frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 13 March 2024
DOI 10.3389/fcomp.2024.1294985



OPEN ACCESS

EDITED BY
Paolino Di Felice,
University of L'Aquila, Italy

REVIEWED BY
Abdallah Namoun,
Islamic University of Madinah, Saudi Arabia
Patricia Martin-Rodilla,
University of A Coruña, Spain

*CORRESPONDENCE
Kyrill Meyer
✉ kyrill.meyer@hs-mittweida.de

RECEIVED 15 September 2023
ACCEPTED 21 February 2024
PUBLISHED 13 March 2024

CITATION
Memon AB, Sootahar DK, Luhana KK and Meyer K (2024) A corpus-based real-time text classification and tagging approach for social data. *Front. Comput. Sci.* 6:1294985. doi: 10.3389/fcomp.2024.1294985



# A corpus-based real-time text classification and tagging approach for social data

Atia Bano Memon[1,2], Dileep Kumar Sootahar[2], Kirshan Kumar Luhana[2] and Kyrill Meyer[3,4]*

[1]Department of Computer Science, University of Leipzig, Leipzig, Germany, [2]Department of Computer Science, University of Sindh, Jamshore, Sindh, Pakistan, [3]Mittweida University of Applied Sciences, Mittweida, Germany, [4]Institute for Digital Technologies, Leipzig, Germany

With the rapid accumulation of large amounts of user-generated content through social media, social data reuse and integration have gained increasing attention recently. This has made it almost obsolete for software applications to collect, store, and work with their own data stored on local servers. While, with the provision of Application Programming Interfaces from the leading social networking sites, data acquisition and integration has become possible, the meaningful usage of such unstructured, non-uniform, and incoherent data collections needs special procedures of data summarization, understanding, and visualization. One particular aspect in this regard that needs special attention is the procedures for data (text snippets in the form of social media posts) categorization and concept tagging to filter out the relevant and most suitable data for the particular audience and for the particular purpose. In this regard, we propose a corpus-based approach for searching and successively categorizing and tagging the social data with relevant concepts in real time. The proposed approach is capable of addressing the semantical and morphological similarities, as well as domain-specific vocabularies of query strings and tagged concepts. We demonstrate the feasibility and application of our proposed approach in a web-based tool that allows searching Facebook posts and provides search results together with a concept map for further navigation, filtering, and refining of search results. The tool has been evaluated by performing multiple search queries, and resultant concept maps and annotated texts are analyzed in terms of their precision. The approach is thereby found effective in achieving its stated goal of classifying text snippets in real time.

KEYWORDS

**social data, text classification, document tagging, text matching, social networking sites, social data integration, social data analysis**

## 1 Introduction

Nowadays, with the rapid introduction and advancement of social media platforms on one hand and the growing attention of the user community toward digitalization and social networking on the other hand, the world has transformed into a global village whereby everyone has a sort of digital presence and is actively participating in social networking activities (Bullinger et al., 2010). The services offered by social media platforms have expanded so well that they have provided their users with extensive possibilities for exchanging information in various fields such as education, health, culture, entertainment, sports, and other domains of knowledge (Mislove et al., 2007; Ghosh et al., 2012). As a result, with the passage of time, a huge amount of user-generated data through social

networking (i.e., social data) is being accumulated on social media servers (Ingole et al., 2018). Social data is defined as large amounts of multi-structured user-generated data being collected at social network data servers as a result of user interactions within social networking and online community sites (Memon et al., 2021). This implies that every time a user tweets on Twitter, posts an update on Facebook or LinkedIn, records Check-ins on Foursquare, and shares videos on YouTube, photos on Instagram, and pins on Pinterest, either way, they are voluntarily or involuntarily generating a significant amount of social (meta) data. Research advocates that this social data, as being collected from user communities, have much potential to support third-party applications for specific audiences and purposes (Memon et al., 2017). Thus, social platform users can be regarded as prosumers who produce and also consume user-generated data simultaneously. This is particularly important as social data can easily be categorized as its sibling concept of big data (George et al., 2014). Big data is categorized by its five properties, also referred to as the 5Vs of big data. This includes variety (different forms of data), volume (huge amount of data), velocity (rapid speech of data accumulation), veracity (structured data), and value (importance of data). Social data (aka community data) also encompasses these features and is thus particularly referred to as social big data or big social data (George et al., 2014), constituting about 95% of big data sources available in this decade (Gandomi and Haider, 2015).

Formally, social big data is defined as "any high-volume, high-velocity, high-variety and/or highly semantic data that is generated from technology-mediated social interactions and actions in the digital realm, and which can be collected and analyzed to model social interactions and behavior" (Olshannikova et al., 2017). This social big data offers numerous opportunities for data researchers and analysts to re-use the already available data in different ways and generate various relevant insights for specific purposes (Chen H. et al., 2012; Memon and Meyer, 2015; Ali et al., 2016; Tian, 2020). For example, students' casual talks on these media can be helpful in understanding their experiences, opinions, feelings, and emotions about the learning and educational institutes (Ingole et al., 2018). Similarly, users' reviews and informal discussions over these media regarding a particular product or service can be helpful in understanding the feedback and determining the acceptance or needs for improvement for the business organization. This provides a medium of direct conversation between customers and businesses and assists the businesses in the course of various functionalities encompassing business intelligence such as personalized recommendation systems, opinion analysis, expertise retrieval, and computational advertising (Chen H. et al., 2012; Abu-Salih et al., 2021). Many studies have suggested the effectiveness and benefits of leveraging social big data in supporting various information retrieval and analysis tasks for predictive analysis, such as for political trending and forecasting or stock market prediction (Nguyen et al., 2015), opinion mining or sentimental analysis during any epidemic analysis (Culotta, 2010), disaster relief (Beigi et al., 2016) or crime identification (Gerber, 2014), business intelligence as in targeted advertisement and marketing strategies (Maurer and Wiegmann, 2011), social network analysis for understanding the community and its features (Tabassum et al., 2018), and topic modeling and data analytics as in human–machine interactions (Bello-Orgaz et al., 2016). According to Vongkusolkit and Huang (2021), utilizing social big data for such applications as opposed to the data collected through traditional media of communication is more favorable as (1) it offers convenience in assessing the information and expressed emotions, (2) it allows the insights to the situation from different perspectives and locations, (3) is collected through participation in interactive, decentralized, uncontrolled, and large-scale discussions, and (4) is a cheap and efficient way to receive real-time updates in an unofficial and irregular manner.

While, with the provision of Application Programming Interfaces from the leading social networking sites, the acquisition, integration, and re-use of user data have become possible (Felt and Evans, 2008; Memon and Meyer, 2020), the meaningful usage of such unstructured, non-uniform, and incoherent yet semantically rich data collections need special procedures of data integration and summarization, data analysis and understanding, knowledge representation, and data visualization (George et al., 2014; Memon et al., 2017). It is believed that social big data analysis is an interdisciplinary field spanning various related fields. It draws upon the intersection of three major areas: social media as a natural source of data and data analysis, big data as a parallel and massive processing paradigm, and data analysis algorithms and methods for extracting and analyzing the data (Bello-Orgaz et al., 2016; Olshannikova et al., 2017). One particular aspect in this regard that needs special attention is the procedures for data (text snippets in the form of social media posts) categorization and concept tagging to filter out the relevant and most suitable data for the particular audience and for the particular purpose. Given such large heterogeneous datasets, manual processing, organizing, and analysis tend to be very difficult, time-consuming, less accurate, and much more expensive. Hence, drawing valuable and relevant knowledge from such data is a challenging task (Chen H. et al., 2012), which has not yet been completely solved by the research community.

In this regard, this study proposes a corpus-based approach for searching and successively categorizing and tagging the social data with relevant concepts in real time. The proposed approach is capable of addressing the semantic and morphological similarities and domain-specific vocabularies of query strings and tagged concepts. As a case-based approach, the proposed approach is implemented in a web-based tool that allows searching Facebook posts and provides search results together with a concept map for further navigation, filtering, and refining of search results. The tool has been evaluated by performing multiple search queries, and resultant concept maps and annotated text segments are analyzed in terms of precision by the researchers adopting an expert evaluation method. The approach is thereby found effective in achieving its stated goal of classifying text snippets in real time. In conclusion, we contribute a novel approach to social data categorization that can be integrated into various data-intensive applications that leverage textual social data and are adapted to different real-life scenarios.

The remainder of the article is structured as follows. Section 2 offers background information regarding different ways of text classification and thereby describes the problem statement and research objectives of the present article. The research approach of the current article is described in section 3. The proposed

approach is presented in section 4, section 5 presents the case implementation, and section 6 presents the evaluation results of the proposed approach. The article concludes in section 7 with an outlook on future research agenda.

## 2 Background

Text classification and concept tagging have been the focus of researchers for a long time (e.g., Lai et al., 2015; Kowsari et al., 2017, 2018) and have been studied in many real-life applications such as news filtering and organization, information retrieval and management, opinion mining, and email classification and spam filtering, to name a few (Aggarwal and Zhai, 2012; Kowsari et al., 2019). Research on the analysis of structured data has somehow matured recently; however, the analysis of huge amounts of unstructured data, which prevailed in recent years, still brings challenges for researchers and application developers. Such data contain unrelated and diverse information, thereby offering no outline to identify any defined patterns or structural manifestations (Chugani et al., 2021). Text classification, a task of natural language processing, refers to assigning categories to the textual data to facilitate its effective search, management, and navigation (Lashkarashvili and Tsintsadze, 2022). Text classification often assigns multiple categories to individual pieces of text, which might or might not be mutually exclusive. The essence of text classification is the process of adding tags or annotations to unstructured data for further analysis. Usually, text tagging can be achieved manually or automatically. However, manually tagging unstructured data can be a difficult, ineffective, time-consuming, and expensive task (Ingole et al., 2018; Kowsari et al., 2018). Therefore, the researchers have made substantial efforts to devise approaches for the automatic classification of unstructured data (Kowsari et al., 2019). This is particularly relevant whenever comprehensive human coding is not feasible due to the amount of data or because immediate classification information is required. Related work in the information retrieval domain has focused on search engine fundamentals such as indexing and improved feedback and query reformulation (Schütze et al., 2008). At the same time, recent work in this domain has focused on data mining and machine learning paradigms (Kowsari et al., 2015). The existing approaches can be categorized into three broad categories: rule-based approaches, machine learning-based approaches, and hybrid approaches.

### 2.1 Rule-based text classification approaches

Rule-based text classification approaches consist of manually created rules that define the classification scheme. The rules are often described in terms of taxonomy (as a thesaurus, dictionary, or ontology) or folksonomy. The taxonomies are usually predefined, hierarchical in nature with a top-down approach, and domain-specific, while the folksonomy-based approaches are non-hierarchical in nature with a bottom-up approach that allows the users to tag their data with relevant terms that they find appropriate according to their content (Al-Khalifa and Davis, 2007). When taxonomies exist, which essentially offer a list of possible classes of data, there are various approaches suggested to determine the particular class of individual data (Wang et al., 2007, 2012; Sanchez-Pi et al., 2016; Li et al., 2017). The major problem with taxonomy-based techniques is that they might offer good accuracy, but creating those taxonomies takes deep understanding and a lot of time. They require an authoritative source for their creation as well as validation. These methods also tend to be less scalable as new data might need new rules, and they are difficult to code again with existing rules. Thus, with time, they tend to produce more false-positives and false-negatives (Kaur et al., 2021). In comparison, folksonomies are open-ended and agree more with human understanding (Bai et al., 2009; Cantador et al., 2011; Godoy and Corbellini, 2016). Folksonomies offer benefits of ease of implementation and reflection of intended user vocabulary; however, the problem with folksonomies is the uncontrolled vocabulary leading to ambiguity due to the use of synonyms and noise tags by the users (Cai et al., 2016). Alternatively, rule-based approaches include the techniques to extract relevant keywords from the text itself through various statistical techniques such as Bag of Words, TF-IDF, N-grams, etc., as the textual features of the data (Chen Y. et al., 2012; Nobata et al., 2016). These approaches rely on word frequencies, and the keywords extracted therein are not always relevant and are not necessarily based on the concepts present in the text. An interesting approach in this regard was presented by Sahlgren and Coaster (2004), where they computed concepts based on word co-occurrence data, which they claimed improved the performance of their text classification task.

### 2.2 Machine learning-based text classification approaches

Recent research in the area of text classification has suggested various machine-learning approaches that offer relatively higher accuracy and efficiency for large unstructured datasets. Machine language approaches in general involve building classifiers from different features, which are then trained for automatic classification from instances of already labeled data. The machine learning approaches are classified into three types based on how they develop and train the classifiers, i.e., supervised, semi-supervised, and unsupervised approaches (Thangaraj and Sivakami, 2018). Supervised learning-based techniques explore various features of given content and find a suitable set of features through binary classification. The training data is labeled by human annotators who manually interpret and categorize the training data (Nasteski, 2017). The most common algorithms of supervised learning applied for text classification are based on a Support Vector Machine (Huang et al., 2018), decision tree (Hang and Fong, 2010; Ville, 2013; Sharma et al., 2016), k-Nearest-Neighbors (Guo et al., 2003), and other binary classification techniques. While supervised learning approaches have achieved good performance (LeCun et al., 2015; Yang et al., 2016), the major drawback of such approaches is that their performance is highly dependent on the training data, and thus, they require a huge amount of labeled data for the training of classifiers. In addition, labeling data for training by human annotators is also a time-consuming and expensive

process. Therefore, these approaches are not suitable for the domains where standard labeled datasets are not available. The high dependence on training data also makes the classifiers difficult to adapt to new settings and real-world applications (Xie et al., 2020). Alternatively, there are unsupervised learning techniques that do not require any labeled data. Unsupervised learning algorithms determine how data can be organized into different groups called clusters. This involves learning from unlabeled data to classify new inputs. For example, Di Capua et al. (2016) suggested an unsupervised approach based on self-organizing maps to cluster the social data containing bully traces. Similarly, Chatzakou et al. (2017) used K-means algorithms to identify users' behavior and classify them accordingly. Unsupervised learning is useful when labeled data are not accessible; however, the performance is not always good. The merge of supervised and unsupervised techniques is the semi-supervised approach for text classification. The semi-supervised algorithms leverage a mixture of labeled data (as in supervised) and unlabeled data (as in unsupervised) to build learning models. Semi-supervised approaches are used when there are few labeled data and a lot of unlabeled data (Linmei et al., 2019). They thus allow to scale and adopt them to any application for real-time text analysis. For example, Xiang et al. (2012) adopted a semi-supervised approach for identifying offensive content in the Twitter corpus. They used the bootstrapping approach to label the unlabeled data based on very small set of generic information. Similarly, Meng et al. (2020) proposed to generate pseudo-labeled documents for training of the model from seed information. It involves the human efforts to just generate the labels and not to label the documents. One of the problems experienced by the semi-supervised techniques is the performance degradation over time as unlabeled data are added to the fixed set of labeled data.

## 2.3 Hybrid text classification approaches

Owing to the relative merits and demerits of rule-based and machine learning-based text analysis, many researchers have proposed hybrid approaches. For example, Zubiaga et al. (2015) adopted a hybrid approach for the trend analysis of Twitter data. They constructed a topology of trending topics as feature selection and then used automatic classification of tweets to appropriate trends. Similarly, Cai et al. (2016) adopted a hybrid approach by using a general concept list as a baseline and then constructed ontology based on the collaborative tags by the users. Similarly, Ingole et al. (2018) implemented a hybrid model by combining the Naïve Bayes classifier and the Support Vector Machine to identify the problems experienced by engineering students from Twitter data. They claim that the combination has yielded a sufficient reduction in training time and improved accuracy compared to applying them individually. While hybrid models have been successful in solving particular text classification problems, their performance on large text corpora depends on the test datasets. Therefore, there is no guarantee that the performance achieved on one dataset will be the same as with any other dataset (Irfan et al., 2015). Moreover, they require several parameters to be defined and initialized in advance. Thus, selecting a hybrid classification approach on social data totally depends upon the dataset and the problem being investigated (Miao et al., 2009). Similarly, research has suggested a few approaches devised around the concept of pooling wherein the selected documents from the given dataset are annotated with relevant concepts by the human experts to create a document collection (pool) of relevant documents for future referencing, reuse, and other information retrieval related tasks (Aslam et al., 2003; Losada et al., 2017; Otero et al., 2021). This implies that once a pool of relevant documents is collected, a machine-learning algorithm can be used to annotate future documents. While such approaches cut off human efforts to some extent by requiring only a subset of documents to be annotated manually, they still require human judgment and tend to be domain specific as they are applied to a single dataset at a time and required topic list to be supplied at the beginning of retrieval process.

## 2.4 Problem statement and research objectives

Despite the previous research work and substantial efforts of the research community, as discussed above, the task of real-time text classification of social data is still challenging and an emerging trend. People on social media use irregular data patterns to communicate and do not always use structured sentences, correct grammar, and spelling (Salloum et al., 2017). The use of informal language, character limitations, and symbols and slang makes the social data analysis a challenging task. In addition, the unpresented amount of unstructured data being generated in real time does not allow much time and a well-defined taxonomy to be processed. This presents two interrelated challenges: One, the use of semantical equivalents used by different users. Two, the lower level concepts that can be applied by higher level concepts. For example, the data with 'ICT' can be discussed by different users, such as ICT, information and communication technology, computers, and technology. Furthermore, the data discussing 'ICT' can be more related to lower-level concepts, such as ICT implementation, ICT adoption, ICT development, and ICT industry. In addition, the problems presented by user-generated data are multifold, including out-of-dictionary slang words, short forms, punctuation omissions, phonetic spellings, and misspellings (Clark et al., 2010). Thus, the majority of existing approaches still require some form of labeled data or human annotation to perform text classification efficiently and are designed for specific types of data and a particular domain (Rogers et al., 2021). These approaches designed for a particular domain are not directly implementable in another domain as this is not a fit for all tasks. For example, Wanichayapong et al. (2011) adopted an approach for real-time classification of traffic information from social data by leveraging a dictionary of related topics and a lexical analyzer. However, given the dictionary component, the application cannot be adopted to any other domain where a known list of words cannot be determined beforehand. Therefore, there is a need to develop more general information processing methods for the classification of data pertaining to a broad range of data types (Ingole et al., 2018; Kowsari et al., 2018).

In this realm, this study presents an approach for real-time classification of social data through corpus-based concept generation and text tagging. The approach is built upon the ideas

of feature extraction based on word semantics (as in Shi et al., 2005) and syntactic features of words (Wanichayapong et al., 2011). Hence, the proposed approach is capable of addressing the semantical and morphological similarities, as well as domain-specific vocabularies of query strings and tagged concepts in a general manner. The approach is different from existing approaches in a way that it does not require any dictionary, concept list, or labeled data beforehand. This makes it adaptable irrespective of any domain and can easily be implemented in any general-purpose application relating to social data retrieval and classification. Wherein, if need be, the application can also be made domain-specific just by referencing any suitable corpus related to that particular domain. Moreover, the approach can be applied to any language given the suitable corpora and corpus processing methods/services for that respective language. Thus, to the best of our knowledge, this approach is one of its kind that addresses the task of social data classification in a general way by offering a hybrid on-the-go method for concept generation and tagging of social data in real time.

## 3 Research design

The present study was conducted using the design science research methodology. DSRM is an applied science methodology that aims to develop technology-related innovative artifacts that address real-world problems. The resultant artifacts deliver utility and novelty (Hevner et al., 2004). The DSRM employed herein targets to develop a new corpus-based text classification and tagging approach for social data. While there are various models for undertaking DSRM research, this study has been conducted in five steps drawn upon the studies by Takeda et al. (1990) and Peffers et al. (2007), i.e., problem identification, objective description, solution design, test implementation, and evaluation (cf. Figure 1).

In doing so, first, the problem description was solicited using a problem-driven approach (Wieringa, 2009) from the literature review and state-of-the-art research area. In this regard, the nature and requirements of the classification of social big data were analyzed and compared with existing approaches to text classification. As a result, the shortcomings of existing approaches were outlined, and a problem statement was defined and specified (cf. Section 2.4). The problem statement was later used to outline the objectives of the intended approach that could guide its development and evaluation. Successively, the proposed solution was designed and iteratively improved to address the identified problem. The solution was designed in light of the identified capabilities of SNS APIs and other existing text similarity and expansion services (cf. Section 4). Successively, a test prototype of a web application that implements the proposed approach was developed to demonstrate the proof-of-concept that the proposed approach works to solve the instances of identified problems (cf. Section 5). The approach, through its instantiation, was evaluated on sample keywords to measure its efficiency and performance. The application was used to generate the concepts and tag the text snippets accordingly, which were then evaluated manually by the researchers to determine their relevancy in terms of precision measure. The evaluation was carried out formatively during the development process with controlled experiments (cf. Cleven et al., 2009) as the application was with the researchers who carried out the evaluation. The focus of these episodes of formative evaluations was to measure the developed approach against the identified research gap (cf. Section 6).

Following the DSRM concepts, the resultant approach presented in this study is designed in iterations. In doing so, a build-and-evaluate loop was exercised with formative evaluation during the development phase to seek feedback on the proposed approach and determine areas of improvement and refinement. This resulted in a re-iteration of the design process. Accordingly, this study consists of three iterations of the DSRM process model (cf. Figure 2). The first iteration comprises most of the work presented herein and starts by identifying the problem and setting solution objectives based on previous literature. Following this, the solution was designed and instantiated wherein the search of relevant data and concept generation and tagging for the retrieved data were performed based on only keywords given in the search query. The resulting first version of the proposed approach was then evaluated by the researchers in terms of its response rate, number of concepts generated, and precision of concept tagging. It was then observed that the performance was not good. Thus, the second iteration started from the solution design stage, wherein the solution was revised to use the original keywords given in the search query and their semantic equivalents. The implementation and evaluation of this iteration showed that the approach has improved in terms of response rate and the number of concepts generated. However, it was revealed that the approach still has a low response rate as it could not understand similarities between different forms of a word. Therefore, many text snippets having a different form of a given keyword were not recognized as relevant and returned by SNS APIs. This informed the third iteration, wherein keywords, their semantic equivalents, and morphological forms were also used for text retrieval, whereas concept generation and tagging were carried out with the keywords and their semantic equivalents only.

## 4 Proposed approach

The proposed corpus-based real-time text classification and tagging approach is given in Figure 3. According to the proposed approach, the intended goal of searching the social data and its tagging with relevant two-worded and three-worded concepts is achieved in five steps: keyword preprocessing, keyword expansion, data search, concept map generation, and data classification and tagging.

### 4.1 Step 1 – keyword preprocessing

In the first step, the given string of one or more keywords is cleaned and preprocessed for relevant search. Herein, at first, the search string is tokenized and broken into individual keywords. This is achieved through simple whitespace tokenization whereby a string is split into words from whitespaces. Successively, the bag of tokenized words (unigrams) is normalized, which refers to converting the keywords to their most canonical (standard) form

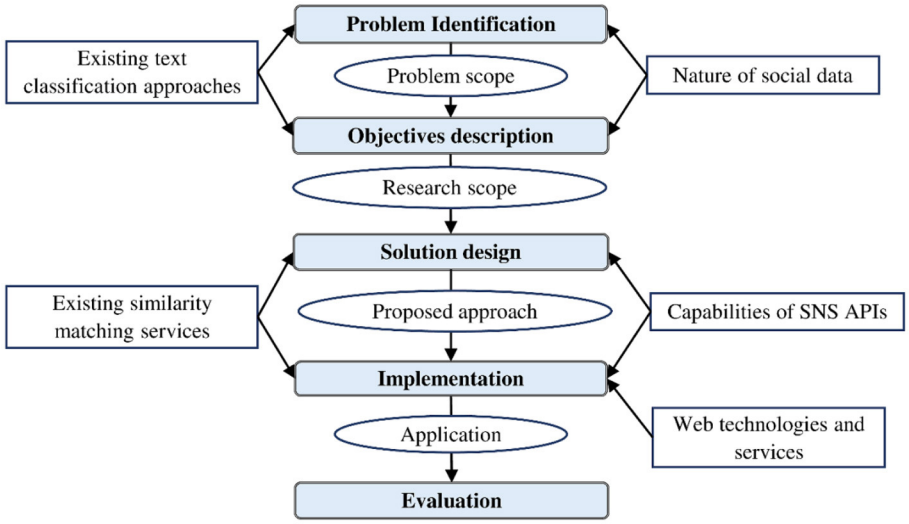


FIGURE 1
Overall research procedure of the study.

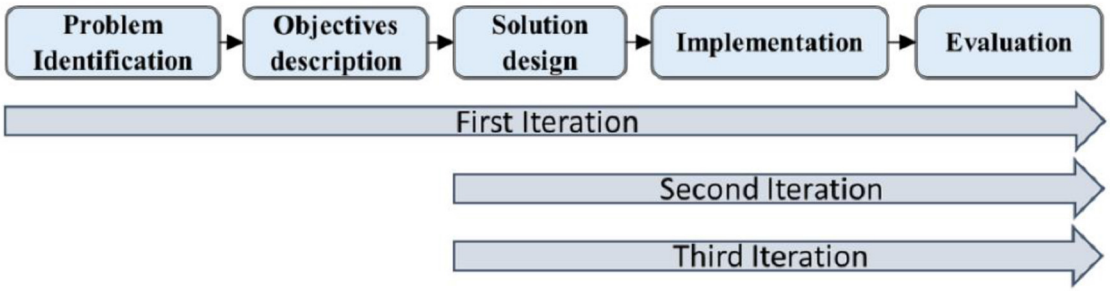


FIGURE 2
Iterations of the DSRM process model of the study.

through the process of lemmatization and stemming techniques (Clark and Araki, 2011). This is necessary to address the problems of user-generated data, including slang words, short forms, punctuation omissions, phonetic spellings, and misspellings (Clark et al., 2010). Normalizing words to their base form helps to reduce the number of unique tokens and remove the word variations and redundancies. The normalized words are then parsed and filtered to discard any stop words. Stop words refer to the most common words with grammatical functions (syntactic in nature more than semantic) and do not relate more with the actual content (Salton, 1989). Research shows that removing such words from search keywords improves the performance of information retrieval systems (Wilbur and Sirotkin, 1992) in terms of efficiency and effectiveness (El-Khair, 2017). To do so, a stop word list is identified containing both the domain-independent (of any specific language) and domain-specific stop words (of a particular domain), and then search tokens are matched with these identified stop words through any suitable text-matching approach (El-Khair, 2017). As a result, the resultant bag of words contains the most appropriate, relevant, and standardized tokens that can be used to search the relevant data.

## 4.2 Step 2 – keyword expansion

In the second step, the bag of keywords generated in step 1 is semantically and morphologically expanded to retrieve more relevant data and improve the information retrieval process. The purpose of the semantical expansion of query keywords is to include all other alternate words in the query that might mean the same as the given keywords to increase the search recall and precision rates (Jain et al., 2021). This is to address the semantical similarities of words in domain-independent as well as domain-dependent terminologies. Therefore, the corpus-based semantic word generation and expansion is levied here. Using a corpus-based approach helps to capture all similarities (domain-specific as well, given that domain-related literature is referenced) and makes the approach and resultant web tool easily adaptable to any domain and any language by just referencing an appropriate domain corpus

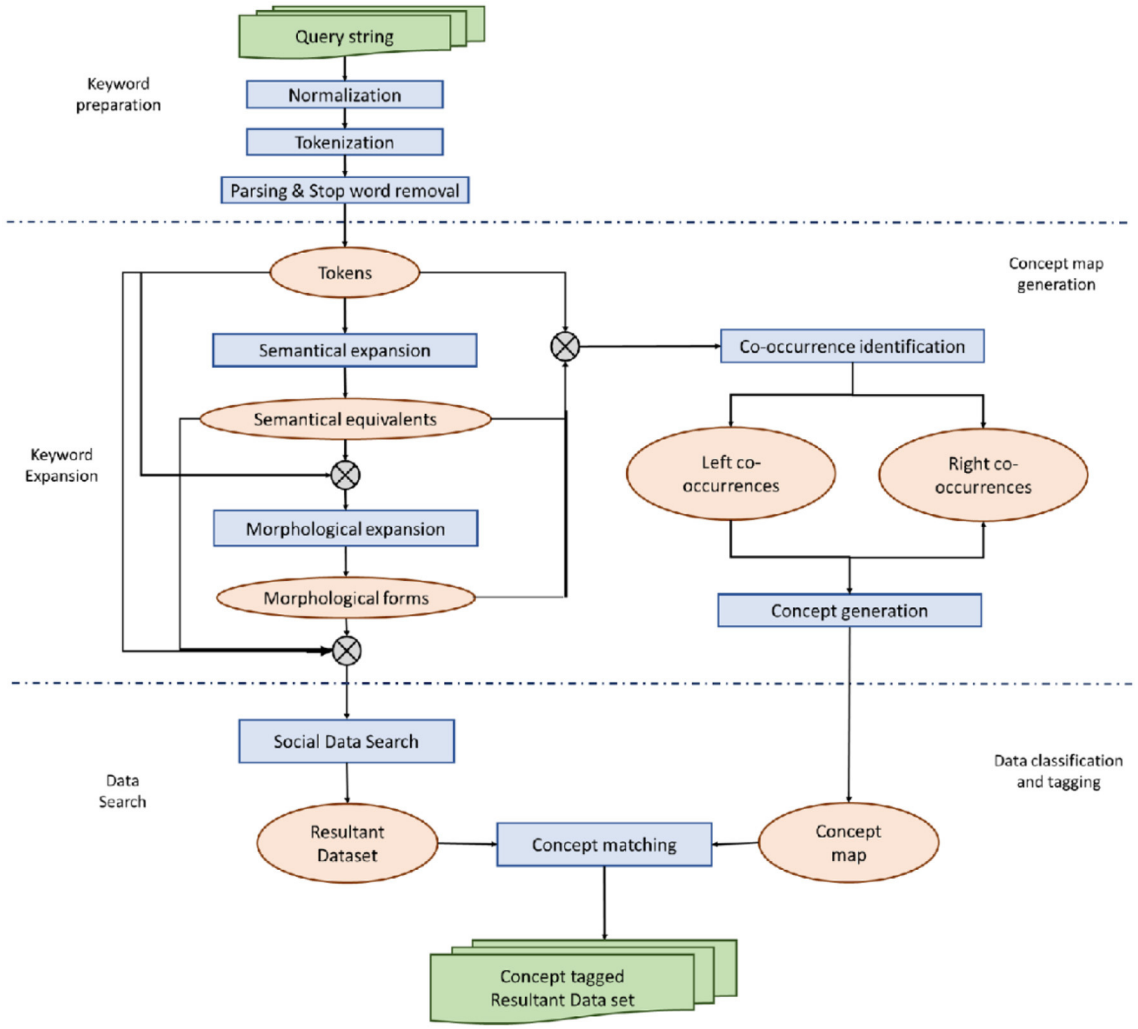


FIGURE 3
Proposed corpus-based real-time text classification and tagging approach for social data.

in a particular language. This results in a collection of actual keywords preprocessed in the previous step and all semantically equivalent words. The resultant bag of words is then processed for the morphological extension. Morphological expression refers to different forms of a particular word depending on the language lexical, which are referred to as 'morphs' of a particular word. This is necessary as the text similarity measures, especially vector-based text similarity approaches, work on the exact matching of keywords in the underlying text data sources. Thus, including all forms of a particular word again improves the precision and recall of information retrieval. This results in a bag of words containing actual preprocessed keywords, all semantically equivalent words of given keywords, and all morphological forms of keywords and their semantical equivalents. Such that, taking each keyword at a time (say $K_i$), its semantically similar words ($K_iS_1$ to $K_iS_n$) are generated, and subsequently, all its morphological forms ($K_iM_1$ to $K_iM_n$) and morphological forms of its semantical equivalents ($K_iS_1M_1$ to $K_iS_nM_n$) are generated. This step results in a bag of words (T) for the search string with N keywords as $T = \sum_{i=1toN} (K_i, K_iM_1, \ldots, K_iM_n, K_iS_1, \ldots, K_iS_n, K_iS_1M_1, \ldots, K_iS_1M_n, \ldots\ldots\ldots, K_iS_nM_1, \ldots, K_iS_nM_n)$.

## 4.3 Step 3 – searching social data

In the third step, the bag of words generated in the second step is used to query and retrieve the relevant data from respective social networks. For searching social networks, a suitable data search methodology is adopted like the one proposed by Memon and Meyer (2020). Accordingly, the data are searched through five modules (cf. Figure 4): (1) query processor module that prepares the semantic search based on given keywords to eliminate the vocabulary mismatch problem and improve search recall, (2) query formulator module, wherein the query is formulated according to the structure and query format of the respective SNS APIs and proper authentication information is attached to it, (3) data fetcher module that issues repetitive SNS API calls to fetch all relevant data while addressing particular data limits of each API, (4) data mapper

module that removes platform dependencies from retrieved data and maps all data to a uniform data model, and (5) data integrator model that aggregates all the data retrieved from different SNSs into a single dataset while discarding any potential duplicates.

### 4.4 Step 4 – concept map generation

In the fourth step, the actual keywords and their semantic equivalent words are used to generate data concepts. Herein, semantically similar words are used to refer to all relevant words of given keywords. The resultant bag of words is used to identify the left and right co-occurrences of each word (also termed as neighboring terms) based on the semantic analysis of a suitable corpus. In this regard, either the domain-independent or the domain-dependent corpus can be used according to the specific search application. After retrieving the possible left and right neighboring terms of each word, its collations with each left neighboring term (left co-occurrence concept), its each right neighboring term (right co-occurrence concept), and different combinations of left and right neighboring terms (three-worded concepts) are created. As a result, taking each keyword at a time (say $W_i$), its left neighboring terms ($L_{wi}$) and right neighboring terms ($R_{wi}$) are generated. Following this, concepts are generated, which result in a concept map containing the concepts as $\sum_{i=1toN}$ ($L_{wi+}$ $R_{wi+}$ $L^*_{wi}R_{wi}$).

### 4.5 Step 5 – data classification and tagging

For determining the similarity of each concept with the text snippets, the cosine similarity measure is used as it is regarded as the most widely used for determining text similarity (Huang, 2008) and provides the foundation for document classification in many studies (Shrestha, 2011). The process of determining the text similarity through cosine similarity measure is to transform each category tag and each text snippet in the retrieved dataset into a vector in some high dimensional space so that we find the similar strings being close to each other. The cosine value of the angle shows how similar the two texts are and thus represents the similarity of the two texts.

Accordingly, a high dimensional space V is created where each term in each concept in concept map T1 and text snippet T2 defines an independent dimension. The texts are then transformed into their respective binary vectors, V1 and V2. With respect to the positive quadrant of the Euclidean space, wherein no term is assigned a negative value, vector values are used to indicate whether a particular term in each text string is present (1 – identical vector) or not (0 – orthogonal vector). Finally, the cosine of the angle between V1 and V2 is computed, which is identical to their normalized inner product, such that $Sim_{cos}$ (T1, T2) $= \frac{V1 \,.\, V2}{\sqrt{V1^2}\sqrt{V2^2}}$.

## 5 Implementation

The approach presented in Figure 3 is a test implemented in designing a web-based application that enables the search for social data from the Facebook platform as well as from the application's native text resources in an aggregate manner. The approach facilitates the synthesis of well-defined query strings by suggesting possible left and right co-occurrences and semantically similar terms. Successively, the application allows searching text resources from connected data servers and then generates concepts and classifies and tags the retrieved text resources accordingly.

### 5.1 User interface

The initial user interface of the application is divided into 4 parts (cf. Figure 5):

- The first segment (top panel) provides a search box for entering keywords; as the user types in the keyword and hits a blank space, signaling the end of a word, the application presents alternative semantically similar terms of a given keyword on the top of the search box. The user can then select any alternative word instead of the given word if found more suitable. In addition, as the user finishes a word, the application also suggests the various possible left and right neighboring terms so that the user may add more relevant collations to the search query. The final search button also finalizes the search string and issues a search command to the server.
- The second segment (left panel) presents a concept map generated according to the proposed approach and organized around given search keywords. In addition to each concept, the number of resources is indicated that match the particular concept. Clicking on a particular concept allows filtering the result set to show only those resources that match the given concept.
- The third segment (middle panel) lists the retrieved resources matching the search query. The resources are organized with respect to different data sources.
- The fourth segment (right panel) shows the details of a single selected resource. As an additional function, this part enables the user to interact with the particular resource and thereby like or comment on the retrieved social media post directly within the application.

### 5.2 Technical implementation

The application is configured in a typical three-tier style using a WAMP server architecture. The application is implemented in PHP as a server-side scripting language and HTML and JavaScript with jQuery library[1] as client-side scripting languages using a relational MySQL database for general user management tasks. Dynamic content per user request is generated through Asynchronous JavaScript calls with JavaScript Object Notation, and JavaScript Underscore library[2] is used for results sorting and filtering at the client side. The local database is accessed with SQL commands in PHP scripts. The interconnectivity with the Facebook platform is achieved through Facebook Graph API.[3]

1 JQuery library: https://jquery.com/.

2 Underscore library: http://underscorejs.org/.

3 Facebook graph API: https://developers.facebook.com/docs/graph-api/.

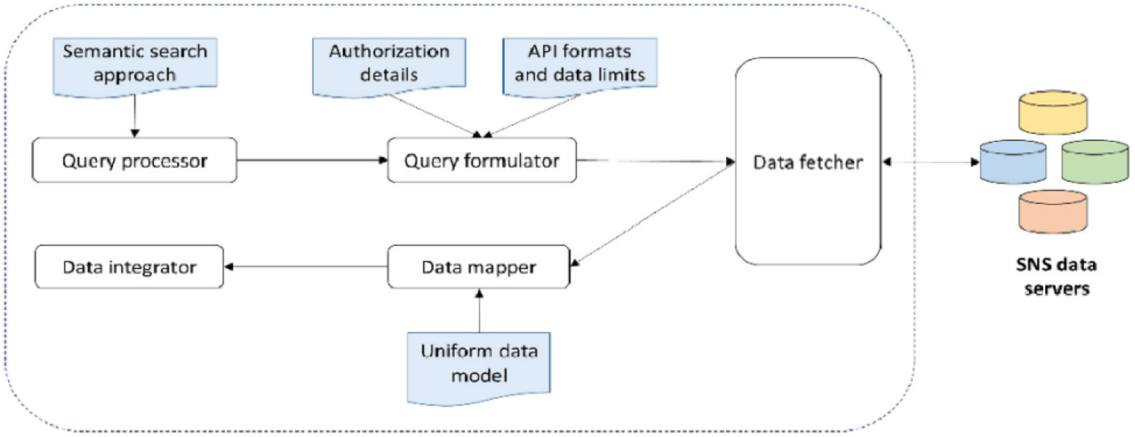


FIGURE 4
Social data search procedure by Memon and Meyer (2020).

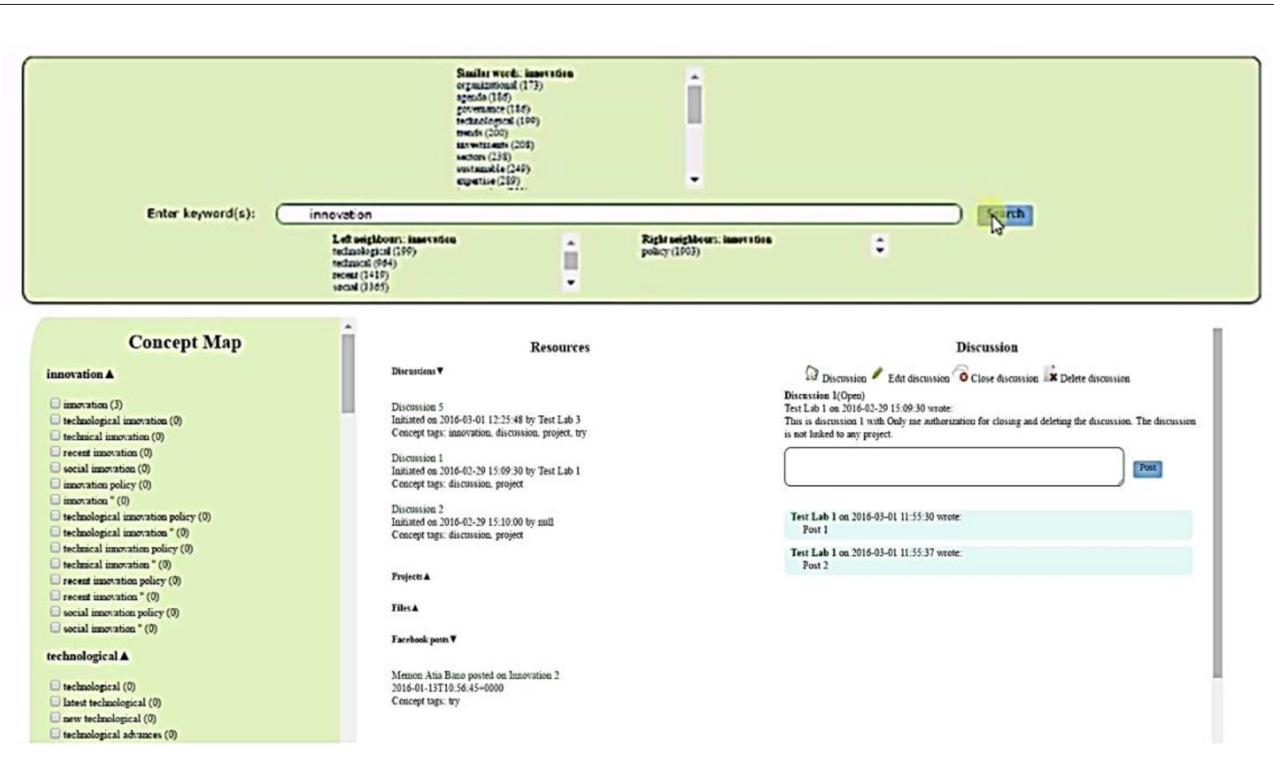


FIGURE 5
A screenshot of a text application implementing the proposed approach.

For the morphological expansion of keywords, the application uses the PHP-based phpMorphy library,[4] which is a dictionary-based morphological service for different languages, including English, Russian, German, Ukrainian, and Estonian. The phpMorphy service provides three types of information for each given word, including the lemma, which is the base form of the word, all morphological forms of the word, and grammatical information of the word. Herein, the lemma and morphological forms of the given keyword are retreived and sued by the application.

For the semantical expansion of keywords and during concept generation, the application leverages the corpus-based web services offered by Leipzig Corpora Collection[5] using its REST API.[6] LCC collects the large corpus using different data collection

4 PhpMorphy library: http://phpmorphy.sourceforge.net/dokuwiki/.

5 Leipzig Corpora Collection: http://corpora.uni-leipzig.de/.

6 Leipzig Corpora Collection API: https://api.wortschatz-leipzig.de/ws/swagger-ui/index.html.

methods and in different languages from freely available electronic resources such as the web, newspapers, and Wikipedia. It processes the collected corpus and calculates a full-form dictionary with frequency information for each word. In addition, it precomputes statistics related to the significant co-occurrence/neighboring terms and stores two kinds of co-occurrence information, i.e., words that occur together in sentences and words that are found as immediate (left or right) neighbors (Richter et al., 2006; Biemann et al., 2007). The LCC collects corpora for different languages, and for every language, different releases are available. The application references their English language corpus based on Wikipedia with 1M sentences (eng_wikipedia_2012_1M). The complete algorithm for the concept map generation and text snippet tagging is given in Listing 1.

**Algorithm:** Generates concepts on the basis of co-occurrences and successively tags the resources with relevant concepts

**Input:** list of keywords *Keywords*, list of retrieved resources *Resources,* list of stop words *Stop_words*
**Output:** List of concepts *Concepts,* list of tagged resources *Resources*

```
/* Generate Concepts */
foreach word W in Keywords do
    Array Semantical_equivalents = retrieve semantically similar words of W;
    foreach semantical equivalent S in Semantical_equivalents do
        if S is not in Keywords then
            add S in Keywords
        end
    end
 end
foreach word W in Keywords do
    Array Left_neighbours = get left neighbouring terms of W;
    Array Right_neighbours = get right neighbouring terms of W;
    l = length of Left_neighbours;
    for i = 1 to l do
        if Left_neighbours[i] is not in Stop_words then
            add Left_neighbours[i] + W in Concepts;
        end
    end
    r = length of right_neighbours;
    for j = 1 to r do
        if Right_neighbours[i] is not in Stop_words then
            add W + Right_neighbours[j] in Concepts;
        end
    end
    for i = 1 to l do
        if Left_neighbours[i] is not in Stop_words then
            for j = 1 to r do
                if Right_neighbours[i] is not in Stop_words then
                    add Left_neighbours[i] + W + Right_neighbours[j] in Concepts;
                end
            end
        end
    end
 end
/* Add concept tags to resources */
foreach resource Resource in Resources do
    Array Concept_tags;
    foreach concepts Concept in Concepts do
        Concept_terms = tokenize Concept;
        n = length of Concept_terms;
        for i = 1 to n do
            Sim = match Concept_terms [i] and Resource;
            Sim > 0 then
                continue;
            end
            else
                break;
            end
        end
        if i = = n and Sim > 0 then
            add Concept in Concept_tags;
        end
    end
    add Concept_tags to Resource metadata;
end
return Concepts, Resources;
```

Listing 1. Algorithm for the concept map generation and tagging of text snippets.

## 6 Evaluation

The designed application has been tested and validated by integrating it into the prototype of a larger collaboration platform that supports interconnectivity among a specific type of organization throughout the collaboration process. The results of a sample execution of the proposed approach are shown in Figure 6. For the sample execution, the keyword "Innovation" is used, and therein, 2-worded and 3-worded concepts are generated to verify the functionality of the application. The application was able to generate 41 two-worded concepts (32 with right co-occurrences and 9 with left co-occurrences) and 288 three-worded concepts.

The proposed approach and the implementing prototype have been evaluated formatively during the continuous design and development cycles. The study adopted a controlled experiment evaluation strategy in an interpretive manner (Venable et al., 2012, 2016), wherein the researchers served as human evaluators for the different versions of the approach. As discussed in Section 3, the resultant approach is devised in three main iterations. At the end of each iteration, the concept map generated by the approach was analyzed by three researchers to assess the performance and outline the improvements for the next iteration. Herein, precision was calculated as the performance indicator. The precision of the concept map is defined as the percentage of meaningful concepts out of the total concepts generated such that

$$\text{Precision} = \frac{\textit{Number of relevant concepts generated}}{\textit{Total number of concepts generated}} * 100$$

Accordingly, to assess the concept map, the concept list generated by the approach was manually checked by the evaluators to judge if a particular concept is related to the given keyword. A particular concept was considered relevant if it was marked relevant by at least two of the evaluators.

As described in the methodology section, in the first iteration of development, only the given keyword was used to create concepts. The evaluation showed that the performance was good in terms of precision; however, the number of identified concepts was not good, as the evaluators could identify several concepts that were not generated by the system. Those unidentified concepts were mainly oriented around semantic equivalents of given keywords. For example, given the word "innovation," the system was able to identify "technological innovation"; however, it failed to create a concept as "new technology" while they both are different variants of the same concept. This initiated the second iteration, whereby the word and its semantic equivalents were used to generate concepts. This resulted in a rise in the number of identified concepts, but it was observed that there is still room for improvement. For instance, given the keyword "computer," the system was not able to generate concepts involving the word "computerized" even though it is just another form of the given base term. Thus, in the third iteration of development, the morphological forms of keywords were also used to generate concepts in addition to the word and its semantic equivalents. Contrary to the expectations, this addition increased the number of identified concepts to some extent; it lowered the precision because of more number of meaningless concepts created by different forms of the words. Thus, it was decided to identify concepts based on keywords and semantically equivalent terms only. The test results of each version of the proposed approach for the three keywords are shown in Table 1.

Subsequently, the accuracy and relevancy of tagged concepts were tested on a set of three keywords, and for each keyword, a set of three relevant text documents was used for tagging. The documents and tagged concepts were then manually reviewed by the evaluators to determine the relevancy of tagged concepts and calculate the precision. The precision of tagged concepts is defined as the percentage of relevant tags out of the total number of tags supplied to a particular text snippet such that

$$\text{Precision} = \frac{\textit{Number of relevant tags}}{\textit{Total number of tags predicted}} * 100.$$

In doing so, the text segments and the associated tags were reviewed by the evaluators to judge if a particular tagged concept was relevant to the given text. A particular tag was considered relevant if it was marked relevant by at least two of the evaluators. The tests are given in Table 2.

It is important to mention here that the number of identified concepts is highly dependent on the completeness and relatedness of the referenced corpora, as is the case with all corpus-based approaches. It is also interesting to note from the results that as the number of related concepts increased (by incorporating more synonyms of given keyword and their co-occurrences), the precision of generated concepts decreased. However, this can be controlled by referencing a good quality and more relevant corpus for identifying the semantically equivalent terms at the beginning of the proposed approach. Another configuration in this regard could be to use a linguistic dictionary (for general-purpose applications) or a pre-identified synonym list for defined keywords (for domain-specific applications). Nevertheless, as given in Tables 1, 2, the proposed approach is able to achieve an average precision of 82.25% for the list of concepts identified and a precision of 91.04% for the concept tags as annotated to particular text segments. Thus, it can be concluded that it is effective in achieving its stated goal of classifying text snippets in real time.

## 7 Conclusion

The current study has proposed a corpus-based approach for searching and successively categorizing and tagging social data with relevant concepts in real time. The proposed approach is capable of addressing the semantic and morphological similarities and domain-specific vocabularies of the query string and tagged concepts. The feasibility and application of the proposed approach are demonstrated in a web-based tool that allows searching the post from the Facebook platform and local data resources of the application and provides search results together with a concept map for further navigation, filtering, and refining of search results. The tool has been evaluated by performing multiple search queries, and resultant concept tags are analyzed in terms of relevancy and precision in an interpretive manner by the researchers themselves in controlled settings. The approach is thereby found effective

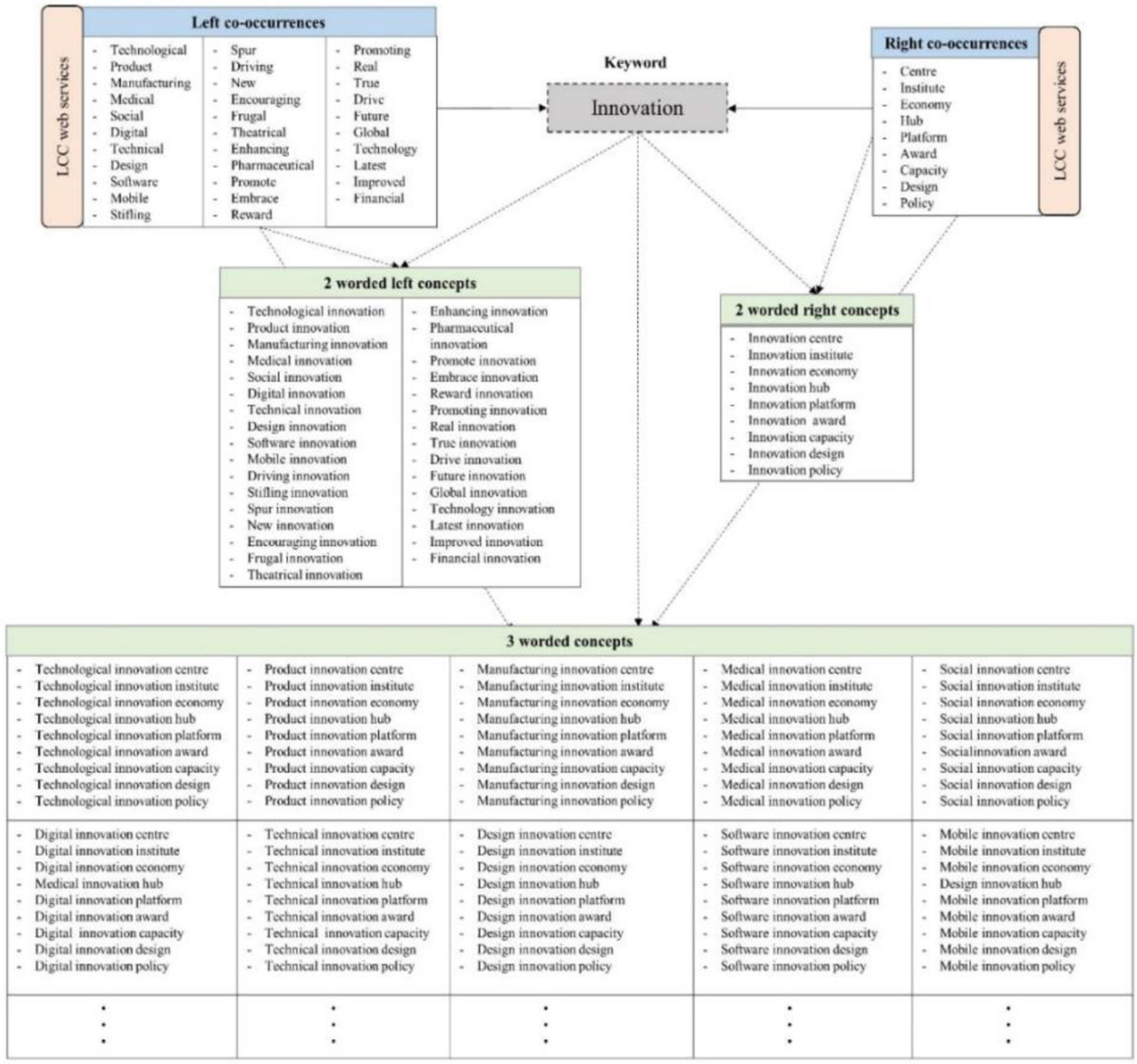


FIGURE 6
Sample concept generation for the keyword "Innovation."

in achieving its stated goal of classifying text snippets in real time.

## 7.1 Contributions and implications

The study presented herein has two contributions that have implications for the theory and practice.

One, the proposed approach adds to the literature on the topics of real-time text classification of social data and corpus-based text classification and tagging approaches. It suggests how corpus-based hybrid approaches can be designed to achieve the task in real time without much human effort and processing time. In addition to offering a solution to an identified problem, it opens a new research direction that calls attention from the research community to investigate the task on a deeper level. This would serve to bring new innovative approaches to the literature and also provide the grounds to cross-compare, assess, and debate on the utility and effectiveness of different approaches in different contexts. In addition, this study adds to the literature on design science research in a way that offers an example of how design science research methodology can suggest and guide the design and development of such approaches. It indicates how the formative evaluation incorporated during the development of technical artifacts can yield a better understanding of the problem and a suitable solution to the identified problem.

10.3389/fcomp.2024.1294985

TABLE 1 Evaluation results of different versions of the proposed approach (* as compared to baseline with keyword only).

| Keyword | Keyword only | | | Keyword + Semantical equivalents (Proposed approach) | | | Keyword + Semantical equivalents + Morphological forms | | |
|---|---|---|---|---|---|---|---|---|---|
| | No. of identified concepts | No. of meaningful concepts | Precision (%) | No. of identified concepts | No. of meaningful concepts | Precision (%) | No. of identified concepts | No. of meaningful concepts | Precision (%) |
| Computer | 198 | 176 | 88.88 | 329 | 287 | 87.23 | 412 | 326 | 79.12 |
| Business | 312 | 268 | 85.89 | 506 | 435 | 85.96 | 598 | 463 | 77.42 |
| Development | 296 | 232 | 78.37 | 424 | 312 | 73.58 | 534 | 352 | 65.91 |
| Average | | | 84.38 % | | | 82.25 % | | | 74.15 % |
| Average number of identified concepts* | | | | + 56.20% | | | + 91.56% | | |
| Average precision of concepts* | | | | −2.13% | | | −10.23% | | |

Two, the approach and specifically the developed prototype showcases how social data can be retrieved through SNS APIs and processed to be leveraged in designing data-intensive applications. The application further shows how corpus-based services can be used to serve different tasks in an on-the-go fashion. Accordingly, this offers substantial motivation and foundational grounds for the application designers and developers to conceptualize and implement data-intensive applications based on social data to solve various real-life tasks in real time. The application developed herein, encompassing the proposed approach, is intended to serve the end users in the real world toward searching and accessing social data in an effective manner in different contexts. As the approach provides a way to classify textual social data, it can be utilized standalone or incorporated in any other large-level web-based platform or natural language processing system dealing with information search and retrieval. When integrated within any system, the approach may aid in achieving a sub-task of social data classification, and its output (i.e., concept map and tagged data) can then be exploited differently based on the purpose of the enclosing platform.

## 7.2 Limitations and outlook

The current study has two limitations, which in turn have implications for further research. First, while the approach presented herein substantially offers a solution to the identified problem of real-time text classification of social data and opens a new direction for the application developers to conceptualize and implement more such approaches, it is worth to be mentioned that data search and retrieval from social media as proposed in this approach highly relies on the capabilities of SNS APIs and data permissions and is thus vulnerable to their changes in future. Given the current state of authorization standards of existing SNS APIs and provided that the terms and conditions of their data usage are properly followed, the data can be accessed and leveraged for different purposes. However, there will be a need to continuously update such applications in line with any future changes in SNS data policy. In addition, the approach has been evaluated with only Facebook data so far, and thus, there is a need to test the approach with data retrieved from other social networks as well to demonstrate its utility for varying types of text datasets.

Second, the application has been evaluated in an isolated manner by the researchers adopting a controlled and interpretive strategy. The aim of the episodes of formative evaluation undertaken during this study was to establish the utility and efficacy of the proposed approach by showing how well it performs. While we believe that approach substantially solves the identified problem, there is still a need to deploy it in real-world settings so that a summative evaluation on a larger level with real-end users can be sought. The feedback from such evaluation will be useful to see its efficacy in broader terms and also indicate how the approach performs over time for different users with varying search and tagging requirements. In addition, as the scope of the presented study at the current level was to develop a generic application that can be leveraged in a general context irrespective of any domain specificity, it has not been evaluated in terms of

TABLE 2 Evaluation results of concept tagging of the proposed approach.

| Keyword | No. of generated concepts | | | | Document | No. of tagged concepts | No. of correctly tagged concepts | Precision (%) |
|---|---|---|---|---|---|---|---|---|
| | 2-worded left concepts | 2-worded right concepts | 3-worded concepts | Total concepts | | | | |
| Computer | 33 | 9 | 297 | 339 | Text 1 | 315 | 297 | 94.28 |
| | | | | | Text 2 | 291 | 272 | 93.47 |
| | | | | | Text 3 | 287 | 235 | 81.88 |
| Business | 12 | 38 | 456 | 506 | Text 1 | 413 | 376 | 91.04 |
| | | | | | Text 2 | 487 | 472 | 96.91 |
| | | | | | Text 3 | 476 | 447 | 93.90 |
| Development | 24 | 16 | 384 | 424 | Text 1 | 388 | 339 | 87.37 |
| | | | | | Text 2 | 354 | 312 | 88.13 |
| | | | | | Text 3 | 397 | 367 | 92.44 |
| Average | | | | | | | | 91.04 % |

any domain-specific attributes. Furthermore, to the best of our knowledge, there is no existing approach that can classify text snippets in real time without requiring any sort of predefined list of topics/concepts relating to a particular field. Therefore, it has not been cross-compared with other approaches as they are all suitable for respective domains. Thus, adopting the proposed approach to different domains (by referencing suitable domain-specific corpora) would be beneficial in assessing if and how the approach performs for the text classification of domain-specific datasets. This will also enable to cross-compare the approach with other approaches designed and applied to that particular domain to test and showcase its relative utility and effectiveness.

## 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

AM: Conceptualization, Validation, Writing – original draft, Methodology, Software. DS: Writing – review & editing, Data curation, Formal analysis, Investigation. KL: Investigation, Software, Validation, Writing – review & editing. KM: Conceptualization, Funding acquisition, Resources, Supervision, Validation, Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The authors are thankful to the Mittweida University of Applied Sciences for the financial support.

## 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



## References

Abu-Salih, B., Wongthongtham, P., Zhu, D., Chan, K. Y., and Rudra, A. (2021). Social big data: an overview and applications. *Soc. Big Data Anal. Prac. Tech. Appl.* 2, 1–14. doi: 10.1007/978-981-33-6652-7_1

Aggarwal, C. C., and Zhai, C. (2012). A survey of text classification algorithms. *Mining Text Data* 22, 163–222. doi: 10.1007/978-1-4614-3223-4_6

Ali, A., Qadir, J., Rasool, R. U., Sathiaseelan, A., Zwitter, A., Crowcroft, J., et al. (2016). Big data for development: applications and techniques. *Big Data Anal.* 1, 1–24. doi: 10.1186/s41044-016-0002-4

Al-Khalifa, H. S., and Davis, H. C. (2007). Exploring the value of folksonomies for creating semantic metadata. *IJSWIS* 3, 12–38. doi: 10.4018/jswis.2007010102

Aslam, J. A., Pavlu, V., and Savell, R. (2003). "A unified model for metasearch, pooling, and system evaluation," in *Proceedings of the Twelfth International Conference on Information and Knowledge Management* (New York, NY: ACM), 484–491.

Bai, R., Wang, X., and Liao, J. (2009). "Folksonomy for the blogosphere: Blog identification and classification," in *2009 WRI World Congress on Computer Science and Information Engineering.* Piscataway, NJ: IEEE.

Beigi, G., Hu, X., Maciejewski, R., and Liu, H. (2016). An overview of sentiment analysis in social media and its applications in disaster relief. *Sentiment Anal. Ontol. Eng. Environ. Comput. Int.* 12, 313–340. doi: 10.1007/978-3-319-30319-2_13

Bello-Orgaz, G., Jung, J. J., and Camacho, D. (2016). Social big data: recent achievements and new challenges. *Inf. Fusion* 28, 45–59. doi: 10.1016/j.inffus.2015.08.005

Biemann, C., Heyer, G., Quasthoff, U., and Richter, M. (2007). The Leipzig corpora collection-monolingual corpora of standard size. *Proceedings of Corpus Linguistic* 2007.

Bullinger, A. C., Hallerstede, S. H., Renken, U., Soeldner, J. H., and Moeslein, K. M. (2010). *Towards research collaboration-a taxonomy of social research network sites. Stem Cells* 42, 107–115.

Cai, Y., Chen, W. H., Leung, H. F., Li, Q., Xie, H., Lau, R. Y. K., et al. (2016). Context-aware ontologies generation with basic level concepts from collaborative tags. *Neurocomputing* 208, 25–38. doi: 10.1016/j.neucom.2016.02.070

Cantador, I., Konstas, I., and Jose, J. M. (2011). Categorising social tags to improve folksonomy-based recommendations. *J. Web Semantic.* 9, 1–15. doi: 10.1016/j.websem.2010.10.001

Chatzakou, D., Kourtellis, N., Blackburn, J., de Cristofaro, E., Stringhini, G., and Vakali, A. (2017). "Hate is not binary: studying abusive behavior of# gamergate on twitter," in *Proceedings of the 28th ACM Conference on Hypertext and Social Media* (New York, NY: ACM), 65–74.

Chen, H., Chiang, R. H. L., and Storey, V. C. (2012). Business intelligence and analytics: from big data to big impact. *MIS Q.* 12, 1165–1188. doi: 10.2307/41703503

Chen, Y., Zhou, Y., Zhu, S., and Xu, H. (2012). "Detecting offensive language in social media to protect adolescent online safety," in *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing.* Piscataway, NJ: IEEE.

Chugani, M., Vatsal, S., Ramena, G., Purre, N., and Moharana, S. (2021). "On-device tag generation for unstructured text," in *2021 IEEE 15th International Conference on Semantic Computing (ICSC).* Piscataway, NJ: IEEE.

Clark, E., and Araki, K. (2011). Text normalization in social media: progress, problems and applications for a pre-processing system of casual English. *Procedia-Soc. Behav. Sci.* 27, 2–11. doi: 10.1016/j.sbspro.2011.10.577

Clark, E., Roberts, T., and Araki, K. (2010). "Towards a pre-processing system for casual English annotated with linguistic and cultural information," in *Proceedings of the Fifth IASTED International Conference* (Maui, HI), 44–84.

Cleven, A., Gubler, P., and Hüner, K. M. (2009). "Design alternatives for the evaluation of design science research artifacts," in *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology* (New York, NY: ACM), 1–8.

Culotta, A. (2010). "Towards detecting influenza epidemics by analyzing Twitter messages," in *Proceedings of the First Workshop on Social Media Analytics* (New York, NY: ACM), 115–122.

Di Capua, D., Di Nardo, M. E., and Petrosino, A. (2016). "Unsupervised cyber bullying detection in social networks," in *2016 23rd International Conference on Pattern Recognition (ICPR).* Piscataway, NJ: IEEE.

El-Khair, I. A. (2017). Effects of stop words elimination for Arabic information retrieval: a comparative study. *arXiv [Preprint].* arXiv:1702.01925. Available online at: https://arxiv.org/ftp/arxiv/papers/1702/1702.01925.pdf

Felt, A., and Evans, D. (2008). *Privacy Protection for Social Networking APIs.* Web 2.0 Security and Privacy (W2SP'08).

Gandomi, A., and Haider, M. (2015). Beyond the hype: big data concepts, methods, and analytics. *Int. J. Inf. Manage.* 35, 137–144. doi: 10.1016/j.ijinfomgt.2014.10.007

George, G., Haas, M. R., and Pentland, A. (2014). Big data and management (No. 2). academy of management Briarcliff Manor, NY. *Acad. Manage. J.* 57:4002. doi: 10.5465/amj.2014.4002

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Dec. Supp. Syst.* 61, 115–125. doi: 10.1016/j.dss.2014.02.003

Ghosh, S., Viswanath, B., Kooti, F., Sharma, N. K., Korlam, G., Benevenuto, F., et al. (2012). "Understanding and combating link farming in the twitter social network," in *Proceedings of the 21st International Conference on World Wide Web* (New York, NY: ACM).

Godoy, D., and Corbellini, A. (2016). Folksonomy-based recommender systems: a state-of-the-art review. *Int. J. Int. Syst.* 31, 314–346. doi: 10.1002/int.21753

Guo, G., Wang, H., Bell, D., Bi, Y., and Greer, K. (2003). "KNN model-based approach in classification," in *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy.* Cham: Springer.

Hang, Y., and Fong, S. (2010). "An experimental comparison of decision trees in traditional data mining and data stream mining," in *2010 6th International Conference on Advanced Information Management and Service (IMS). Symposium conducted at the meeting of IEEE.* Piscataway, NJ: IEEE.

Hevner, A. R., March, S. T., Park, J., and Ram, S. (2004). Design science in information systems research. *MIS Q.* 21, 75–105. doi: 10.2307/25148625

Huang, A. (2008). "Similarity measures for text document clustering," in *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008),* Christchurch, New Zealand.

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., Xu, W., et al. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genom. Proteom.* 15, 41–51. doi: 10.21873/cgp.20063

Ingole, P., Bhoir, S., and Vidhate, A. V. (2018). "Hybrid model for text classification," in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA).* Piscataway, NJ: IEEE.

Irfan, R., King, C. K., Grages, D., Ewen, S., Khan, S. U., Madani, S. A., et al. (2015). A survey on text mining in social networks. *The Knowledge Eng. Rev.* 30, 157–170. doi: 10.1017/S0269888914000277

Jain, S., Seeja, K. R., and Jindal, R. (2021). A fuzzy ontology framework in information retrieval using semantic query expansion. *International Journal of Inf. Manage. Data Insights* 1, 100009. doi: 10.1016/j.jjimei.2021.100009

Kaur, S., Singh, S., and Kaushal, S. (2021). Abusive content detection in online user-generated data: a survey. *Proc. Comput. Sci.* 189, 274–281. doi: 10.1016/j.procs.2021.05.098

Kowsari, K., Brown, D. E., Heidarysafa, M., Meimandi, K. J., Gerber, M. S., Barnes, L. E., et al. (2017). "Hdltex: Hierarchical deep learning for text classification," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA).* Piscataway, NJ: IEEE.

Kowsari, K., Heidarysafa, M., Brown, D. E., Meimandi, K. J., and Barnes, L. E. (2018). "Rmdl: random multimodel deep learning for classification," in *Proceedings of the 2nd International Conference on Information System and Data Mining* (Lakeland, FL). doi: 10.1145/3206098.3206111

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., Brown, D., et al. (2019). Text classification algorithms: a survey. *Information* 10:150. doi: 10.3390/info10040150

Kowsari, K., Yammahi, M., Bari, N., Vichr, R., Alsaby, F., and Berkovich, S. Y. (2015). Construction of fuzzyfind dictionary using golay coding transformation for searching applications. *arXiv [Preprint].* arXiv:1503.06483. Available online at: https://arxiv.org/ftp/arxiv/papers/1503/1503.06483.pdf

Lai, S., Xu, L., Liu, K., and Zhao, J. (2015). "Recurrent convolutional neural networks for text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 333* (Austin, TX), 2267–2273.

Lashkarashvili, N., and Tsintsadze, M. (2022). Toxicity detection in online Georgian discussions. *Int. J. Inf. Manage. Data Insights* 2:100062. doi: 10.1016/j.jjimei.2022.100062

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Li, J., Cai, Y., Cai, Z., Leung, H., and Yang, K. (2017). "Wikipedia based short text classification method," in *Database Systems for Advanced Applications: DASFAA 2017 International Workshops: BDMS, BDQM, SeCoP, and DMMOOC, Suzhou, China, March 27-30, 2017.* Cham: Springer.

Linmei, H., Yang, T., Shi, C., Ji, H., and Li, X. (2019). "Heterogeneous graph attention networks for semi-supervised short text classification," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong: ACL), 4820–4829. doi: 10.18653/v1/D19-1488

Losada, D. E., Parapar, J., and Barreiro, A. (2017). Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems. *Inf. Proc. Manage.* 53, 1005–1025. doi: 10.1016/j.ipm.2017.04.005

Maurer, C., and Wiegmann, R. (2011). *Effectiveness of Advertising on Social Network Sites: A Case Study on Facebook. Information and Communication Technologies in Tourism 2011.* Cham: Springer, 485–498.

Memon, A. B., Bhutto, A., Luhana, K. K., Abro, A. H., and Meyer, K. (2021). Towards social networks integrated domain-specific business directories. *Quaid-E-Awam Univ. Res. J. Eng. Sci. Technol.* 19, 28–34. doi: 10.52584/QRJ.1901.04

Memon, A. B., and Meyer, K. (2015). "*CoDiT: an integrated business partner discovery tool over SNSs,*" in *Working Conference on Virtual Enterprises.* Cham: Springer.

Memon, A. B., and Meyer, K. (2020). Affordances of business pages on social networking sites: towards an integration model. *IJSMOC* 12, 21–41. doi: 10.4018/IJSMOC.2020070102

Memon, A. B., Zinke, C., and Meyer, K. (2017). "*A semantics-based approach for business categorization on social networking sites*," in *Working Conference on Virtual Enterprises*. Cham: Springer.

Meng, Y., Zhang, Y., Huang, J., Xiong, C., Ji, H., Zhang, C., et al. (2020). Text classification using label names only: a language model self-training approach. *arXiv [Preprint]*. arXiv:2010.07245. Available online at: https://arxiv.org/pdf/2010.07245.pdf

Miao, D., Duan, Q., Zhang, H., and Jiao, N. (2009). Rough set based hybrid algorithm for text classification. *Exp. Syst. Appl.* 36, 9168–9174. doi: 10.1016/j.eswa.2008.12.026

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. (2007). "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement* (New York, NY: ACM), 29–42.

Nasteski, V. (2017). An overview of the supervised machine learning methods. Horizons. *B* 4, 51–62. doi: 10.20544/HORIZONS.B.04.1.17.P05

Nguyen, T. H., Shirai, K., and Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Exp. Syst. Appl.* 42, 9603–9611. doi: 10.1016/j.eswa.2015.07.052

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). "Abusive language detection in online user content," in *Proceedings of the 25th International Conference on World Wide Web* (Geneva), 145–153.

Olshannikova, E., Olsson, T., Huhtamäki, J., and Kärkkäinen, H. (2017). Conceptualizing big social data. *J. Big Data* 4, 1–19. doi: 10.1186/s40537-017-0063-x

Otero, D., Martin-Rodilla, P., and Parapar, J. (2021). Building cultural heritage reference collections from social media through pooling strategies: the case of 2020's tensions over race and heritage. *ACM JOCCH* 15, 1–13. doi: 10.1145/3477604

Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. (2007). A design science research methodology for information systems research. *J. Manage. Inf. Syst.* 24, 45–77. doi: 10.2753/MIS0742-1222240302

Richter, M., Quasthoff, U., Hallsteinsdóttir, E., and Biemann, C. (2006). Exploiting the leipzig corpora collection. *Proc. IS-LTC* 27, 68–73.

Rogers, D., Preece, A., Innes, M., and Spasić, I. (2021). Real-time text classification of user-generated content on social media: systematic review. *IEEE Trans. Comput. Soc. Syst.* 9, 1154–1166. doi: 10.1109/TCSS.2021.3120138

Sahlgren, M., and Coaster, R. (2004). "Using bag-of-concepts to improve the performance of support vector machines in text categorization," in *20th International Conference on Computational Linguistics (COLING'04)* (Geneva: Association for Computational Linguistics), 487.

Salloum, S. A., Al-Emran, M., Monem, A. A., and Shaalan, K. (2017). A survey of text mining in social media: Facebook and Twitter perspectives. *Adv. Sci. Technol. Eng. Syst. J.* 2, 127–133. doi: 10.25046/aj020115

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Reading. New York, NY: Addison-Wesley,* 169.

Sanchez-Pi, N., Marti, L., and Garcia, A. C. B. (2016). Improving ontology-based text classification: An occupational health and security application. *J. Appl. Logic* 17, 48–58. doi: 10.1016/j.jal.2015.09.008

Schütze, H., Manning, C. D., and Raghavan, P. (2008). *Introduction to Information Retrieval, Vol. 39*. Cambridge, MA: Cambridge University Press Cambridge.

Sharma, H., and Kumar, S., and others (2016). A survey on decision tree algorithms of classification in data mining. *IJSR* 5, 2094–2097. doi: 10.21275/v5i4.NOV162954

Shi, Z., Gu, B., Popowich, F., and Sarkar, A. (2005). "Synonym-based query expansion and boosting-based re-ranking: A two-phase approach for genomic information retrieval," in *The Fourteenth Text REtrieval Conference (TREC 2005). NIST, Gaithersburg, MD.*

Shrestha, P. (2011). "Corpus-based methods for short text similarity," in *Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues* (Montpellier), 297.

Tabassum, S., Pereira, F. S. F., Fernandes, S., and Gama, J. (2018). Social network analysis: an overview. *Wiley Interdis. Rev. Data Mining Knowledge Disc.* 8:e1256. doi: 10.1002/widm.1256

Takeda, H., Veerkamp, P., and Yoshikawa, H. (1990). Modeling design process. *AI Magazine* 11:37.

Thangaraj, M., and Sivakami, M. (2018). Text classification techniques: a literature review. *Interdis. J. Inf. Knowledge Manage.* 13:117. doi: 10.28945/4066

Tian, T. (2020). Social big data: techniques and recent applications. *IJCSS* 14, 224–235.

Venable, J., Pries-Heje, J., and Baskerville, R. (2012). "A comprehensive framework for evaluation in design science research," in *Design Science Research in Information Systems. Advances in Theory and Practice: 7th International Conference, DESRIST 2012, Las Vegas, NV, USA, May 14-15, 2012*. Cham: Springer.

Venable, J., Pries-Heje, J., and Baskerville, R. (2016). FEDS: a framework for evaluation in design science research. *Eur. J. Inf. Syst.* 25, 77–89. doi: 10.1057/ejis.2014.36

Ville, B. (2013). Decision trees. *Wiley Interdis. Rev. Comput. Stat.* 5, 448–455. doi: 10.1002/wics.1278

Vongkusolkit, J., and Huang, Q. (2021). Situational awareness extraction: a comprehensive review of social media data classification during natural hazards. *Annal. GIS* 27, 5–28. doi: 10.1080/19475683.2020.1817146

Wang, B. K., Huang, Y. F, Yang, W. X., and Li, X. (2012). Short text classification based on strong feature thesaurus. *J. Zhejiang Univ. Sci.* 13, 649–659. doi: 10.1631/jzus.C1100373

Wang, P., Hu, J., Zeng, H. -J., Chen, L., and Chen, Z. (2007). "Improving text classification by using encyclopedia knowledge," in *Seventh IEEE international Conference on Data Mining (ICDM 2007)*. Piscataway, NJ: IEEE.

Wanichayapong, N., Pruthipunyaskul, W., Pattara-Atikom, W., and Chaovalit, P. (2011). "Social-based traffic information extraction and classification," in *2011 11th International Conference on ITS Telecommunications*. Piscataway, NJ: IEEE.

Wieringa, R. (2009). "Design science as nested problem solving," in *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology, DESRIST, Philadelphia* (New York, NY: ACM), 1–12.

Wilbur, W. J., and Sirotkin, K. (1992). The automatic identification of stop words. *J. Inf. Sci.* 18, 45–55. doi: 10.1177/016555159201800106

Xiang, G., Fan, B., Wang, L., Hong, J., and Rose, C. (2012). "Detecting offensive tweets via topical feature discovery over a large scale twitter corpus," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (New York, NY: ACM), 1980–1984. doi: 10.1145/2396761.2398556

Xie, Q., Dai, Z., Hovy, E., Luong, T., and Le, Q. (2020). Unsupervised data augmentation for consistency training. *Adv. Neur. Inf. Proc. Syst.* 33, 6256–6268.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (San Diego, CA: Association for Computational Linguistics), 1480–1489.

Zubiaga, A., Spina, D., and Martictor (2015). Real-time classification of twitter trends. *J. Assoc. Inf. Sci. Technol.* 66, 462–473. doi: 10.1002/asi.23186