longitudinal nature and compare social networks built through the reply function at different points in time: We take (A) all tweets from May 2017 up to June 2018 (refresher workshop), (B) all tweets from July 2018 up to March 2019 (end of the intervention), and lastly (C) all tweets since April 2019.

The three networks are shown in Figure 7. Thicker edges reflect multiple replies. In the time frame from May 2017, 40 different nodes are part of the reply network (across the complete time span, there are 43 nodes). In network (A), these nodes share 141 edges; in network (B), they share 53 edges and in network (C), they share 101 edges. The number of edges shared between two nodes can vary as multiple edges are allowed: In (A), one influencer has received 42 replies from a national leader. The maximum of replies exchanged between the same two actors in a timeframe are 9 and 11 in time frames (B) and (C), respectively.

In the earliest timeframe, the network is split up into two components (containing 18 and 7 nodes) while the other 15 are isolates. After the refresher workshop, in timeframe (B), we observe 2 larger components containing 14 nodes and 5 nodes, respectively, 3 dyads, and again 15 isolates. Officials have still replied since the end of the project as seen from network (C) which is made up of 2 larger components (containing 22 and 4 nodes respectively), 2 dyads, and 10 isolates.

Across all time frames, the largest components refer to 13 different influencers and 4 animators which were in exchange with several different officials. Particularly, many replies were exchanged between three different national politicians and the influencers and animators surrounding them. They are pictured in the bottom left of the plotted networks. One regional duty bearer in particular has been replying to many tweets posted by influencers. While we cannot argue that this is caused by the refresher workshop—given that the number of replies is generally very small, and we look at long time frames with many unobserved characteristics—we do observe an increasing number of actors being involved in this discussion network across time and that these discussions have not stopped when the project did. Beyond the cluster in the bottom left, the nodes in the top right capture the fact that Oxfam and its partners have been in conversation with a number of animators. The separated dyad in the upper left corner reflects conversations between an influencer and an NGO or CSO coalition.

Since May 2017, 27 animators and influencers have received replies from official accounts; all others have not. Which tweets are most likely to receive replies? And which tweets receive the most engagement from the public? We have tried to explain whether tweets receive any engagement and whether they receive replies by officials using logit models as described in section Methods.

The results are shown in Table 5. Models 1.1.1 and 1.1.2, which only include tweets posted since the start of the project, suggest that tweets made in the later phases of the project (second phase after the refresher workshop and after the project as a whole) are more likely to receive any engagement (at least on the 10 percent significance level). Using the term *chukua hatua* makes tweets more likely to attract engagement, even more so in the second phase of the project (model 1.1.1). However, this is only observable when not controlling for other

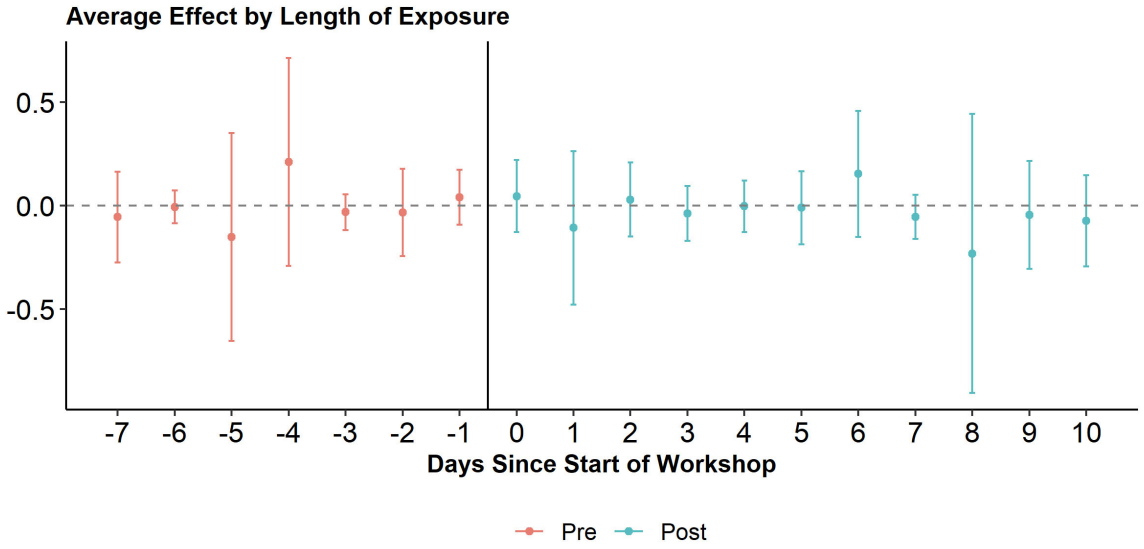


FIGURE 6
Average effect of Oxfam workshop on engagement received by animators. Including 95 percent confidence interval.

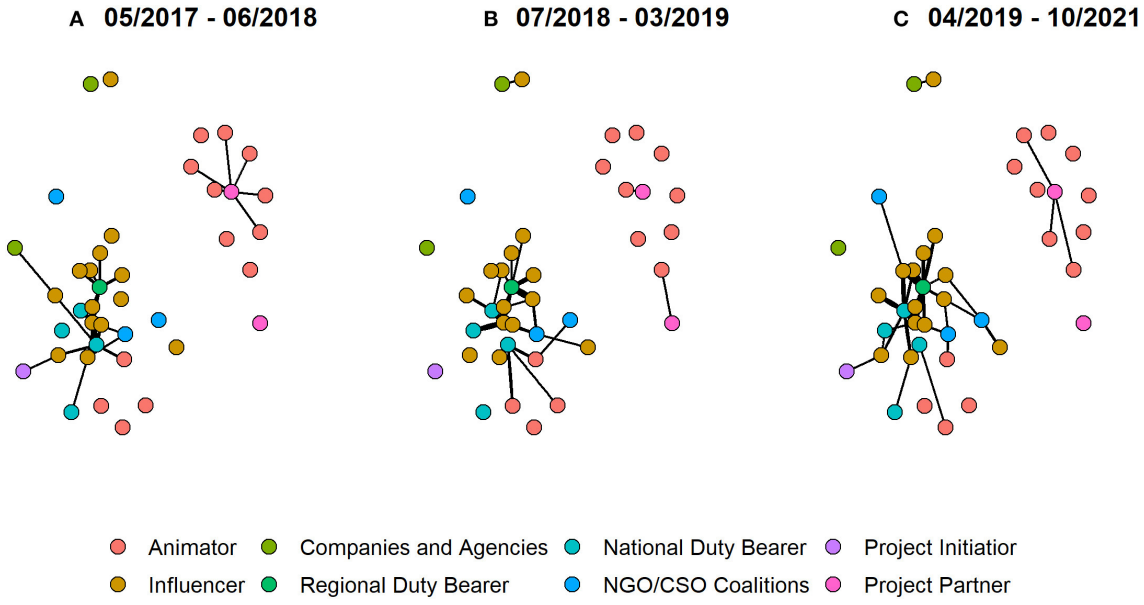


FIGURE 7
Reply network between officials and animators/influencers across three different time frames. Part **(A)** refers to tweets which were posted during the first phase of the project (May 2017 to June 2018), part **(B)** refers to tweets of the second phase of the project (July 2018 to March 2019), and part **(C)** refers to tweets posted after the project ended (April 2019 to October 2021).

user- and tweet-level features (model 1.1.2). There, we only observe that tweets mentioning *chukua hatua* and sent after the project are less likely to receive engagement. Across all observations (models 1.2.1 and 1.2.2), we see that tweets made before the project started are much less likely to receive any likes or retweets than those sent in the first phase, and that, again, those made in the second phase or after the project are, on average, more likely to receive engagement from the public. Working with the complete set of tweets now, we observe a positive project word effect even when controlling for other features. Over time, making use of the project word seems especially positive in the second phase of the project, but can have a negative effect after the project (model 1.2.2 only). These patterns suggest that since the project started and also since the first phase of the project, engagement with tweets from the general public has increased. Making an explicit reference

to the project is generally positive (positive main effect of mentioning *chukua hatua*) during the project timeline, but not since (negative interaction effect of mentioning *chukua hatua* and after project). Our control variables suggest differences between animators and influencers, regional differences, and strong effects of the technical Twitter features (tweets using hashtags are much more likely to receive engagement, while those that mention users are less likely).

Turning to the second set of models focusing on replies, we find a different effect of time: Officials were much less likely to reply to tweets in the later phase of and after the project compared to the first phase. Further, while mentioning *chukua hatua* positively influences the probability to receive engagement by the public, it has a very strong negative effect when explaining replying behavior. Officials were much less likely to reply to tweets using the hashtag and this only varied slightly over time as the interaction effects suggest; compared to the first phase of the project, officials were still less likely to tweets mentioning the term *chukua hatua* and to those not including the term, but a little less so. When focusing on tweets sent during the project time, we find that those which received more engagement are more likely to receive a reply from an official (models 2.1.1 and 2.1.2). Again, the control variables suggest regional variations as well as differences between animators and influencers in the probability to receive replies, and an effect of tweet-specific features.

While the models in Table 5 suggest that officials are more likely to reply to tweets that have received higher levels of engagement, they also show that mentioning *chukua hatua* is negatively affecting a tweet's probability of receiving a reply while it increases its probability to receive some level of engagement in the form of a like or retweet. These findings seem conflicting at first sight, but it might well be that tweets mentioning *chukua hatua* are more likely to be at least liked or retweeted once (potentially by other project participants) while it is the viral tweets which are replied to by the officials and which might not mention the project name.

The models on replying behavior of officials also suggest that tweets are less likely to receive replies since the end of the project. Looking at the data in more detail, we find that officials still regularly reply to influencers and that the last occurrence where an animator received a reply from an official dates back to August 2019, when an animator was in a conversation with an official. In the next section, we will look in more detail at the tweets generating replies and high levels of engagement to gain a better in-depth understanding of the data analyzed.

Close-reading of big data

The previous sections have shown potential directions on how to analyze large numbers of tweets using quantitative and automatic—distant—methods of data (and text) analysis. The tweets written and published as part of the project do not only need to be analyzed in such a distant way, but much can be gained from a close-reading and in-depth analysis of single tweets and actors. The quantitative analysis can be the starting point for this endeavor.

Tweets written by animators which generated high levels of engagement (over 100 likes/retweets) were, for example, concerned with the issue of water, and read: "#ChukuaHatua "Water is life" is a statement that was made by our leaders in the eighties. However, to date most citizens, especially in our region Geita in Mbogwe district, have water scarcity. The World Bank did a survey at Masumbwe but water availability is still a challenge. What does the ministry say with regard to that. @[mention of account]" (original in Swahili, own translation). Another often retweeted and liked tweet from Arusha featured an image of a school building that fell into a state of disrepair and called for attention, while a popular tweet from Mtwara was concerned with unfavorable feedback received at a village meeting. Looking into the tweets that have generated replies, we, for example, find a tweet written by an animator who also highlights water being an issue in a specific village and who is asking for help. A member of parliament then asked to clarify where the village is located; from the digital trace data, however, we do not know how this continued and whether any action was taken. In another tweet, a minister is being thanked as destroyed bridges and roads are becoming unblocked, and he replies with a positive message for the future. Qualitative evidence can come in to highlight the effectiveness of Twitter in actual cases. In one of the interviews conducted for the study on which this paper builds, an animator remarked: "We tweeted about the shortage of teachers in primary schools and in <3 months, three teachers were posted". This was not the only incident, as another animator also mentioned that: "We managed to tweet about the land conflict that occurred at our village then the leaders from the districts came to rescue the situation" (Pretari et al., 2019, p. 51–52). Relying on this interview information, we can state that the project has resulted in real-life offline impact even though it cannot be directly seen from quantitative online evidence. While quantitative analysis of digital trace data has been useful to create a greater overall picture, more in-depth insights can be gained by a close-reading of the tweets produced and in combination with qualitative data sources.

Conclusion

The Oxfam-led "Governance and Accountability through Digitalization" project in Tanzania has integrated digital technologies into traditional animation approaches, in collaboration with three Tanzanian regional organizations. We have analyzed the content created on Twitter and how the public reacted to it. The analyses have shown that animators signed up to Twitter and started to post on the platform about

TABLE 5 Logit models on receiving engagement and replies.

	Receiving engagement				Receiving replies			
	Model 1.1.1 Excluding before project observations	**Model 1.1.2 Excluding before project observations**	**Model 1.2.1 All observations**	**Model 1.2.2 All observations**	**Model 2.1.1 Excluding before project observations**	**Model 2.1.2 Excluding before project observations**	**Model 2.2.1 All observations**	**Model 2.2.2 All observations**
Project phase								
(Ref.: first phase)								
Bef. project			−1.605*** (0.247)	−1.519*** (0.191)			0.344 (0.319)	0.548+ (0.314)
Second phase	0.281* (0.114)	0.403*** (0.082)	0.281* (0.114)	0.399*** (0.087)	−0.600* (0.295)	−0.804* (0.338)	−0.588 (0.299) *	-0.803* (0.332)
After project	0.269+ (0.144)	0.622*** (0.099)	0.269+ (0.144)	0.583*** (0.083)	−0.668* (0.291)	−0.735** (0.255)	−0.690 (0.291) *	-0.812** (0.255)
Mentioned *chukua hatua*	0.947* (0.373)	0.478 (0.338)	0.947* (0.373)	0.890** (0.318)	−11.681*** (0.380)	−12.369*** (0.315)	−10.713 (0.377) ***	−11.432*** (0.307)
Before project x *chukua hatua*			−0.170 (0.394)	−0.293 (0.385)			−0.340 (0.499)	−0.582 (0.440)
Second phase x *chukua hatua*	0.902*** (0.265)	0.447 (0.331)	0.902*** (0.265)	0.530+ (0.306)	0.573* (0.280)	0.374 (0.380)	0.587* (0.283)	0.341 (0.349)
After project x *chukua hatua*	−0.366 (0.437)	−1.264* (0.601)	−0.366 (0.437)	−1.218* (0.528)	0.663* (0.311)	0.116 (0.396)	0.690* (0.307)	0.161 (0.409)
Engagement (log)					0.192*** (0.031)	0.207*** (0.034)	0.008 (0.057)	−0.008 (0.069)
Influencer		−3.277*** (0.757)		−3.556*** (0.659)		−2.774*** (0.684)		−1.720* (0.713)
(Ref.: Animator)								
Region								
(Ref.: Kigoma)								
Arusha		−0.405 (0.356)		−0.392 (0.300)		−12.977*** (0.647)		−12.161*** (0.584)
Geita		−0.655* (0.276)		−0.683* (0.266)		−1.143 (1.097)		-1.421 (1.086)
Mtwara		0.562+ (0.336)		0.431 (0.310)		0.197 (0.513)		0.112 (0.473)
Used hashtags		2.118*** (0.178)		1.397*** (0.177)		−0.750*** (0.221)		−0.559** (0.205)
Mentioned users		−0.828*** (0.123)		−0.737*** (0.117)		−0.090 (0.219)		1.223*** (0.305)
Follower count		0.567*** (0.118)		0.605*** (0.101)		0.188 (0.160)		0.104 (0.164)
Following count		−0.057 (0.115)		−0.082 (0.069)		0.248+ (0.150)		0.064 (0.142)
Tweet count (log)		−0.171 (0.122)		−0.232+ (0.125)		0.029 (0.202)		0.431+ (0.229)
Intercept	0.340 (0.297)	0.394 (1.667)	0.340 (0.297)	1.292 (1.561)	−8.275*** (0.344)	−9.689*** (2.733)	−7.868*** (0.301)	−14.506*** (3.030)
Log Likelihood	−281,978.60	−244,819.50	−54,9038.59	−485,484.62	−2,214.06	−2,167.52	−5,287.43	−5,113.42
AIC	563,969.20	489,669.00	1,098,093.18	97,1003.24	4,442.12	4,367.04	10,592.86	10,262.84
BIC	564,035.05	489,833.60	1,098,187.19	971,203.01	4,524.41	4,555.13	10,703.45	10,484.02
Num. obs.	430,866	430,866	937,645	937,645	941,764	941,764	1,603,717	1,603,717

Cluster robust standard errors in parentheses; + p<0.10, * p<0.05, ** p<0.01, *** p<0.001.

10.3389/fsoc.2022.909458

project-related issues. Additional workshops during the project timeline do not seem to have had an effect on tweeting activity or on engagement received. Animators and influencers have started conversations with key stakeholders, but results suggest that influencers and especially animators only rarely received replies to their tweets. Tweets were more likely to receive replies if they also received high levels of engagement. While explicitly referencing the project in a tweet increased the probability to receive at least one like or retweet, it decreased the probability to receive a reply from a key stakeholder. Even though the overall effect of the project seems to be small according to the analyses discussed—while tweeting activity takes off with the project, reply networks are small in scope, and since the intervention ended, activity is minimal—raising issues on Twitter has shown to lead to positive changes. Nevertheless, despite the positive examples of real-life offline impact mentioned in the previous section, it needs to be acknowledged that the digital component of the project alone did not seem to result in the creation of "large amounts" of bridging social capital, effective network governance and changes in power dynamics considering the relatively low level of engagement of authorities with citizen.

The usage of Twitter data in our study on governance in Tanzania allowed an additional perspective on a developmental project. It has shown to be a valuable complement to the more traditional qualitative and quantitative approaches, specifically allowing to time- and cost-effectively obtain an impression of the potential long-term effects of the project. This has provided us with a unique opportunity to assess the sustainability of the project. While animators have started being active Twitter users, most have stopped tweeting since 2020. However, there is still an exchange of tweets between key stakeholders and influencers, as well as some animators. Our case study also has a number of further limitations and challenges. It is important to keep in mind that we only analyze Twitter data; this does not capture activity on all online channels as WhatsApp is another popular digital tool specifically in Arusha at the regional and district level. Further, while we compare four different regions, it is important to remember their specific dynamics and contexts. These differences can hinder the valid comparison between regions. Also, our analyses do not consider accounts that have been deleted or deactivated from Twitter. This means, we do not know whether animators and influencers have been in active conversation with an official account which has been deleted since. We also tend to underestimate the general level of interaction between animators/influencers and officials by only focusing on replies.

Notwithstanding its limitations, our case study functions as a valuable example highlighting the potential of big data in development sociology. While more in-depth analysis can shed further light on the interesting patterns and dynamics in the project context, we aimed to showcase a starting point for digital trace data in intervention studies.

Concluding remarks and recommendations

In this paper, we pointed to the advantages and disadvantages of big data analysis in development research, highlighted examples with a sociological perspective, and provided a more in-depth case study taking place in Tanzania serving as an example application with a value-added of digital technologies. Clearly, big data in its various forms can help to shed more light on development issues and there are innovative approaches to measure and predict important indicators related to poverty, social inequality, etc. Solving such measurement problems is certainly an important contribution to development research and practice. However, regarding analyzing and explaining the effectiveness of development programs and interventions, a key aim of development initiatives and research, the contribution of big data is less straightforward. The reason is that quantitative impact evaluations rely on causal inference built on counterfactual logic, operationalized through treatment and control group(s), and this is not a given if big data are "just" used as an additional data source. In other words: To provide useful insights regarding the effectiveness of interventions, big and digital trace data need to be considered in the research design phase of development research. Our example has limitations regarding such a causal analysis (e.g., regions in our study do not only vary regarding the social media intervention but also other characteristics). Yet it also demonstrates how a causal analysis can be implemented as part of an impact evaluation. Furthermore, our study exemplifies one major advantage of using (big) digital data as part of an intervention study: Digital data allow the study of long-term effects of interventions which is certainly a limitation in most intervention studies. In our case study, social media activity is significantly decreasing in the long term. Yet, this picture would have looked quite different when only considering the actual time when the intervention project was running.

At first sight, our case study might suggest a limited effect of social media on governance in the corresponding regions in Tanzania. However, we also find some sustained social media activity as well as testimonies of online activity as part of the intervention sparking actual changes at the community level. This underscores the importance of combining different data sources and strategies of analysis in development research. Quantitative analysis of big data is not meant to replace other approaches but to complement them, and there is a need for cross-validation. From a sociological perspective, it is furthermore evident that phenomena such as citizenship and governance are complex and multifaceted and hence their study demands a comprehensive research design, combining qualitative and quantitative research components. In fact, our social media case study was part of a much larger evaluation that comprised several components (see Pretari et al., 2019).

TABLE 6 Issues/challenges of big data analysis in the context of development sociology and recommendations.

Issues/challenges	Recommendations
Uncovering causal effects: A control-treatment group design is not inherent to big data.	An experimental setup should be considered in the planning phase of development research.
Estimating long-term effects: Treatment effects often fade away briefly after the intervention which is difficult to measure.	The advantage of big data to be more easily collected repeatedly should be used. This can shed light on the persistence of intervention effects.
Evaluating the impact: Big data might fall short of capturing all aspects of impact.	Typically, big data analysis is one part of a (complex) development research project and should be complemented with other forms of qualitative and/or quantitative data analysis.
Deriving representative conclusions: Big data such as social media data might be based on biased samples.	Researchers should acknowledge and try to estimate a potential sample bias of big data such as social media data.
Collecting ethically sound data: Typically, big data is not produced for research purposes.	Whenever possible, researchers need to obtain informed consent from research participants such as social media users. In any case, they need to consider ethical aspects such as respect for persons, beneficence, justice, and respect for law and public interest.
Engaging people who produced the (big) data: There is a danger that research participants have no say in the research process, as well as of power and racial dynamics to be reproduced in the way knowledge is generated.	Research teams should find ways for the data producers to become the researchers of the data they produce and participate in the research process for example by a strong community engagement.
Providing public access to data: Many studies using big data cannot be replicated due to lack of access to data and code.	In line with general developments toward open data in development research, researchers at both universities and NGOs should strive for making data and code publicly available (while considering ethical and legal restrictions).

Sociology also invites us to reflect on who is considered a knowledge producer and to be attentive to the power imbalances between different epistemologies. We discuss below the potential for big data research to promote a different role for data producers in the research process and provide a table of recommendations (see Table 6).

There are various important aspects that need to be addressed in big data analysis. Particularly, the representativeness and ethical aspects of this type of analysis must be discussed (Lazer et al., 2014; Ruths and Pfeffer, 2014; Townsend and Wallace, 2016). Representativeness can be affected by potential biases in the available data, but also by the fact that users of social media platforms can differ in their characteristics from the general population. It is important to be aware of digital divides and inequalities within the study population when making general claims. From an ethical standpoint, (big) digital data comes with new questions and uncertainties which are also best addressed and considered in the research design phase. Even if only publicly available information is accessed (as is generally the case with Twitter), it is important to keep in mind that users of social media sites make their data available for the purpose of social networking, not to be harvested and used for research purposes. Salganik (2018, Chapter 6) advises to be guided by four principles when facing ethical uncertainty in digital social science research: respect for persons, beneficence, justice, and respect for law and public interest. Townsend and Wallace (2016) also highlight that the terms of conditions of social media sites, the ethical guidelines of the researcher's institution, the privateness of the social media site, the vulnerability of the users, the sensitivity of the research topic, and the potential for anonymization and data sharing must be considered when making use of digital trace data (see also Zook et al., 2017 for responsible big data research). In the ideal case and whenever possible, researchers should obtain consent of their participants. This is especially indispensable when wanting to access more private digital spaces like WhatsApp groups or similar. The use of big data in sociological research also raises the broader question of how to engage the many people who produced the data in the research process. Going beyond the consent stage, engagement means for data producers to shape the research questions, the analysis, interpretation, and ultimately its use. In the setting of development research, the racial division of labor has been documented [see for a recent example the "(Silent) Voices Bukavu series" blog[5]]. We acknowledge that the team of researchers involved in this research is primarily white, living in

5 A blog series which highlights 'the premeditated violence that persists in the process of academic knowledge production', arguing 'that this process is, among other things, responsible for the dehumanization and the erasure of researchers from the Global South, available here https://www.gicnetwork.be/silent-voices-blog-bukavu-series-eng/

and from the so-called Global North. While big data research and computational social science in general promote a culture of open access of codes and availability of data, more effort needs to be devoted to challenge power dynamics and avoid structural inequalities to be reproduced in and through the research field. In particular, big data development sociology as a sector and sociologists as individuals need to challenge racialized power dynamics between research teams and data producers, and find ways for the data producers to become the researchers of the data they produce and participate in the research process.

Our paper aimed at presenting the potential of different types of big data for development sociology and an example case study integrating social media in an intervention program. This can be seen as a starting point for more systematic usage of big data in development research with a sociological focus. It should be clear that this sociological focus comprises a vast number of sociologically relevant topics that can be studied with big data, as well as different techniques such as network or text analysis which can be applied to big data in the context of development research. A sociological focus also entails the integration of a sociological perspective in inter- and transdisciplinary research including a critical reflection on the use of big data in the development sector. We hope that our paper paves the way for much more research on these topics.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

For the Oxfam-led project informed consent was obtained from all participants. While no informed consent was obtained for the Twitter data analysis, only publicly available information was collected, i.e. profiles which were made private or deleted after the project were not collected, and deleted tweets were not collected. Further, Twitter data were not merged with other personal data.

Author contributions

WM (with Oxfam in Tanzania and partner organizations) designed and carried out the project. AP and SL designed and implemented its impact evaluation, in collaboration with WM (and Oxfam in Tanzania and partners). NS, UL, AP, and SL designed the Twitter data analysis. NS collected and analyzed the data. NS and UL prepared the draft manuscript with input from all authors. All authors reviewed the results and approved the final version of the manuscript.

Funding

The presented case study was funded by the Belgian Directorate-General for Development Cooperation and Humanitarian Aid. UL acknowledges support by the Warwick ESRC Impact Acceleration Account.

Acknowledgments

The authors want to thank Oxfam in Tanzania and the following partner organizations for their efforts in designing and carrying out the project: Pastoral Livelihood Support and Empowerment Programme in Arusha, Capacity Building Initiatives for Poverty Alleviation in Geita, Mtwara Society Against Poverty in Mtwara.

Conflict of interest

Authors AP, WM, and SL were employees of Oxfam at the time of the project.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fsoc.2022.909458/full#supplementary-material

References

Abreu Lopes, C., Bailur, S., and Barton-Owen, G. (2018). *Can big data be used for evaluation? A UN Women feasibility study*. Available online at: https://www.unwomen.org/en/digital-library/publications/2018/4/can-big-data-be-used-for-evaluation (accessed September 01, 2022).

Ali, D. A., Collin, M., Deininger, K., Dercon, S., Sandefur, J., and Zeitlin, A. (2016). Small price incentives increase women's access to land titles in Tanzania. *J. Dev. Econ.* 123, 107–122. doi: 10.1016/j.jdeveco.2016.06.001

Allen, C., Smith, M., Rabiee, M., and Dahmm, H. (2021). A review of scientific advancements in datasets derived from big data for monitoring the Sustainable Development Goals. *Sustain. Sci.* 16, 1701–1716. doi: 10.1007/s11625-021-00982-3

Angrist, J. D., and Pischke, J. S. (2008). *Mostly harmless econometrics*. Princeton: Princeton University Press. doi: 10.2307/j.ctvcm4j72

Barrie, C., and Ho, J. (2021). academictwitteR: an R package to access the Twitter Academic Research Product Track v2 API endpoint. *J. Open Source Softw.* 6, 3272. doi: 10.21105/joss.03272

Bevir, M. (2012): *Governance: A Very Short Introduction*. Oxford: Oxford University Press. doi: 10.1093/actrade/9780199606412.001.0001

Bliss, C. A., Kloumann, I. M., Harris, K. D., Danforth, C. M., and Dodds, P. S. (2012). Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *J. Comput. Sci.* 3, 388–397. doi: 10.1016/j.jocs.2012.05.001

Briceño, B., Coville, A., Gertler, P., and Martinez, S. (2017). Are there synergies from combining hygiene and sanitation promotion campaigns: evidence from a large-scale cluster-randomized trial in rural Tanzania. *PloS ONE.* 12, e0186228. doi: 10.1371/journal.pone.0186228

Callaway, B., and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *J. Econom.* 225, 200–230. doi: 10.1016/j.jeconom.2020.12.001

Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A., and Menczer, F. (2011a). "Predicting the political alignment of twitter users," in *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing* p. 192–199. doi: 10.1109/PASSAT/SocialCom.2011.34

Conover, M. D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., and Flammini, A. (2011b). "Political polarization on twitter," in *Proceedings of the international AAAI conference on web and social media* p. 89–96.

Criado, J. I., Rodrigo, S.-A., and Gil-Garcia, J. R. (2013). Government innovation through social media. *Gov. Inf. Q.* 30, 319–326. doi: 10.1016/j.giq.2013.10.003

Csardi, G., and Nepusz, T. (2006): The igraph software package for complex network research. *Int. J. Complex Syst.* 1695, 1–9.

Data2X (2019). *Big Data, Big Impact? Towards Gender-Sensitive Data Systems*. Report (pp. 13-14). Available online at: https://data2x.org/resource-center/big-data-report/ (accessed September 01, 2022).

DataReportal (2017). *Digital 2017 Tanzania*. Report. Available online at: https://datareportal.com/reports/digital-2017-tanzania (last accessed 2022-09-01).

DataReportal. (2022). *Digital 2022 Tanzania*. Report. https://datareportal.com/reports/digital-2022-tanzania (accessed September 01, 2022).

De Chaisemartin, C., and d'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *Am. Econ. Rev.* 110, 2964–2996. doi: 10.1257/aer.20181169

De Walque, D., Dow, W. H., Nathan, R., Abdul, R., Abilahi, F., Gong, E., et al. (2012). Incentivising safe sex: a randomised trial of conditional cash transfers for HIV and sexually transmitted infection prevention in rural Tanzania. *BMJ Open.* 2, e000747. doi: 10.1136/bmjopen-2011-000747

Edelmann, A., Wol, T., Montagne, D., and Bail, C. A. (2020). Computational social science and sociology. *Annu. Rev. Sociol.* 46, 61–81. doi: 10.1146/annurev-soc-121919-054621

Fatehkia, M., Kashyap, R., and Weber, I. (2018). Using Facebook ad data to track the global digital gender gap. *World Dev.* 107, 189–209. doi: 10.1016/j.worlddev.2018.03.007

Gaisbauer, F., Pournaki, A., Banisch, S., and Olbrich, E. (2021). Ideological differences in engagement in public debate on Twitter. *PloS ONE.* 16, e0249241. doi: 10.1371/journal.pone.0249241

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *J. Econom.* 225, 254–277. doi: 10.1016/j.jeconom.2021.03.014

Granovetter, M. S. (1973). The strength of weak ties. *Am. J. Sociol.* 78, 1360–1380. doi: 10.1086/225469

Green, D. (2015). *The Chukua Hatua Accountability Programme, Tanzania*. Oxford: Oxfam GB. Available online at: https://policy-practice.oxfam.org/resources/the-chukua-hatua-accountability-programme-tanzania-338436/ (accessed September 01, 2022).

Green, D. P., Groves, D., Manda, C., and Jones, R. (2018). *Mass Media Experiments to Reduce Violence Against Women in Tanzania*. Available online at: https://www.povertyactionlab.org/evaluation/mass-media-experiments-reduce-violence-against-women-tanzania (accessed September 01, 2022).

Green, D. P., Wilke, A. M., and Cooper, J. (2020). Countering violence against women by encouraging disclosure: a mass media experiment in rural Uganda. *Comp. Political Stud.* 53, 2283–2320. doi: 10.1177/0010414020912275

Green, K., Girault, P., Wambugu, S., Clement, N. F., and Adams, B. (2014). Reaching men who have sex with men through social media: a pilot intervention. *Digit. Cult. and Edu.* 6, 208–213.

Hofer, M., and Aubert, V. (2013). Perceived bridging and bonding social capital on Twitter: Differentiating between followers and followees. *Comput. Hum. Behav.* 29, 2134–2142. doi: 10.1016/j.chb.2013.04.038

Howison, J., Wiggins, A., and Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. *J. Assoc. Inf. Syst.* 12, 767–797. doi: 10.17705/1jais.00282

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, A. (2016). Combining satellite imagery and machine learning to predict poverty. *Science* 353, 790–794. doi: 10.1126/science.aaf7894

Jeong, D. (2021). *Creating (Digital) Labor Markets in Rural Tanzania. SSRN [Preprint]*. Available online at: http://dx.doi.org/10.2139/ssrn.4043833

Kadushin, C. (2002). The motivational foundation of social networks. *Soc. Netw.* 24, 77–91. doi: 10.1016/S0378-8733(01)00052-1

Kashyap, R., Fatehkia, M., Al Tamime, R., and Weber, I. (2020). Monitoring global digital gender inequality using the online populations of Facebook and Google. *Demogr. Res.* 43, 779–816. doi: 10.4054/DemRes.2020.43.27

Keast, R. (2022). "*Network Governance*," in *Handbook on Theories of Governance*, Ansell, C., and Torfing, J. (Eds.). Cheltenham, UK: Edward Elgar.

Keuschnigg, M., Lovsjö, N., and Hedström, P. (2017). Analytical sociology and computational social science. *J. Comp. Soc. Sci.* 1, 3–14. doi: 10.1007/s42001-017-0006-5

Khan, M. R. (2019). Educational Inequality and Mobile Phone Data. Case Study 3 in Data2X. *Big Data, Big Impact? Towards Gender-Sensitive Data Systems*. Available online at: https://data2x.org/resource-center/big-data-report/ (accessed September 01, 2022).

Koehler-Derrick, G. (2013). Quantifying anecdotes: google search data and political developments in Egypt. *PS Polit. Sci. Polit.* 46, 291–298. doi: 10.1017/S1049096513000267

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Qual. Quant.* 47, 2025–2047. doi: 10.1007/s11135-011-9640-9

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. *Science.* 343, 1203–1205. doi: 10.1126/science.1248506

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., et al. (2009). Computational social science. *Science.* 323, 721–723. doi: 10.1126/science.1167742

Lazer, D., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., et al. (2020). Computational social science: obstacles and opportunities. *Science.* 369, 1060–1062. doi: 10.1126/science.aaz8170

Lees, S., Marchant, M., Selestine, V., Mshana, G., Kapiga, S., and Harvey, S. (2021). The transformative effects of a participatory social empowerment intervention in the MAISHA intimate partner violence trial in Tanzania. *Cult. Health Sex.* 23, 1313–1328. doi: 10.1080/13691058.2020.1779347

Lin, N. (2001). *Social Capital: A Theory of Social Structure and Action*. New York, NY: Cambridge University Press. doi: 10.1017/CBO9780511815447

Marres, N. (2017). *Digital Sociology. The Reinvention of Social Research.* Cambridge, UK: Polity Press.

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., and Rajani, R. (2019). Inputs, incentives, and complementarities in education: experimental evidence from Tanzania. *Q. J. Econ.* 134, 1627–1673. doi: 10.1093/qje/qjz010

Okeleke, K. (2019). *Digital transformation in Tanzania: The role of mobile technology and impact on development goals*. GSM Association. Available online at: https://data.gsmaintelligence.com/api-web/v2/research-file-download?id=39256224&file=2736-180319-Tanzania.pdf (accessed September 01, 2022).

Pansardi, P., and Bindi,. M. (2021). The new concepts of power? Power-over, power-to and power-with. *J. Political Power*. 14, 51–71. doi: 10.1080/2158379X.2021.1877001

Park, P. S., Blumenstock, J. E., and Macy, M. W. (2018). The strength of long-range ties in population-scale social networks. *Science*. 362, 1410–1413. doi: 10.1126/science.aau9735

Pressgrove, G., McKeever, B. W., and Jang, S. M. (2018). What is contagious? Exploring why content goes viral on Twitter: a case study of the ALS Ice Bucket Challenge. *Int. J. Nonprofit Volunt. Sect. Mark.* 23, e1586. doi: 10.1002/nvsm.1586

Pretari, A., Towa, E., Ndiego, B., Wango, E., Mhina, M., Schwitter, N., et al. (2019). *Active Citizenship in Tanzania: Impact Evaluation of the 'Governance and Accountability Through Digitalization' Project*. Oxford: Oxfam GB. Available online at: https://policy-practice.oxfam.org/resources/active-citizenship-in-tanzania-impact-evaluation-of-the-governance-and-accounta-620855/ (accessed September 01, 2022).

Putnam, R. D. (2000). *Bowling Alone. The Collapse and Revival of American Community*. New York, NY: Simon & Schuster. doi: 10.1145/358916.361990

Rowlands, J. (2014). *Making the Impossible Possible: An overview of governance programming in fragile contexts*. Oxford: Oxfam GB. Available online at: https://policy-practice.oxfam.org/resources/making-the-impossible-possible-an-overview-of-governance-programming-in-fragile-331984/ (accessed September 01, 2022).

Ruths, D., and Pfeffer, J. (2014). Social Media for large studies of behavior. *Science* 346, 1063–1064. doi: 10.1126/science.346.6213.1063

Salganik, M. J. (2018). *Bit by Bit*. Princeton, New Jersey: Princeton University Press.

Smith, B. C. (2007). *Good Governance and Development*. New York: Palgrave. doi: 10.1007/978-1-137-06218-5

Smith, R. D., and Kishekya (2013). *Effectiveness Review: Chukua Hatua Tanzania*. Oxford: Oxfam GB. Available online at: https://policy-practice.oxfam.org/resources/effectiveness-review-chukua-hatua-tanzania-303755/ (accessed September 01, 2022).

Sousa, D., Sarmento, L., and Mendes Rodrigues, E. (2010). "Characterization of the twitter@ replies network: are user ties social or topical?" in *Proceedings of the 2nd international workshop on Search and mining user-generated contents* p. 63–70. doi: 10.1145/1871985.1871996

Stier, S., Breuer, J., Siegers, P., and Thorson, K. (2019). Integrating survey data and digital trace data: key issues in developing an emerging field. *Soc. Sci. Comp. Rev.* 38, 503516. doi: 10.1177/0894439319843669

Sun, L., and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *J. Econom.* 225, 175–199. doi: 10.1016/j.jeconom.2020.09.006

Townsend, L., and Wallace, C. (2016). *Social Media Research: A Guide to Ethics*. University of Aberdeen. Available online at: www.gla.ac.uk/media/media_487729_en.pdf

Verweij, P. (2012). Twitter links between politicians and journalists. *Journalism Pract.* 6, 680–691. doi: 10.1080/17512786.2012.667272

Viterna, J., and Robertson, C. (2015). New directions for the sociology of development. *Annu. Rev. Sociol.* 41, 243–269. doi: 10.1146/annurev-soc-071913-043426

Wellman, B. (2001). Physical place and cyberplace: the rise of personalized networking. *Int. J. Urban Reg. Res.* 25, 227–252 doi: 10.1111/1468-2427.00309

York, P., and Bamberger, M. (2020). *Measuring results and impact in the age of big data: The nexus of evaluation, analytics, and digital technology*. New York: The Rockefeller Foundation. Available online at: https://www.rockefellerfoundation.org/wp-content/uploads/Measuring-results-and-impact-in-the-age-of-big-data-by-York-and-Bamberger-March-2020.pdf (accessed September 01, 2022).

Zadeh, A., and Sharda, R. (2022). How can our tweets go viral? Point-process modelling of brand content. *Inf. Manag.* 59, 103594. doi: 10.1016/j.im.2022.103594

Zook, M., Barocas, S., boyd, d., Crawford, K., Keller, E., Gangadharan, S. P., et al. (2017). Ten simple rules for responsible big data research. *PLoS Comput. Biol.* 13, e1005399. doi: 10.1371/journal.pcbi.1005399

frontiers | Frontiers in Sociology

TYPE Original Research
PUBLISHED 29 November 2022
DOI 10.3389/fsoc.2022.907199

OPEN ACCESS

EDITED BY
Roger Berger,
Leipzig University, Germany

REVIEWED BY
Ruben Bach,
University of Mannheim, Germany
Jessica Daikeler,
GESIS Leibniz Institute for the Social Sciences, Germany

*CORRESPONDENCE
Juliane Kühn
juliane.kuehn@fau.de

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Sociological Theory, a section of the journal Frontiers in Sociology

RECEIVED 29 March 2022
ACCEPTED 20 October 2022
PUBLISHED 29 November 2022

CITATION
Eberl A, Kühn J and Wolbring T (2022) Using deepfakes for experiments in the social sciences - A pilot study. *Front. Sociol.* 7:907199. doi: 10.3389/fsoc.2022.907199

Using deepfakes for experiments in the social sciences - A pilot study

Andreas Eberl†, Juliane Kühn*† and Tobias Wolbring†

Chair of Empirical Economic Sociology, Friedrich-Alexander University Erlangen-Nürnberg, Nuremberg, Germany

The advent of deepfakes - the manipulation of audio records, images and videos based on deep learning techniques - has important implications for science and society. Current studies focus primarily on the detection and dangers of deepfakes. In contrast, less attention is paid to the potential of this technology for substantive research - particularly as an approach for controlled experimental manipulations in the social sciences. In this paper, we aim to fill this research gap and argue that deepfakes can be a valuable tool for conducting social science experiments. To demonstrate some of the potentials and pitfalls of deepfakes, we conducted a pilot study on the effects of physical attractiveness on student evaluations of teachers. To this end, we created a deepfake video varying the physical attractiveness of the instructor as compared to the original video and asked students to rate the presentation and instructor. First, our results show that social scientists without special knowledge in computer science can successfully create a credible deepfake within reasonable time. Student ratings of the quality of the two videos were comparable and students did not detect the deepfake. Second, we use deepfakes to examine a substantive research question: whether there are differences in the ratings of a physically more and a physically less attractive instructor. Our suggestive evidence points toward a beauty penalty. Thus, our study supports the idea that deepfakes can be used to introduce systematic variations into experiments while offering a high degree of experimental control. Finally, we discuss the feasibility of deepfakes as an experimental manipulation and the ethical challenges of using deepfakes in experiments.

KEYWORDS

deepfakes, face swap, deep learning, experiment, physical attractiveness, student evaluations of teachers

Introduction

Since the end of 2017, the creation and distribution of deepfakes have increased sharply. This phenomenon started on the platform *reddit* with a user called "deepfake" - a symbiosis between *deep learning* and *fakes* - who created the same name forum on this platform. By making the computer code available, other users could produce deepfakes themselves and contribute their results through the platform, leading to their immense popularity (Kietzmann et al., 2020). Besides the manipulation of audio records, deepfakes provide the ability to swap one person's face onto another in a picture or a video based on

artificial intelligence applying deep learning techniques. The specific algorithm, which creates these fake videos, learns and improves by constantly mimicking gestures and facial expressions (Maras and Alexandrou, 2019). While image editing packages have only enabled adding, replicating, or removing objects on images (Verdoliva, 2020), *video manipulations* become more realistic using artificial intelligence. The most common examples are videos that include celebrities or politicians whose faces have been swapped with those of other persons or individuals whose facial attributes or styles (e.g., hair) have been altered (Langguth et al., 2021). Moreover, deepfakes also include sophisticated *image manipulations* based on artificial intelligence. Besides the possibility to create images based on the semantic layout (Park et al., 2019), sketches (Isola et al., 2017), or text (Reed et al., 2016), it is also feasible to modify images, such as changing the color scheme (Zhu et al., 2017) or background (Isola et al., 2017), without affecting their realistic perception (see also Tolosana et al., 2020).

This high degree of realism of deepfakes and their indistinguishability from original videos and images for the inattentive human mind lead to the perception of deepfakes as a threat to human society, democracy, and public discourse as well as a potential driver of societal radicalization, polarization and conflict (Borges et al., 2019; Qayyum et al., 2019; Westerlund, 2019). Therefore, it is not surprising that buzzwords like manipulation, abuse, and political influence often appear in news reports and scientific pieces covering deepfakes. Examples could especially be seen in the U.S., where deepfakes were used to spread fake news (Westerlund, 2019). Having those examples from everyday life in mind, many argue that threats related to deepfakes outweigh their benefits (e.g., Fallis, 2020). While deepfakes carry the potential of disinformation and manipulative use, they cannot be dismissed exclusively as a threat, because differentiations exist concerning their ethical principles. As de Ruiter (2021) puts it: "deepfake technology and deepfakes are morally suspect, but not inherently morally wrong" (p. 1328). In her opinion, three factors condition the immoral use of deepfakes: representation of persons to which they would not consent, deliberate deception of viewers, and harmful intention. Considering these specific factors, a morally acceptable use of deepfakes is not entirely out of the question. Nevertheless, are deepfakes solely a threat to social cohesion or can they also help to advance social science knowledge?

Until now, most scientific papers dealing with deepfakes focus either on the extension of algorithms to improve the graphical results, solutions to detect those deepfakes, or their threats to society. However, less attention is paid to the potential of deepfakes for substantive research - especially as an approach for experimental manipulation with a high degree of control in the social sciences. In this paper, we aim to address this research gap and argue that deepfakes can be a valuable tool for conducting social science experiments. To demonstrate some of the potentials and pitfalls of deepfakes, we conducted a pilot study on the effects of physical attractiveness on students' evaluations of teaching. For this purpose, a deepfake video was created from two individuals with varying physical attractiveness. Students watched one of the two randomly assigned videos and rated the presentation, the instructor, and the video. Besides providing suggestive evidence on potential mechanisms of discrimination at work, we also conducted this experiment as an attempt to test the possibility of using the deepfake technology for experimental variation in sociological research. However, before we go into the details of this pilot study, we discuss previous research that has used deepfakes for answering social science research questions. While - to the best of our knowledge - only one study exists which uses deepfake videos in a similar way as our pilot (Haut et al., 2021), providing some background on existing research hopefully contributes to a better understanding of the potentials and pitfalls of the technique in the social sciences.

Previous studies using deepfakes

The amount of literature on deepfakes has increased sharply since 2017, and many of these papers warn primarily about their dangers (e.g., Fallis, 2020). Rather than just reporting on these threats to society and democracy, we will take a broader social science perspective in this paper. Therefore, we will also address the potential of deepfakes in scientific research. Accordingly, this section also covers studies that used deepfakes as a treatment in experiments or surveys to answer social science research questions. Please note that this is not a systematic review (for systematic reviews on deepfakes, see: Westerlund, 2019; Gamage et al., 2021; Godulla et al., 2021).

Due to the threat potential attributed to deepfakes, several studies deal with the computer-assisted *detection of deepfakes*, i.e., automated detection through machine learning (e.g., Zhang et al., 2017; Matern et al., 2019; Fagni et al., 2021; Mehta et al., 2021; Trinh and Liu, 2021). Other studies focus on human detection of deepfakes by conducting experiments to determine whether individuals can reliably detect deepfaked content (images and videos). The upshot of these studies is that individuals fail to detect deepfaked images. For example, Nightingale and Farid (2022) show that artificial intelligence (AI) synthesized faces are indistinguishable from real faces. Experiments using manipulated videos point in the same direction corroborating the claim that people cannot reliably detect deepfakes (Khodabakhsh et al., 2019; Köbis et al., 2021; Ternovski et al., 2021). Possible reasons for this insufficient detection rate are that deepfakes are sometimes perceived as more authentic than the original videos (Köbis et al., 2021) and that AI-synthesized images are perceived as more trustworthy than real faces (Nightingale and Farid, 2022).

Deepfakes can therefore be seen as a further step as compared to manipulations that only have a *human-like* appearance, such as robots or avatars. A distinction in this respect is made by de Ruiter (2021): "While real person deepfakes attribute digitally produced forms of speech and behavior to real individuals, avatar[s] […] attribute actual speech and behavior of real persons to digitally produced avatars" (p. 1316). Nevertheless, researchers claim that head-talking avatars also reduce confidence in AI-generated results, while uncanny valley expectations act as a mediator (Weisman and Peña, 2021). The term "uncanny valley" refers to the feeling of unease due to conflicting information resulting from visual impressions that are neither clearly artificial nor clearly human (Mori et al., 2012). In this ambiguous context, two options arise, either the avoidance of human likeness (so that robots are clearly recognized as such) or the perfectionism of human likeness (so that robots cannot be distinguished from humans) (Welker et al., 2020). For the latter, deepfakes seem to be a suitable means.

However, deepfakes not only help to overcome eerie feelings, but they also show *influence on (social) media and trust*. For example, Vaccari and Chadwick (2020) use an existing political deepfake video (Obama/Peele video) in their experiment to investigate whether deepfakes are recognized as such by individuals and how this affects respondents' trust in the media. The results show that political deepfakes do not deceive individuals because they realize that the person in the video would never have said anything like that. However, watching the deepfake video increases uncertainty, reducing general trust in social media and the news. This finding is supported by Ahmed (2021), who uses survey data and shows that skepticism toward the media is increasing due to deepfakes. Going one step further, Dobber et al. (2021) investigate in an online experiment how a political deepfake (manipulated video and audio) affects political attitudes. The results indicate that deepfakes could be used to stage a political scandal. While attitudes toward the depicted politician are significantly lower after watching the deepfake video, attitudes toward the politician's party are not affected. Additionally, the authors show that political microtargeting techniques can intensify the effects of a deepfake. More general, the results by Hughes et al. (2021) suggest that deepfake videos influence viewers' attitudes and intentions in the same way, as is true for original (not faked) videos.

A simple solution to buffer the harmful consequences of deepfakes could be to raise awareness for the existence of deepfakes. However, warning individuals of deepfakes can further decrease trust in information and the media in general. In this context, Ternovski et al. (2021) use online experiments to warn voters of the existence and dangers of deepfakes before watching selected political videos. After receiving a warning regarding deepfake videos, the results show that individuals begin to distrust all political video footage presented in the experiment, even the original (not faked) videos. Thus, their results illustrate that deepfakes pose a problem not simply through the spread of misinformation but also through the delegitimization of true information.

To the best of our knowledge, there is only one study that leverages deepfakes for examining discrimination. Haut et al. (2021) show an image of a black person vs. an image of a white person using the same audio record in their experiment. The authors measure credibility as the percent of participants who believed the speaker was telling the truth. The results reveal that changing a person's race in a static image has no impact on credibility. In a second step, Haut et al. (2021) test the effect of showing either an original video or a manipulated video where the person's appearance in the original video is manipulated to appear more "white." The original video shows a South Asian speaker, whereas the altered video shows a more "white" speaker. Unlike the presentation of an image, manipulation in a video significantly increases credibility.

To sum up, this literature review reveals that only a limited number of studies used deepfakes to investigate social science research questions beyond their effects on trust in media and politics. While previous research has mainly focused on the dangers of deepfakes or their detection by algorithms or humans, few studies address their potential, e.g., to study the discrimination of different groups of people like Haut et al. (2021). However, this lack of studies is surprising, as deepfakes have specific advantages for social science research. Deepfakes enable the systematic variation of visual and audio stimulus materials in experiments, while holding all else constant. In particular, the simultaneous manipulation of visual and acoustic materials represents an extension of previous techniques. For example, researchers can manipulate a person's face while keeping all other video elements like the audio record and its speed, background, clothing, and hairstyles identical. Influences outside the individual, which also affect their perception (Keres and Chartier, 2016), can be kept stable across experimental conditions, minimizing biases in estimates of physical attractiveness effects. Consequently, deepfakes offer a high degree of experimental control and thus appear to be a promising method to identify causal effects in experiments by systematically varying only one factor at a time.

Motivation and theoretical background of pilot study

In order to fill the research gap identified in the previous section, we conducted a pilot study to explore the feasibility of using deepfakes for social science research, especially experiments on discrimination. In this pilot, we build on previous research of one of the authors (Wolbring and Riordan, 2016) on the effects of instructors' physical attractiveness on students' evaluations of teaching (SET). The basic idea is that physical attractive instructors might profit from a *beauty premium* in the form of better SET scores (e.g.,

Hamermesh and Parker, 2005). Different theoretical mechanisms might cause this effect, including an attention boost to physical attractive instructors (e.g., Mulford et al., 1998), the ascription of positive stereotypes to good looking faculty (e.g., Dion et al., 1972) and the beauty glamor effect which can buffer the consequences of misconduct and bad performance to some degree (e.g., Bassili, 1981). However, this premium can also turn into a *beauty penalty* (e.g., Andreoni and Petrie, 2008) if positive stereotypes are disappointed by conflicting behavior or if the activated stereotypes do not match with the demands of the context (e.g., physically attractive female managers). Thereby, current research shows for various contexts that men consistently benefit from physical attractiveness, while the picture is more differentiated for women, who may profit or be disadvantaged (Hosoda et al., 2003; Paustian-Underdahl and Walker, 2016; Pajunen et al., 2021).

In the current literature on the effect of physical attractiveness on teaching evaluations, there are two opposing approaches. To some extent, our approach takes a middle ground combining strengths from both approaches. One group of studies relies on field data collected in real teaching contexts (Felton et al., 2008). The other group of studies uses experimental data collected in the context of laboratory experiments (Wolbring and Riordan, 2016). While in the first approach, based on observational data, the singular effect of physical attractiveness is hard to separate from other nuisances, in the second approach, based on experimental data, there is no real classroom situation. So far, image and audio material had to be separated from each other, as their simultaneous manipulation was not possible.

This is where the deepfake technology comes in, bringing exactly this advantage. By using deepfakes, it is possible both to vary the physical attractiveness in a targeted manner and to combine manipulated image material with an audio record, thus creating a realistic (online) teaching situation. Thereby, deepfakes can also account for the fact that some researchers assume that the evaluation of static and dynamic faces is based on different evaluation schemes (Riggio et al., 1991; Rubenstein, 2005). In order to achieve the most realistic assessment of teaching, we argue that videos should be given preference over images. Another advantage is a high degree of experimental control which helps to isolate the effect of physical attractiveness, since the audio records and background conditions of the original and the deepfake are identical, while other nuisances are addressed by means of randomization.

Guided by our theoretical framework, we created a deepfake based on two persons with varying physical attractiveness and conducted a small experiment among student subjects. In the experiment, we focus, on the one hand, on practical and methodological aspects such as the effort needed to manipulate the videos for social scientists without a strong background in computer science, the challenges we encountered when implementing the deepfakes, and the realism of the resulting videos according to participants of the study. On the other hand, we provide suggestive evidence on a substantive research question by exploring whether there are differences in the SET scores of a more and of a less physically attractive instructor. Given that the deepfakes allow us to control all other nuisances, finding such differences would point toward a beauty premium or beauty penalty. However, it is important to note that this is only suggestive evidence due to the small number of videos ($N = 2$) and subjects ($N = 37$) which also limits the possibilities to dig deeper into the underlying mechanisms at work.

Creation of the deepfakes

For the creation of the deepfake video, we used the software *deepfacelab*[1] which relies on the principle of an autoencoder, a special type of neural network. Thereby, an image first passes through an *encoder* that compresses the information provided, resulting in a low-dimensional representation of that input. On this basis the *decoder* tries to restore the original image (Perov et al., 2020). Using this technology, we can systematically vary the stimulus material shown in our experiment. As a starting point, we used the video of a person who is perceived as physically more attractive (original A) and the video of a person who is perceived as physically less attractive (original B) as the source materials. In the creation of the deepfake, we insert the face of the latter (B) into the video of the physically more attractive person (A) resulting in a deepfake.

In order to check whether the instructors actually differ in terms of their physical attractiveness, the pictures of those two individuals were evaluated in advance. So on the one hand, person A of the original video and on the other hand, person B whose face will be used for the creation of the deepfake. In order to avoid suspicion among the participants of the actual experiment on the deepfaking of videos, we asked 32 external reviewers from a snowball sample in our personal network for physical attractiveness ratings on a seven-point Likert scale from 1 = not at all physically attractive to 7 = very physically attractive. Each reviewer only rated one picture to avoid mutual influence or anchor effects. The physical attractiveness ratings of person A and B differ by almost two scale points (mean for person A: 4.93, mean for person B: 3.06).

In a next step, we recorded the videos and generated the deepfakes. To facilitate the creation of the deepfake, both videos

1 Here, we used *DeepFaceLab_DirectX12_build_11_20_2021.exe* which can be downloaded on https://github.com/iperov/DeepFaceLab/. Thereby we used the following procedure: 1) clear workspace; 2) extract images from video data_src; 3) extract images from video data_dst; 4) data_src faceset extract; 4.1) data_src view aligned result; 5) data_dst faceset extract; 5.1) data_dst view aligned results; 6) train Quick96; 7) merge Quick96. Advanced settings regarding the merging were determined via various test trials with different videos and people.

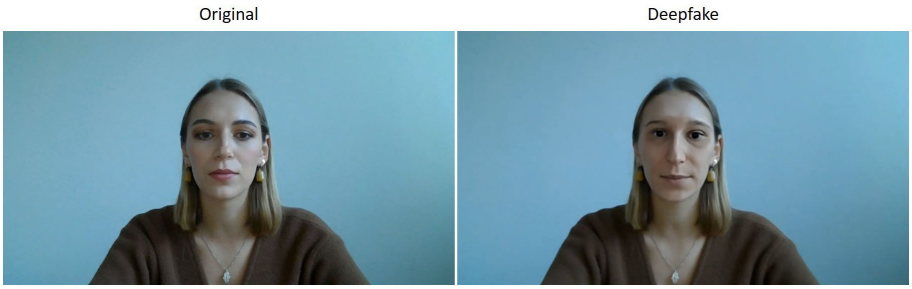


FIGURE 1
Experimental stimuli = **Original**: Physically more attractive instructor; **Deepfake**: physically less attractive instructor.

were shot under the same conditions (e.g., camera position, recording device, etc.) and with neutral background. After the recording, we switched to *deepfacelab* for extracting the images from both videos. We then started the training running three weeks[2] until we reached over 250,000 iterations. For the merging, we adjusted the corresponding settings (size of the mask, face size, color, etc.) so that the deepfake becomes as realistic as possible. After the complete merge, we visually checked that no artifacts were visible in the deepfake video[3]. Following this procedure, we carried out various tests in order to be able to select the best result and gain experience with the software. The decisive factors for choosing the final video were, on the one hand, that the deepfake appears as credible and convincing as possible, and no visual artifacts are recognizable. On the other hand, a second important criterion was that the people in the source material are rated as differently as possible concerning their physical attractiveness to secure sufficient variation in physical attractiveness. So, if there are differences in the ratings of the videos, we can likely attribute them to the different appearances of the individuals.

In order to ensure that the person depicted in the deepfake video was indeed physically less attractive than the person in the original video, we also asked 18 external reviewers to rate the physical attractiveness of the hypothetical person shown in the deepfake. This rating matches almost perfectly to the one of the physically less attractive person (mean for person in deepfake: 2.89 as compared to 3.06 for the real person B). With those results, we can ensure that the treatment group evaluates the physically less attractive person (deepfake), while the control group assesses the physically more attractive person (original). An image of the stimulus material used in the experiment is shown in Figure 1 (videos in German language are available upon request).

Experimental setting and questionnaire

The experiment was embedded in an online bachelor course at Friedrich-Alexander University Erlangen-Nürnberg with 39 students. All students received the same instruction, explaining that after watching a video of a hypothetical teaching situation, they would have to rate both the presentation and the instructor. By having only one person giving the instruction, interviewer effects were avoided[4]. A voucher worth 20 euros was raffled among the students who performed best in a test on the content of the presentation. In this way, we wanted to ensure that the students focus on the content of the video.

After one instruction for all respondents, the students were randomly assigned into one of two groups (after data cleaning: original $N = 19$, deepfake $N = 18$) and were not able to switch between groups. The treatment group watched the deepfake video, which contained the physically less attractive instructor, and the control group watched the original video with the physically more attractive instructor. Accordingly, each participant watched either the deepfake video *or* the original video. The 2-minute video was an introduction to the topic of social inequality based on Solga et al. (2009). Following the study by Wolbring and Riordan (2016), the students then received the corresponding SET questionnaire, which they filled out online and anonymously. At the end of the survey, respondents were

2 Please note that this time varies with the hardware used and the exact specification of the algorithm.

3 Visual artifacts are errors in deepfakes, such as brief moments in which the original face is recognizable or visible attributes in the deepfaked face that belong to the original face, such as eyebrows or earrings (Verdoliva, 2020). In this context, attentive viewing of the videos by several people has proven to be a suitable method for us.

4 In order to be able to respond to any questions or problems the students might encounter, there was an experimenter present in each group. These persons had the same name in both conditions and had a switched-off camera to avoid different visual stimuli. The option to ask questions was not used in either subgroup. Influences of the experimenter can therefore be excluded.

redirected to another page for the lottery to ensure that personal information and survey responses cannot be linked.

The questionnaire started with the *evaluation of the presentation,* including four items - the structure of the presentation, its argumentation chain, its speed, as well as its effect on students' interest in the topic - in random order. Afterwards, the *evaluation of the instructor* was based on nine items, which were also displayed in random order. The instructor's competence was evaluated with two items, followed by questions on her rhetoric and leadership qualities. In addition, the students were asked to assess the instructor's preparation, reliability, likeability, open-mindedness, and enthusiasm for the subject. The ratings of the presentation and the instructor are based on a Likert scale from 1 = does not apply at all to 7 = fully applies. To complete this evaluation, students assigned an *overall grade* for both the presentation and the instructor with all values including decimals between 1.00 = poor to 5.00 = excellent. Finally, five *knowledge questions* as well as *questions about the experiment* and *socio-demographics* formed the last part of the survey.

After data quality control and the subsequent deletion of two persons from the sample (final $N = 37$), our analyses concerning covariate balance suggest that the *randomization* was successful. Regarding interest in the topic, differences in average ratings are less than half a scale point (on a seven-point Likert scale: $t = 0.99$; $p = 0.34$). Similarly, prior knowledge of the topic differs between the treatment and the control group by only half a scale point ($t = 1.59$; $p = 0.12$). Given those small differences, we checked the robustness of the reported results by controlling for interest, prior knowledge, and the number of correct test answers in a linear regression model. Despite the small number of cases, we follow the request of a reviewer to report results from significance testing, but want to emphasize that due to the low statistical power of our study the results of significance testing should be treated with caution. In particular, conclusions about statistical significance should not be mixed with the strength of substantial relevance of an effect (Bernardi et al., 2017).

Results on the credibility of the deepfake video

In order to evaluate whether we were able to generate a credible deepfake for the experiment, the subsequent analyses in this section focus on three aspects. First, we asked the respondents to summarize the *study's aim* in their own words. On the one hand, part of the answers by the students referred to the content of the presentation, namely "social inequality." On the other hand, part of the students suspected that this study was about teaching evaluations. None of the answers addressed the video itself, nor did any comment suspect a possible manipulation of the instructor.

Second, we evaluated the *video quality*. In this context, we suspected that the deepfake may not be obvious to the students but that they may notice a deteriorating quality, for example, by perceiving the video as jerky or distorted. However, the results displayed in Figure 2 show that the video quality is rated comparably in both conditions. The average ratings of the video quality (1 = very poor to 7 = very good) hardly differ, with a difference of 0.1 (mean of original: 3.58; mean of deepfake: 3.67; $t = 0.23$; $p = 0.82$).

Finally, we analyzed the perceived *authenticity* following Haut et al. (2021). This question generated the largest differences among all inspected variables, although surprisingly not in the expected way: the deepfake video was rated more authentic than the original video differing by 1.28 scale points (1 = not authentic at all to 7 = very authentic). Accordingly, we find the same effect with regard to authenticity as Köbis et al. (2021) in their study. As Figure 3 shows, the average authenticity of the

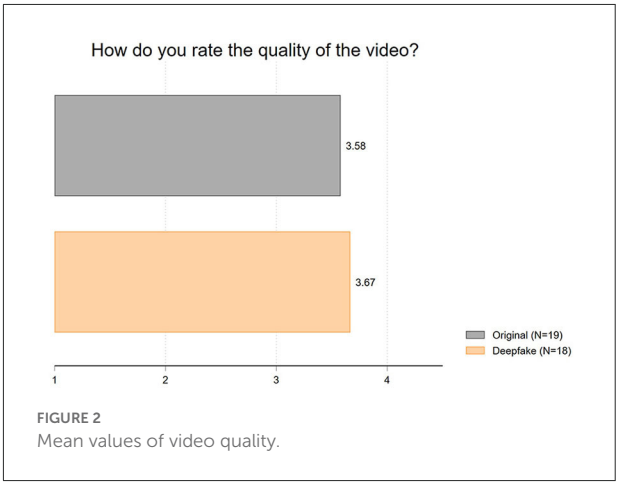


FIGURE 2
Mean values of video quality.

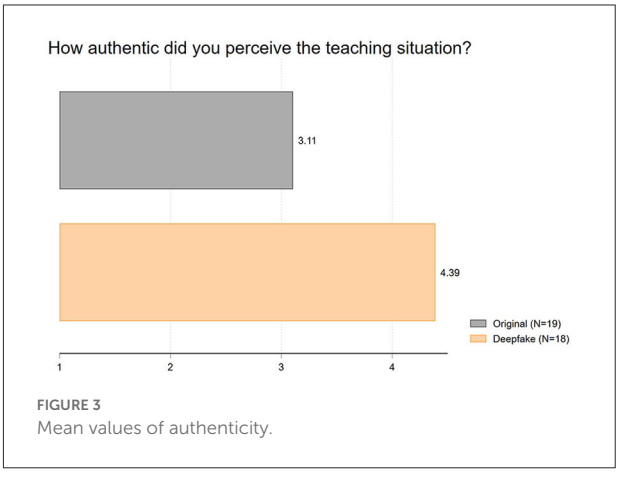


FIGURE 3
Mean values of authenticity.

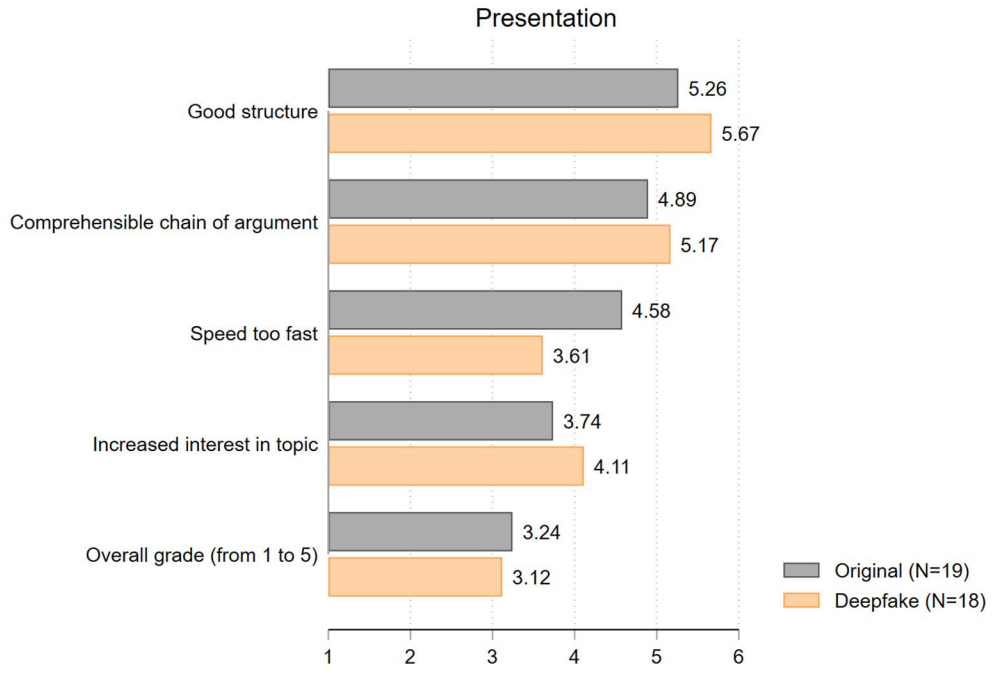


FIGURE 4
Evaluation of presentation (mean values).

original video was rated at 3.11, while the deepfake video was rated with a mean value of 4.39 ($t = 2.93$; $p = 0.01$).[5]

To sum up, we find no indications that the deepfake was detected by the participating students or that the deepfake video was perceived of lower quality than the original video. The deepfake also was not perceived as less authentic, although - as explained in footnote 5 - some concerns remain regarding the exact meaning of this authenticity measure in our study.

Suggestive evidence on the effects of attractiveness on SET

Having ruled out concerns about the potential detection of deepfakes, we can now explore whether differences in the evaluations between treatment and control group using deepfakes exist and whether they point toward a beauty premium or penalty. First, we present the results of students' *evaluation of the presentation*. Since we used the identical video in both experimental groups except for the persons' face, we expect the presentation ratings to be very similar. Our results largely confirm this expectation. As displayed in Figure 4, almost all presentation ratings of the deepfake and the original video differ by no more than half a scale point (*structure*: $t = 0.83$; $p = 0.41$; *comprehensible argumentation*: $t = 0.59$; $p = 0.56$; *impact on interest*: $t = 0.76$; $p = 0.45$). In line with this, the overall ratings of the presentation in grades (from 1.00 = poor to 5.00 = excellent) only slightly differ between the two experimental groups (3.24 vs. 3.12; $t = 0.49$; $p = 0.63$). The only exception concerns student's rating of the speed of the presentation. Surprisingly, students especially expressed that the speed of the original video is too fast compared to the deepfake (mean: 4.58 vs. 3.61; $t = 1.53$; $p = 0.13$). Even if this difference is not statistically significant, the difference is remarkably large given that the original and the deepfake video are based on exactly the same source, involving the identical audio record. We interpret this result as a first indication that the beauty premium does not show in our experiment, while there might be a beauty penalty at work.

Next, we focus on students' *evaluation of the instructor*, where we expect larger SET differences due to the face swap. The results are well in line with this suspicion (see Figure 5). Only the ratings for *likeability* (mean of original: 4.74, mean of deepfake: 5.06; $t = 0.66$; $p = 0.52$) and *good preparation* (5.05 vs. 5.44; t

5 Compared to Haut et al. (2021), we asked about the authenticity of the situation presented rather than the authenticity of the video. By adapting this question to our experiment in order to prevent suspicion about a potential video manipulation, we shifted the focus from the video to the situation and might have changed the meaning of the underlying measure. The observed lack of authenticity of the original video could be due to the fact that the person shown was perceived as too young to be a fully trained instructor, while the person in the deepfake looked older and might fit better in student's mental script of university instructors.

FIGURE 5
Evaluation of instructor (mean values).

= 0.95; $p = 0.35$) are similar in the two groups. All other items differ by almost one scale point or more. The largest differences exist in ratings on *general competence* (4.42 vs. 5.56; $t = 2.75$; $p = 0.01$) and *clear core statements* (3.95 vs. 5.33; $t = 2.41$, $p = 0.02$). Likewise, the perception of the *rhetorical skills* (3.00 vs. 3.94; $t = 1.66$; $p = 0.11$), *leadership qualities* (2.95 vs. 4.06; $t = 2.62$; $p = 0.01$), *open-mindedness* (3.37 vs. 4.61; $t = 2.14$; $p = 0.04$) and *reliability* (4.68 vs. 5.56; $t = 2.20$; $p = 0.03$) are influenced by the appearance of the instructor. Overall, both instructors are perceived to show only average *enthusiasm for their subject*, although here, again, the degree of enthusiasm of the deepfake instructor is rated 0.93 scale points better (2.74 vs. 3.67; $t = 1.69$; $p = 0.10$). Taken together, we see more positive instructor ratings for the deepfake than for the original video. The analysis of the overall instructor ratings in grades (from 1.00 = poor to 5.00 = excellent) points in the same direction (3.31 vs. 3.52) - even though the overall grade does not differ significantly ($t = 0.81$; $p = 0.42$). Therefore, there is no evidence for the existence of a beauty premium here either, but rather some suggestive evidence for a beauty penalty.

Discussion and implications

The nascent technology of deepfakes has important implications for the social sciences, both concerning its substantive research question such as misinformation and media trust and, as we contend, as a potential method for experimental manipulation. Our literature review shows that only a few social science studies address deepfakes while moving beyond the detection and dangers of this technology. In particular, deepfakes so far have been very rarely used as a tool for developing manipulations in experiments. We fill this research gap with our pilot study, and our findings suggest the feasibility of such an approach. Social scientists can successfully create a credible deepfake even without a corresponding education in computer science. Based on different test trials, we acquired appropriate knowledge in a reasonable amount of time that allowed us to create a deepfake using standard software and hardware. Notably, the quality of the deepfake was - in the eyes of the experimental subjects - comparable to our original video. None of the student subjects realized that they are watching a manipulated video.

Our study further underlines that deepfakes are suitable for researching social science issues in general and discrimination in particular. Because deepfakes maximize the videos' comparability (especially by having an identical audio record and the same conditions such as hairstyles, background, clothing, etc.), differences in ratings can be causally attributed to the varied stimulus. Using the case of physical attractiveness, our study supports the idea that deepfakes can be used to introduce systematic variations into experiments, while offering a high degree of experimental control. As a result, there are only small differences in students' evaluation of the presentation in

the two videos, but larger differences in the evaluations of the two instructors. By holding all other factors constant, we can attribute these differences to the appearance of the instructors. However, in contrast to previous studies (e.g., Wolbring and Riordan, 2016), our suggestive evidence points to adverse effects of physical attractiveness. The physically less attractive instructor is rated better than the physically more attractive instructor. One possible way to reconcile this finding with previous results is that the physically more attractive instructor in the original video was rather young. Students might thus have perceived this physically attractive and young instructor as less authentic and competent than the physically less attractive but older instructor (see text footnote 5). However, as this study is based on a small sample, one should not overinterpret the fact that this suggestive evidence from the pilot study pointing toward a beauty penalty conflicts with existing large-scale studies documenting a beauty premium. A replication of our pilot with more subjects and videos is needed. While the homogeneity of our student sample is an advantage for a first test with a small sample, sampling from a broader student population with more diverse backgrounds and majors is worth considering. Such a more heterogeneous and representative sample would help to address concerns about sample selection and to answer questions about the generalizability of our results.

Besides these methodological and substantive insights from this pilot study, the use of deepfakes in social science studies raises more general practical and ethical challenges, concerns, and tensions associated with the application of this technology. Subjects are - by definition - deceived when deepfakes are used in studies *without* actively communicating their use. In the social and behavioral sciences, there are conflicting views on the appropriateness of deception for research purposes, ranging from complete rejection of deception on one side to reinforcement of the benefits associated with deception on the other side (Barrera and Simpson, 2012). Thus, deepfake technology appears to be morally problematic at first glance because it violates social norms like truthfulness and risks undermining people's autonomy.

However, although deepfakes may appear morally suspect, the technology is not inherently morally wrong and, as we contend, there are ways to use deepfakes in empirical research in responsible ways. According to de Ruiter (2021), three factors are important to determine whether deepfakes are morally problematic: (i) would the faked person complain about how she/he is portrayed; (ii) does the deepfake deceive the viewers; (iii) what is the intention with which the deepfake was created. In our study, the faked person was aware of the purpose of the study when videotaping the presentation. The intention of the deepfake was to investigate discrimination and did not cause any harm whatsoever. Finally, one might argue that the viewers were deceived in our study, but we explicitly informed our subjects that they are watching the video of a *hypothetical* teaching situation. Moreover, we decided not to inform subjects after the experiment because, as our literature review has shown, such an active communication that deepfakes are used can harm people's general trust in the media and politics.

Additionally, it can be argued that deepfakes are real enough to avoid a sense of eeriness, which other studies with robots or avatars have shown (de Borst and de Gelder, 2015; Konijn and Hoorn, 2020). On the one hand, the deepfake technology offers great advantages concerning the authenticity of used video materials, whereby this is accompanied by a pleasant feeling when viewing them - in comparison to the problematic feeling of eeriness watching a human-like robot or avatar. On the other hand, the use of this technology evokes the often-discussed danger that the difference between the original and the deepfake is no longer perceptible. In this context, a clear distinction is needed: (a) when are deepfakes used to manipulate and deceive people in order to create harm, so that the lack of distinguishability is also morally reprehensible. And (b) when are deepfakes used as a scientific instrument in order to create optimal experimental conditions. In the latter case, there is an opportunity to make the most of this development. The accompanying lack of distinctiveness creates the conditions for investigating the different treatment of *real* persons based on their appearance and, if applicable, the underlying discrimination mechanism.

While we believe that this approach has circumvented the major concerns when using deepfakes, other studies might warrant other avenues to address these issues. Therefore, more research is not only needed to further explore the possibilities of deepfakes for answering substantive research questions in the social sciences, but also to address the associated ethical challenges of using deepfakes in scientific experiments.

Data availability statement

The data and code for this study are publicly and permanently available at the GESIS Datorium (Eberl et al., 2022).

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Acknowledgments

We would like to thank our student assistants Klara Frankenberger and Selin Atmaca for their support in the creation of the deepfake videos as well as their support throughout the project. AE gratefully acknowledges financial support from the Emerging Talents Initiative (ETI) from the University of Erlangen-Nürnberg.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ahmed, S. (2021). Navigating the maze: deepfakes, cognitive ability, and social media news skepticism. *New Media Soc.* 1–22. doi: 10.1177/14614448211019198

Andreoni, J., and Petrie, R. (2008). Beauty, gender and stereotypes: evidence from laboratory experiments. *J. Econ. Psychol.* 29, 73–93. doi: 10.1016/j.joep.2007.07.008

Barrera, D., and Simpson, B. (2012). Much ado about deception: consequences of deceiving research participants in the social sciences. *Sociol. Methods Res.* 41, 383–413. doi: 10.1177/0049124112452526

Bassili, J. N. (1981). The attractiveness stereotype: goodness or glamour? *Basic Appl. Soc. Psych.* 2, 235–252. doi: 10.1207/s15324834basp0204_1

Bernardi, F., Chakhaia, L., and Leopold, L. (2017). "Sing me a song with social significance": the (mis)use of statistical significance testing in European sociological research, European sociological. *Review.* 33, 1–15. doi: 10.1093/esr/jcx044

Borges, L., Martins, B., and Calado, P. (2019). Combining similarity features and deep representation learning for stance detection in the context of checking fake news. *J. Data Inf. Qual.* 11, 1–26. doi: 10.1145/3287763

de Borst, A. W., and de Gelder, B. (2015). Is it the real deal? Perception of virtual characters versus humans: an affective cognitive neuroscience perspective. *Front. Psychol.* 6, 1–12. doi: 10.3389/fpsyg.2015.00576

de Ruiter, A. (2021). The distinct wrong of deepfakes. *Philos. Technol.* 34, 1–22. doi: 10.1007/s13347-021-00459-2

Dion, K., Berscheid, E., and Walster, E. (1972). What is beautiful is good. *J. Pers. Soc. Psychol.* 24, 285–290. doi: 10.1037/h0033731

Dobber, T., Metoui, N., Trilling, D., Helberger, N., and de Vreese, C. (2021). Do (microtargeted) deepfakes have real effects on political attitudes? *Int. J. Press/Politics.* 26, 69–91. doi: 10.1177/1940161220944364

Eberl, A., Kühn, J., and Wolbring, T. (2022). Data & code: using deepfakes for experiments in the social sciences - a pilot study. *GESIS datorium.* doi: 10.7802/2467

Fagni, T., Falchi, F., Gambini, M., Martella, A., and Tesconi, M. (2021). TweepFake: about detecting deepfake tweets. *PLoS ONE.* 16, e0251415. doi: 10.1371/journal.pone.0251415

Fallis, D. (2020). The epistemic threat of deepfakes. *Philos. Technol.* 34, 1–21. doi: 10.1007/s13347-020-00419-2

Felton, J., Koper, P. T., Mitchell, J., and Stinson, M. (2008). Attractiveness, easiness and other issues: Student evaluations of professors on ratemyprofessors.com. *Assess. Eval. High. Educ.* 33, 45–61. doi: 10.1080/02602930601122803

Gamage, D., Chen, J., and Sasahara, K. (2021). The emergence of deepfakes and its societal implications: a systematic review. *TTO.* 2021, 28—39. Available online at: https://truthandtrustonline.com/wp-content/uploads/2021/11/TTO_2021_proceedings.pdf

Godulla, A., Hoffmann, C. P., and Seibert, D. (2021). Dealing with deepfakes - an interdisciplinary examination of the state of research and implications for communication studies. *SCM Studies in Communication and Media.* 10, 72–96. doi: 10.5771/2192-4007-2021-1-72

Hamermesh, D. S., and Parker, A. (2005). Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity. *Econ. Educ. Rev.* 24, 369–376. doi: 10.1016/j.econedurev.2004.07.013

Haut, K., Wohn, C., Antony, V., Goldfarb, A., Welsh, M., Sumanthiran, D., et al. (2021). Could you become more credible by being white? Assessing impact of race on credibility with deepfakes. *arXiv preprint arXiv:2102.08054.* 1–10. doi: 10.48550/arXiv.2102.08054

Hosoda, M., Stone-Romero, E. F., and Coats, G. (2003). The effects of physical attractiveness on job-related outcomes: a meta-analysis of experimental studies. *Pers. Psychol.* 56, 431–462. doi: 10.1111/j.1744-6570.2003.tb00157.x

Hughes, S., Fried, O., Ferguson, M., Hughes, C., Hughes, R., Yao, X., et al. (2021). Deepfaked online content is highly effective in manipulating people's attitudes and intentions. *PsyArXiv.* 1–6. doi: 10.31234/osf.io/4ms5a

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). "Image-to-image translation with conditional adversarial networks." in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 1125—1134. doi: 10.1109/CVPR.2017.632

Keres, A., and Chartier, C. R. (2016). The biasing effects of visual background on perceived facial trustworthiness. *Psi Chi J. Psychol. Res.* 21, 170–175. doi: 10.24839/2164-8204.JN21.3.170

Khodabakhsh, A., Ramachandra, R., and Busch, C. (2019). "Subjective evaluation of media consumer vulnerability to fake audiovisual content" in: *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX): IEEE.* 1–6. doi: 10.1109/QoMEX.2019.8743316

Kietzmann, J., Lee, L. W., McCarthy, I. P., and Kietzmann, T. C. (2020). Deepfakes: trick or treat? *Bus. Horiz.* 63, 135–146. doi: 10.1016/j.bushor.2019.11.006

Köbis, N. C., DoleŽalová, B., and Soraperra, I. (2021). Fooled twice: people cannot detect deepfakes but think they can. *Iscience.* 24, 1–18. doi: 10.1016/j.isci.2021.103364

Konijn, E. A., and Hoorn, J. F. (2020). Differential facial articulacy in robots and humans elicit different levels of responsiveness, empathy, and projected feelings. *Robotics.* 9, 1–17. doi: 10.3390/robotics9040092

Langguth, J., Pogorelov, K., Brenner, S., Filkukov,á, P., and Schroeder, D. T. (2021). Don't trust your eyes: image manipulation in the age of deepfakes. *Front. Commun.* 6, 632317. doi: 10.3389/fcomm.2021.632317

Maras, M.-H., and Alexandrou, A. (2019). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos. *Int. J. Evid. Proof.* 23, 255–262. doi: 10.1177/1365712718807226

Matern, F., Riess, C., and Stamminger, M. (2019). "Exploiting visual artifacts to expose deepfakes and face manipulations" in: *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*: *IEEE*. 83–92. doi: 10.1109/WACVW.2019.00020

Mehta, V., Gupta, P., Subramanian, R., and Dhall, A. (2021). "Fakebuster: a deepfakes detection tool for video conferencing scenarios" in: *26th International Conference on Intelligent User Interfaces-Companion*. 61–63. doi: 10.1145/3397482.3450726

Mori, M., MacDorman, K. F., and Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robot. Autom. Mag.* 19, 98–100. doi: 10.1109/MRA.2012.2192811

Mulford, M., Orbell, J., Shatto, C., and Stockard, J. (1998). Physical attractiveness, opportunity, and success in everyday exchange. *Am. J. Sociol.* 103, 1565–1592. doi: 10.1086/231401

Nightingale, S. J., and Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proc. Natl. Acad. Sci. U.S.A.* 119, 1–3. doi: 10.1073/pnas.2120481119

Pajunen, T., Kukkonen, I., Sarpila, O., and Åberg, E. (2021). Systematic review of differences in socioeconomic outcomes of attractiveness between men and women. *arXiv*. doi: 10.31235/osf.io/rmcqh

Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. (2019). "Semantic image synthesis with spatially-adaptive normalization." in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Long Beach, CA), 2337–2346. doi: 10.1109/CVPR.2019.00244

Paustian-Underdahl, S. C., and Walker, L. S. (2016). Revisiting the beauty is beastly effect: examining when and why sex and attractiveness impact hiring judgments. *Int. J. Hum. Resour. Manag.* 27, 1034–1058. doi: 10.1080/09585192.2015.1053963

Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., et al. (2020). DeepFaceLab: integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535*. 1–10. doi: 10.48550/arXiv.2005.05535

Qayyum, A., Qadir, J., Janjua, M. U., and Sher, F. (2019). Using blockchain to rein in the new post-truth world and check the spread of fake news. *IT Prof.* 21, 16–24. doi: 10.1109/MITP.2019.2910503

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. (2016). "Generative adversarial text to image synthesis." in: *Proceedings of the 33rd International Conference on Machine Learning*. 1060–1069. doi: 10.48550/arXiv.1605.05396

Riggio, R. E., Widaman, K. F., Tucker, J. S., and Salinas, C. (1991). Beauty is more than skin deep: components of attractiveness. *Basic Appl. Soc. Psych.* 12, 423–439. doi: 10.1207/s15324834basp1204_4

Rubenstein, A. J. (2005). Variation in perceived attractiveness: differences between dynamic and static faces. *Psychol. Sci.* 16, 759–762. doi: 10.1111/j.1467-9280.2005.01610.x

Solga, H., Powell, J., and Berger, P. A. (2009). *Soziale Ungleichheit. Klassische Texte zur Sozialstrukturanalyse*. Frankfurt/New York: Campus Verlag.

Ternovski, J., Kalla, J., and Aronow, P. M. (2021). Deepfake warnings for political videos increase disbelief but do not improve discernment: evidence from two experiments. *OSF Preprint*. 1—12. doi: 10.31219/osf.io/dta97

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., and Ortega-Garcia, J. (2020). Deepfakes and beyond: a survey of face manipulation and fake detection. *Inf. Fusion*. 64, 131–148. doi: 10.1016/j.inffus.2020.06.014

Trinh, L., and Liu, Y. (2021). An examination of fairness of ai models for deepfake detection. *arXiv preprint arXiv:2105.00558*. 1–9. doi: 10.24963/ijcai.2021/79

Vaccari, C., and Chadwick, A. (2020). Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Soc. Media Soc.* 6, 1–13. doi: 10.1177/2056305120903408

Verdoliva, L. (2020). Media forensics and deepfakes: an overview. *IEEE J. Sel. Top. Signal Process.* 14, 910–932. doi: 10.1109/JSTSP.2020.3002101

Weisman, W. D., and Peña, J. F. (2021). Face the uncanny: the effects of doppelganger talking head avatars on affect-based trust toward artificial intelligence technology are mediated by uncanny valley perceptions. *Cyberpsychol. Behav. Soc. Netw.* 24, 182–187. doi: 10.1089/cyber.2020.0175

Welker, C., France, D., Henty, A., and Wheatley, T. (2020). Trading faces: complete AI face doubles avoid the uncanny valley. *Preprint from PsyArXiv*. 1–11. doi: 10.31234/osf.io/pykjr

Westerlund, M. (2019). The emergence of deepfake technology: a review. *Technol. Innov. Manag. Rev.* 9, 39–52. doi: 10.22215/timreview/1282

Wolbring, T., and Riordan, P. (2016). How beauty works. Theoretical mechanisms and two empirical applications on students' evaluation of teaching. *Soc. Sci. Res.* 57, 253–272. doi: 10.1016/j.ssresearch.2015.12.009

Zhang, Y., Zheng, L., and Thing, V. L. (2017). "Automated face swapping and its detection" in: *2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)*: *IEEE*. 15–19. doi: 10.1109/SIPROCESS.2017.8124497

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). "Unpaired image-to-image translation using cycle-consistent adversarial networks." in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 2223–2232. doi: 10.1109/ICCV.2017.244

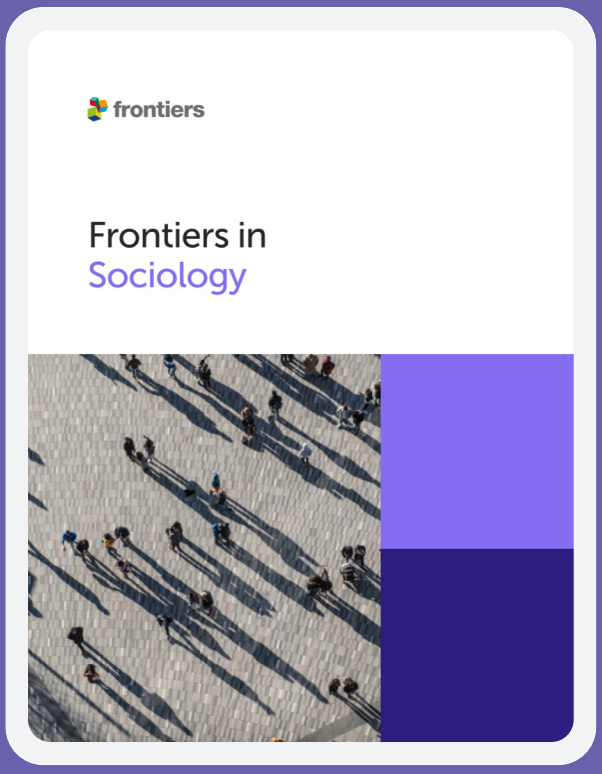
Frontiers in Sociology

Highlights and explores the key challenges of human societies

A multidisciplinary journal which focuses on contemporary social problems with a historical purview to understand the functioning and development of societies.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact