frontiers in Computer Science

ORIGINAL RESEARCH
published: 07 January 2022
doi: 10.3389/fcomp.2021.808507

# Implicit Estimation of Paragraph Relevance From Eye Movements

*Michael Barz*[1,2]*, *Omair Shahzad Bhatti*[1] *and Daniel Sonntag*[1,2]

[1] German Research Center for Artificial Intelligence, Interactive Machine Learning Department, Saarbrücken, Germany,
[2] Applied Artificial Intelligence, Oldenburg University, Oldenburg, Germany

Eye movements were shown to be an effective source of implicit relevance feedback in constrained search and decision-making tasks. Recent research suggests that gaze-based features, extracted from scanpaths over short news articles (g-REL), can reveal the perceived relevance of read text with respect to a previously shown trigger question. In this work, we aim to confirm this finding and we investigate whether it generalizes to multi-paragraph documents from Wikipedia (Google Natural Questions) that require readers to scroll down to read the whole text. We conduct a user study ($n = 24$) in which participants read single- and multi-paragraph articles and rate their relevance at the paragraph level with respect to a trigger question. We model the perceived document relevance using machine learning and features from the literature as input. Our results confirm that eye movements can be used to effectively model the relevance of short news articles, in particular if we exclude difficult cases: documents which are on topic of the trigger questions but irrelevant. However, our results do not clearly show that the modeling approach generalizes to multi-paragraph document settings. We publish our dataset and our code for feature extraction under an open source license to enable future research in the field of gaze-based implicit relevance feedback.

**Keywords: implicit relevance feedback, reading analysis, machine learning, eye tracking, perceived paragraph relevance, eye movements and reading**

OPEN ACCESS

***Edited by:***
*Siyuan Chen,*
*University of New South Wales, Australia*

***Reviewed by:***
*Nora Castner,*
*University of Tübingen, Germany*
*Xi Wang,*
*ETH Zürich, Switzerland*

***Correspondence:**
*Michael Barz*
*michael.barz@dfki.de*

***Specialty section:***
*This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science*

***Received:*** *03 November 2021*
***Accepted:*** *12 December 2021*
***Published:*** *07 January 2022*

***Citation:***
*Barz M, Bhatti OS and Sonntag D (2022) Implicit Estimation of Paragraph Relevance From Eye Movements. Front. Comput. Sci. 3:808507. doi: 10.3389/fcomp.2021.808507*

## 1. INTRODUCTION

Searching for information on the web or in a knowledge base is pervasive. However, search queries to information retrieval systems seldom represent a user's information need precisely (Carpineto and Romano, 2012). At the same time, a growing number of available documents, sources, and media types further increase the required effort to satisfy an information need. Implicit relevance feedback, obtained from users' interaction signals, was proposed to improve information retrieval systems as an alternative to more accurate, but costly explicit feedback (Agichtein et al., 2006). Behavioral signals that were investigated in this regard include clickthrough data (Agichtein et al., 2006; Joachims et al., 2017), dwell time of (partial) documents (Buscher et al., 2009), mouse movements (Eickhoff et al., 2015; Akuma et al., 2016), and eye movements (Buscher et al., 2012). This data may originate from search logs, which can be used to tune the ranking model of a search engine offline, or from real-time interaction data to extend search queries during a search session or to identify relevant text passages. In this work, we aim at identifying relevant paragraphs using real-time eye tracking data as input.

Eye movements play an important role in information acquisition (Gwizdka and Dillon, 2020) and were shown to be an effective source of implicit relevance feedback in search (Buscher et al., 2008a) and decision-making (Feit et al., 2020). However, eye movements highly depend on the user characteristics, the task at hand, and the content visualization (Buchanan et al., 2017). Related approaches use eye tracking to infer the perceived relevance of text documents with respect to previously shown trigger questions (Salojarvi et al., 2003, 2004, 2005a; Buscher et al., 2008a; Loboda et al., 2011; Gwizdka, 2014a; Bhattacharya et al., 2020a,b), and to extend (Buscher et al., 2008b; Chen et al., 2015) or generate search queries (Hardoon et al., 2007; Ajanki et al., 2009). A common disadvantage of approaches for gaze-based relevance estimation is that they are tested using documents with constrained layouts and topics such as single sentences (Salojarvi et al., 2003, 2004, 2005a) or short news articles that fit on the screen at once (Buscher et al., 2008a; Loboda et al., 2011; Gwizdka, 2014a; Bhattacharya et al., 2020a,b). Hence, it is unclear whether related findings generalize to more realistic settings such as those that include Wikipedia-like web documents.

We investigate whether eye tracking can be used to infer the perceived relevance of read documents with respect to previously shown trigger questions in a less constrained setting. We include multi-paragraph documents that exceed the display size and require scrolling to read the whole text. For this, we conduct a user study with $n = 24$ participants in which participants read single- and multi-paragraph articles and rate their relevance at the paragraph level while their eye movements are recorded. Pairs of single paragraph documents and questions are taken from the g-REL corpus (Gwizdka, 2014a). Multi-paragraph documents with corresponding questions are selected from the Google Natural Questions (GoogleNQ) corpus (Kwiatkowski et al., 2019). We assemble a corresponding dataset, the *gazeRE* dataset, and make it available to the research community under an open source license via Github (see section 3.5). Using the gazeRE dataset, we aim for confirming the findings from the literature on short news articles and investigate whether they generalize to the multi-paragraph documents from Wikipedia. We model the perceived relevance using machine learning and the features from Bhattacharya et al. (2020a) as input.

## 2. RELATED WORK

Prior research addressed the question whether eye movements can be linked to the relevance of a read text and how this implicit feedback can be leveraged in information retrieval settings.

### 2.1. Relevance Estimation From Reading Behavior

One group of work addressed the question whether the relevance of a text with respect to a task or trigger questions can be modeled using the user's gaze. For instance, Salojarvi et al. (2003, 2004, 2005a) investigated whether eye tracking can be used to estimate the user's perceived relevance of a document. They used machine learning to predict the relevance using the eye movements from reading the document titles as input. The authors organized a related research challenge, which is described in Salojarvi et al. (2005b). Loboda et al. (2011) presented an approach for gaze-based estimation of sentence relevance using fixations to sentence-terminal words, i.e., words at the end of a sentence, as there is empirical evidence that these words are fixated longer on average. This is known as the sentence wrap-up effect, which is a manifestation of the integrative process in reading. Buscher et al. (2008a) investigated the relation between reading behavior and document relevance using eye tracking technology. They found that the ratio of skimming is higher in irrelevant documents and the ratio of continuous reading behavior is higher for relevant documents. Further, they introduced the concept of attentive documents that keep track of the perceived relevance based on eye movements (Buscher et al., 2012). Gwizdka (2014a,b) modeled the relation between eye movements and perceived document relevance and investigated the cognitive effort involved in the relevance judgement. They introduced the g-REL corpus, a collection of short news stories and corresponding questions, which they used for collecting ground-truth and eye tracking data. The authors could confirm the findings from Buscher et al. (2012) that relevant documents tend to be read continuously, while irrelevant documents are rather skimmed (Gwizdka, 2014a). Akuma et al. (2016) compared gaze-based relevance feedback with implicit relevance feedback from more common sensors such as mouse movements. They found a high correlation between both feedback options and a relationship between gaze-based features and the perceived document relevance. Li et al. (2018) investigated the reading behavior for relevant and irrelevant documents for factual and intellectual tasks. Based on data from a user study, they suggested a two-staged reading model for explaining the cognitive processes inherent in relevance judgements. Jacob et al. (2018) investigated whether eye movements can be used to infer the interest of a reader in a currently read article. Bhattacharya et al. (2020b) encoded fixations from participants' scanpaths over documents from the g-REL corpus and trained a convolutional neural network (CNN) with the perceived relevance as prediction target. This approach is limited to small texts of similar lengths. Further, they suggested novel features based on the convex hull of scanpath fixations to model the participants' perceived relevance (Bhattacharya et al., 2020a). In addition, they simulated the user interaction to investigate whether their approach can be used in real-time scenarios by cumulatively adding fixations of the scanpath and normalizing the convex hull features with the elapsed time of interaction. Other related approaches include, for instance, a generic approach to map gaze-signals to HTML documents at the word level (Hienert et al., 2019). Davari et al. (2020) use this tool to investigate the role of word fixations in query term prediction. Feit et al. (2020) modeled the user-perceived relevance of information views in a graphical user interface for decision-making. They showed room advertisements in a web-based interface via multiple viewports and asked users what information was perceived as relevant for their decision to book a room or not. In this paper, we investigate whether the perceived relevance can be estimated for paragraphs of long Wikipedia-like documents in contrast to

sentences or short articles. This requires to compensate for the scrolling activity, which may distort the gaze signal and fixation extraction, and to develop a method for effectively extracting consecutive gaze sequences to individual paragraphs.

## 2.2. Query Expansion Methods

Other work focused on generating or expanding search queries based on the user's gaze behavior. Miller and Agne (2005) presented a system that extracts relevant search keywords from short texts based on eye movements. Hardoon et al. (2007) and Ajanki et al. (2009) proposed methods for implicitly generating search queries from eye movements during an information retrieval task. The generated query is used to proactively retrieve relevant documents using content-based ranking algorithms. Buscher et al. (2008b) proposed a technique for automatic query expansion and re-ranking for document retrieval. They use relevance estimates to identify recently read paragraphs that are relevant to the user and, eventually, to reformulate the search query. Chen et al. (2015) presented a query expansion method based on eye tracking and topic modeling. They identified fixated terms and modeled the user's latent intent using the Latent Dirichlet Allocation (LDA) for topic modeling.

## 2.3. Factors That Influence Eye Movements

Buchanan et al. (2017) surveyed works in the field of gaze-based implicit relevance feedback. They identified several factors that might influence gaze patterns and, hence, should be considered when building gaze-enhanced information retrieval systems. Key factors include the task type, the task complexity, individual differences such as expertise, and the presentation of the search results. For instance, Cole et al. (2013) showed that "the user's level of domain knowledge can be inferred from their interactive search behaviors." Bhattacharya and Gwizdka (2018) modeled the knowledge-change while reading using gaze-based features: a high change in knowledge coincides with significant differences in the scan length and duration of reading sequences, and in the number of reading fixations. Gwizdka (2017) investigated the task-related differences in reading strategies between word search and relevance decisions during information search. Eickhoff et al. (2015) studied the relationship between the user's visual attention to tokens in a search engine result page (SERP) or document and the corresponding search query: users fixate terms, which are part of their current query more often and longer than others. Further, they found that the semantic proximity of the search query to the user's attention increases for different reformulation strategies such as specialization, generalization, and reformulation.

# 3. USER STUDY

We conduct a user study ($n = 24$) with the goal to collect eye movement data during relevance estimation tasks. The participants are asked to read documents of different lengths and to judge, per paragraph, whether it provides an answer to a previously shown trigger question. We use this data to model the relation between the recorded eye movement data and the perceived relevance using machine learning (see section 4).

## 3.1. Participants

For our study, we invited 26 students (15 female) with an average age of 27.19 years ($SD = 5.74$). Data from two participants had to be discarded, because they withdrew their participation. The remaining participants reported to have normal (11) or corrected to normal (13) vision of which 11 wore eyeglasses and 2 wore contact lenses. Ten of them participated in an eye tracking study before. The participants rated their language proficiency in English for reading texts as native (1), fluent (18), or worse (5). Each participant received 15 EUR as compensation.

## 3.2. Stimuli

The stimuli data used in our study are pairs of trigger questions and documents with one or multiple paragraphs (see **Figure 1**). We use a subset from the g-REL corpus (Gwizdka, 2014a) with single-paragraph documents that fit on one page and selected pairs from the Google Natural Questions (NQ) corpus, which includes multi-paragraph documents that require scrolling (Kwiatkowski et al., 2019). Both corpora include relevance annotations per paragraph to which we refer as system relevance.

### 3.2.1. g-REL Corpus

The g-REL corpus includes a set of 57 trigger questions and 19 short English news texts that fit on one page. Questions include, for instance, "Where is the headquarters of OPEC located?" and "What was Camp David originally named?". The news texts are either irrelevant, topically relevant, or relevant with respect to these questions: the corpus includes three questions per document. If a document is irrelevant, it is off-topic and does not contain an answer to the question. Topically relevant and relevant documents are on topic, but only the relevant texts contain an answer to the question. The original news texts were selected from the AQUAINT Corpus of English News Texts (Graff, 2002) as used in the TREC 2005 Question Answering track.[1] The questions and judgements (system relevance) from TREC data were further revised and tested by Michael Cole and Jacek Gwizdka. Prior results for this corpus have been published in, e.g., Gwizdka (2014a,b, 2017), Bhattacharya et al. (2020a,b). Like Bhattacharya et al. (2020a,b), we consider a binary relevance classification. Hence, the topically relevant document-question pairs are counted as irrelevant ones.

For our user study, we select a balanced subset of 12 distinct documents of which four are relevant, four are topical, and four are irrelevant with respect to the accompanying trigger question. We select two additional documents for the training phase of which one is relevant and one is topical. We select the news texts such that the length distribution is similar to the whole corpus. The mean number of tokens of the selected news texts is 170.5 ($SD = 14.211$). The mean number of tokens, if all documents were included, is 176.404 ($SD = 12.346$). We used a simple whitespace tokenizer, which segments each document into a list words, to determine the number of tokens in each document.

[1] https://trec.nist.gov/data/qa.html

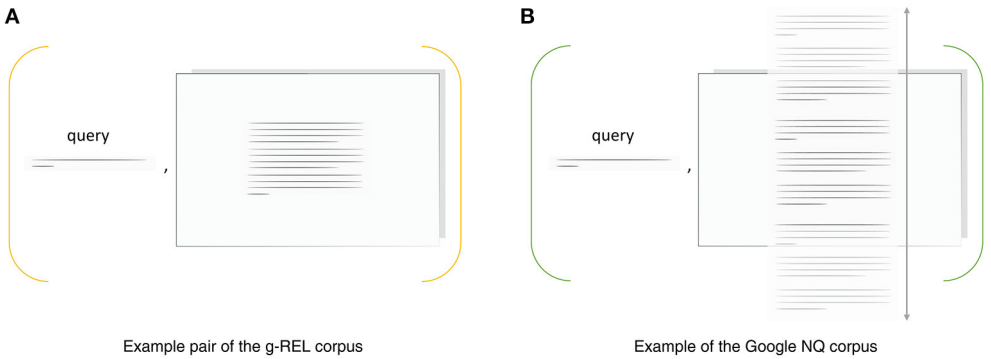


**FIGURE 1 |** We sample stimuli from **(A)** the g-REL corpus, which includes pairs of questions and short English news articles, and **(B)** from the Google NQ corpus, which includes pairs of questions and English Wikipedia articles.

#### 3.2.2. Google Natural Questions Corpus

The Natural Questions (NQ) corpus[2] by Google includes 307k pairs of questions and related English Wikipedia documents (Kwiatkowski et al., 2019). Example questions include "What is the temperature at bottom of ocean?" and "What sonar device let morse code messages be sent underwater from a submarine in 1915?". Each document includes multiple HTML containers such as paragraphs, lists, and tables. Each container that provides an answer to the accompanying question is listed as a *long answer*. We consider this container to be relevant (system relevance). In addition, the corpus provides a *short answer* annotation, if a short phrase exists within a container that fully answers the question. The Google NQ questions are longer and more natural compared to other question answering corpora including TREC 2005 and, hence, g-REL.

For our user study, we select a subset of 12 pairs of documents and questions (plus one for training) from the NQ training data using a set of filters followed by a manual selection. Our filter removes all documents that include at least one container different than a paragraph, because we focus on continuous texts in this work. Further, it selects documents that have exactly one long and one short answer. This means that all but one paragraph per document can be considered to be irrelevant. Also, it removes all documents that have very short (less than 20 tokens) or very long (greater than 200 tokens) documents. Finally, our filter selects all documents with five to seven paragraphs, which leaves 355 of the 307k pairs for manual selection. The manual selection is guided by two factors: the average number of tokens and the position of the relevant paragraph. The remaining documents have an average length of 420.083 ($SD = 54.468$) tokens, which approximately corresponds to two times the height of the display, i.e., participants need to scroll through the document to read all paragraphs. The position of relevant paragraphs is balanced: we select two documents with an answer at position $i$ with $i$ ranging from 0 to 5. On average, each paragraph contains 72.55 tokens.

[2]https://ai.google.com/research/NaturalQuestions

## 3.3. Tasks and Procedure

In the beginning of the study, each participant is asked to sign an informed consent form and to fill in a demography questionnaire. The remainder of the study is divided in two blocks, which follow the same pattern (see **Figure 2**). In each block, stimuli from one of the two corpora are presented (within-subjects design). The starting order is alternating to avoid ordering effects. In the beginning of each block, the experimenter provides block-specific *instructions* and asks the participant to calibrate the eye tracking device. Next, the participant completes a training phase to get familiar with the task, the user interface, and with characteristics of the stimuli from the current corpus. We include two training examples for g-REL and one for Google NQ. The participant is encouraged to ask questions about the system and the task in this phase. Subsequently, the participant completes the main phase of the block, which includes 12 stimuli of the respective corpus. After both blocks are finished, participants receive the compensation payment. The task of participants is to mark all paragraphs of a document as relevant that contain an answer to the previously shown trigger question (query). First, participants read the query and, then, navigate to the corresponding document, which is either a *news article* or a *wiki article*. There is no time constraint for reading the article. Next, participants move to the rating view which enables to enter a binary relevance estimate (perceived relevance) per paragraph. At this stage, the query and the text of the paragraph are available to the participant. For stimuli from the g-REL corpus, participants have to provide one relevance estimate (there is one paragraph). For stimuli from the Google NQ corpus, participants have to provide five to seven relevance estimates (depending on the number of paragraphs).

## 3.4. Apparatus

The study is conducted in a separate room of our lab. We use the Tobii 4C eye tracker[3], a non-intrusive remote eye tracker, which is attached to the lower bezel of a 27-inch screen. This monitor has a resolution of 2560 × 1440 pixels and the attached

[3]https://help.tobii.com/hc/en-us/articles/213414285-Specifications-for-the-Tobii-Eye-Tracker-4C

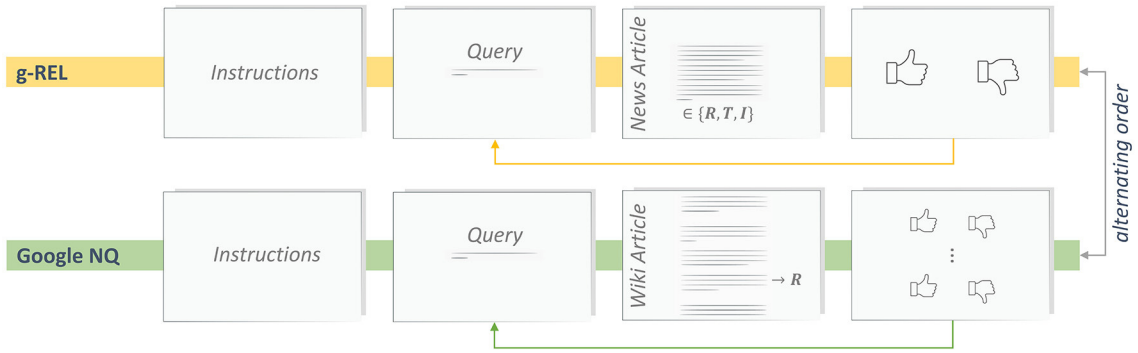


**FIGURE 2 |** Procedure of our user study with one block of tasks per corpus of stimuli: g-REL and Google NQ.

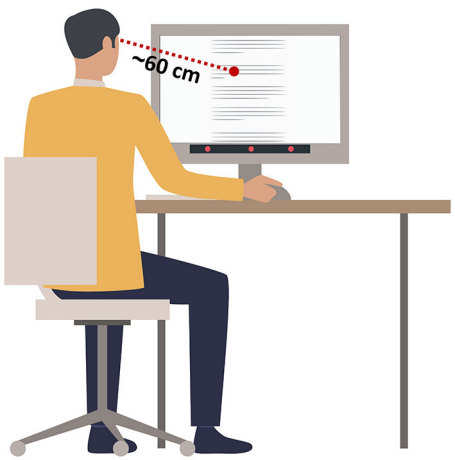


**FIGURE 3 |** Setup of our user study: A user is seated approximately 60 cm from a 27 inch display with the remote eye tracker mounted at its lower bezel.

eye tracker collects the gaze data with a sampling rate of 90 Hz. The monitor and eye tracker are connected to an experimenter laptop running the study software and a monitoring tool. The participants are seated approximately 60 cm in front of the connected display (see **Figure 3**). A mouse is provided to scroll through documents, to navigate between views, and to rate each paragraph for its relevancy. The text-based stimuli are displayed in black, 38-points Roboto font[4] on a white background. Before the user starts executing the tasks, we perform a calibration using the built-in 9-point calibration of the eye tracker. During the calibration process, the user is asked to look at calibration dots on the connected display until they vanish. We use the multisensor-pipeline (Barz et al., 2021), our Python-based framework for building stream processing pipelines, to implement the study software that is responsible to show the stimuli and record the interaction signals according to our experiment procedure.

## 3.5. gazeRE Dataset

We assembled the stimuli and the recorded interaction signals into the *gazeRE* dataset, a dataset for **gaze**-based **R**elevance **E**stimation. It includes relevance ratings (perceived relevance) from 24 participants for 12 stimuli from the g-REL corpus and 12 stimuli from the Google NQ corpus. Also, it includes participants' eye movements per document in terms of 2D gaze coordinates on the connected display. We use the gazeRE dataset for modeling the perceived relevance based on eye tracking in this work and make it publicly available under an open source license on GitHub.[5]

### 3.5.1. Processing of Eye Tracking Data

The gaze data included in the gazeRE dataset is preprocessed and cleaned. We correct irregular timestamps caused by transferring the gaze signal to our study software by resampling the signal with a fixed sampling rate of 83 Hz. Further, we use the gap_fill algorithm, similar to Olsen (2012), which linearly interpolates the gaze signal to close small gaps between valid gaze points, which may occur due to a loss of tracking. In addition, we use the Dispersion-Threshold Identification (I-DT) algorithm to detect fixation events (Salvucci and Goldberg, 2000).

### 3.5.2. Dataset Format

The gazeRE dataset includes synchronized time-series data per document and user. Each record includes a column for timestamps, gaze coordinates (x and y), a fixation ID, if the gaze point belongs to a fixations, the scroll position, and the ID of the paragraph that is hit by the current point of gaze. The origin of the gaze and fixation coordinates is the lower-left corner of the display $(0, 0)$ while $(2560, 1440)$ denotes the upper-right corner. The scroll position reflects the status of the scrollbar and lies between 0 and 1. The position is 1, if the document head is visible, or the document is not scrollable. It is 0, if the tail of the document is visible. We provide the perceived relevance per document and user: *True* is used for positive ratings, i.e., if a paragraph was perceived as relevant, *False* represents irrelevant ratings.

[4]https://fonts.google.com/specimen/Roboto (accessed February 16, 2021).

[5]https://github.com/DFKI-Interactive-Machine-Learning/gazeRE-dataset

#### 3.5.3. Descriptive Statistics

We report descriptive statistics and agreement statistics of the relevance ratings in our dataset. We use Fleiss' $\kappa$ to determine, if there was an agreement in our participants' judgement on whether paragraphs are relevant with respect to a trigger question. If the agreement among participants is low, the rating task might have been too difficult or participants might have given inadequate ratings. Further, we compute Cohen's $\kappa$ to determine the level of agreement between each participant's relevance rating (perceived relevance) and the ground-truth relevance (system relevance). We report the mean agreement over all participants. We expect that the ratings of our participants moderately differ from the system relevance, similar to the findings in Bhattacharya et al. (2020a). For the g-REL corpus, we include a total of 288 trials, i.e., eye movements and a corresponding relevance estimate per paragraph (see **Figure 1**). The 12 different documents include 4 relevant paragraphs (system relevance), while one document corresponds to one paragraph. On average, the participants rated 4.46 ($SD = 1.04$) paragraphs as relevant: they perceived 107 (37%) as relevant and 181 (63%) as irrelevant. Fleiss' $\kappa$ reveals a good agreement for perceived relevance ratings with $\kappa = 0.641$. The mean of Cohen's $\kappa$ of 0.769 ($SD = 0.197$) indicates a substantial agreement between participant and ground-truth relevance ratings. We obtained a total of $1,680$ trials using the Google NQ corpus. The 12 stimuli include 12 relevant paragraphs out of 70. On average, the participants rated 18.75 ($SD = 4.361$) paragraphs as relevant: they perceived 450 (27%) as relevant and $1,230$ (73%) as irrelevant. Fleiss' $\kappa$ reveals a moderate agreement for perceived relevance ratings with $\kappa = 0.576$. Also, the mean of Cohen's $\kappa$ of 0.594 ($SD = 0.126$) indicates a moderate agreement between the perceived and the system relevance.

## 4. GAZE-BASED RELEVANCE ESTIMATION

We investigate different methods for predicting the perceived relevance of a read paragraph based on a user's eye movements. We consider the relevance prediction as a binary classification problem because each paragraph could be marked as either relevant or irrelevant in our user study. Each classification model takes a user's eye movements from reading a paragraph as input to predict the perceived relevance for this paragraph. The explicit user ratings are used as ground truth. In the following, we describe our method for extracting gaze-based features at the paragraph level, we depict our procedure for model training and evaluation, and we report the results based on the gazeRE dataset.

### 4.1. Extraction of Gaze-Based Features

To encode the eye movements of a user for a certain paragraph $p$, we have to extract coherent gaze sequences that lie within the paragraph area. A user might visit a paragraph multiple times during the relevance judgement process. We refer to these gaze sequences as visits $v_p^i \in V_p$ where $i$ indicates the order of visits. We implement an algorithm that extracts all visits to a paragraph with a minimum length while ignoring short gaps. It identifies consecutive gaze samples that lie within the area of the given paragraph and groups them into a visit instance each. As long as there is a pair of two subsequent visits with a gap shorter than 0.2 s, these are merged. Afterwards, all visits that satisfy a minimum length of 3 s are returned as a list. We found that this duration ensures that at least 3 fixations are contained in each visit, which is required to compute the convex hull features.

We use the longest visit per paragraph $v_p^*$ for encoding the eye movements.

To encode eye movements, we implement a set of 17 features that was successfully used to model the perceived relevance of short news articles in Bhattacharya et al. (2020a). This requires to select one visit or to merge them. We decided to use the longest visit under the assumption that the largest consecutive sequence of gaze points has the highest likelihood to capture indicative eye movements. Our feature extraction function $f$ returns a vector of size 17 per visit: $f(v) \rightarrow \mathbb{R}^{17}$. Four of these features are based on fixation events, eight are based on saccadic movements, and five are based on the area spanned by all fixations. **Table 1** provides an overview of all features and describes how they are computed. Some features are normalized by a width factor `w` or a height factor `h`. In Bhattacharya et al. (2020a), these correspond to the display width and height, respectively. We set `w` and `h` to the width and height of the current paragraph, because the display size does not respect the different paragraph sizes and the scrolling behavior.

The absolute reading time of a visit (`scan_time`) is used to compute velocity-based or time-normalized features. The `hull_area`, i.e., the area of the convex hull around all fixations, is used to compute two area-based features.

### 4.2. Model Training and Evaluation

We build and compare several machine learning models that take an encoded paragraph visit $v_p^*$ as input and yield a binary relevance estimate as output. The models are implemented using the scikit-learn machine learning framework (Pedregosa et al., 2011). Model training and testing is done using our gazeRE dataset, which includes eye movements and relevance estimates for documents from the g-REL corpus and from the Google NQ corpus. We refer to these partitions as g-REL data and Google NQ data.

#### 4.2.1. Model Training Conditions

We largely replicate the conditions for model training and evaluation from Bhattacharya et al. (2020a) because we aim for confirming their findings: we group all visits $v \in V^*$ by their relevance rating into three subsets, train each model on 80% of the data of each subset, and evaluate it on the remaining 20% of the data. The grouping yields an *agree* subset, a *topical* subset, and the complete data denoted as *all*. **Table 2** depicts how many relevant and irrelevant samples are included in our dataset per subset. The *agree* subset includes all visits for which the perceived relevance rating agrees with the system relevance. All visits to topical articles, i.e., visits to on-topic articles that are irrelevant, are excluded as well. The *topical* subset includes visits to topical articles only, which are expected to be more difficult to classify. This subset is empty for the Google NQ corpus, because its paragraphs are marked as either relevant or irrelevant. We report

**TABLE 1 |** Overview of the 17 features adapted from Bhattacharya et al. (2020a) based on fixation events, saccadic eye movements, and the scanned area, which we use to encode paragraph visits.

| | Feature | Description |
|---|---|---|
| *fixation-based* | `fixn_n` | Number of fixations |
| | `fixn_dur_sum` | Sum of fixation durations |
| | `fixn_dur_avg` | Mean of fixation durations |
| | `fixn_dur_sd` | Standard deviation of fixation durations |
| *saccade-based* | `scan_dist_h` | Sum of horizontal amplitudes of all saccades, normalized by a factor `w` |
| | `scan_dist_v` | Sum of vertical amplitudes of all saccades, normalized by a factor `h` |
| | `scan_dist_euclid` | Sum of Euclidean distances of normalized amplitudes of all saccades |
| | `scan_hv_ratio` | Ratio of horizontal to vertical amplitudes: `scan_dist_h`/`scan_dist_v` |
| | `avg_sacc_length` | Average saccade amplitude: `scan_dist_euclid`/(`fixn_n` − 1) |
| | `scan_speed_h` | Horizontal saccade velocity: `scan_dist_h`/`scan_time` |
| | `scan_speed_v` | Vertical saccade velocity: `scan_dist_v`/`scan_time` |
| | `scan_speed` | Saccade velocity: `scan_dist_euclid`/`scan_time` |
| *area-based* | `box_area` | Area spanned by summed saccade amplitudes: `scan_dist_h` ∗ `scan_dist_v` |
| | `box_area_per_time` | The `box_area` normalized by the scan time: `box_area`/`scan_time` |
| | `fixns_per_box_area` | Number of fixations per scanned area: `fixn_n`/`box_area` |
| | `hull_area_per_time` | The `hull_area` normalized by the scan time: `hull_area`/`scan_time` |
| | `fixns_per_hull_area` | Number of fixations per convex hull area: `fixn_n`/`hull_area` |

**TABLE 2 |** Number of samples in our dataset per corpus and subset.

| Corpus | Subset | Relevant | Irrelevant | Total |
|---|---|---|---|---|
| g-REL | *agree* | 86 (48%) | 95 (52%) | 181 (63%) |
| | *topical* | 20 (20%) | 76 (80%) | 96 (33%) |
| | ***all*** | **107 (37%)** | **181 (63%)** | **288 (100%)** |
| Google NQ | *agree* | 248 (17%) | 1190 (83%) | 1438 (86%) |
| | ***all*** | **450 (27%)** | **1,230 (73%)** | **1,680 (100%)** |

*The topical subset includes samples for irrelevant paragraphs that are on topic of the trigger questions. The agree subset includes samples for which the participant's relevance rating matches with the system relevance and which is not in topical. Each trial corresponds to one paragraph that was either perceived as relevant or irrelevant.*

the model performance metrics averaged over 10 random train-test splits to estimate the generalization performance. We use the `train_test_split()` function of scikit-learn to split the visits in a stratified fashion with prior shuffling.

#### 4.2.2. Metrics

We include the same metrics than Bhattacharya et al. (2020a): the F1 score, i.e., the harmonic mean of precision and recall, the area under curve of the receiver operator characteristic (ROC AUC), and the balanced accuracy. In addition, we report the true positive rate (TPR) and the false positive rate (FPR), which allow us to estimate the suitability of our models for building adaptive user interfaces similar to Feit et al. (2020).

#### 4.2.3. Model Configurations

We consider the random forest classifier of scikit-learn with default parameters (`n_estimators` = 100) as our baseline model (RF), which turned out to work well in Bhattacharya et al. (2020a). In addition, we investigate the effect of using two pre-processing steps with either a random forest classifier (RF*) or a support vector classifier (SVC*) with default parameters (`kernel` = `"rbf"`, `C` = 1) in an estimator pipeline. First, we apply the oversampling technique SMOTE Chawla et al. (2002) from the imbalanced-learn package Lemaitre et al. (2017) because visits to relevant paragraphs are underrepresented in our dataset (see **Table 2**). Second, we apply a standard feature scaling method that removes the mean and scales features to unit variance. We train separate models for g-REL data and Google NQ data.

#### 4.2.4. Hypotheses

We hypothesize that our models can effectively estimate the perceived relevance of short news articles as shown in Bhattacharya et al. (2020a), but using our newly assembled gazeRE dataset (H1). Confirming this hypothesis would also serve as a validation of our dataset. Further, we assume that the visit-based scanpath encoding enables the prediction of a participants' perceived relevance for individual paragraphs of long Wikipedia articles. In particular, if the participant must scroll through the document to read all contents (H2).

### 4.3. Results

We compare the performance of three models in predicting a user's perceived relevance using our gazeRE dataset, which is based on documents of the g-REL and the Google NQ corpus. The performance scores for each model and subset are shown in **Table 3** (g-REL) and **Table 4** (Google NQ). For the g-REL data, we observe the best performance for the *agree* subset. Models trained on the *topical* subset achieve the worst results. Models for the *all* subset, which includes both other subsets, rank second. Across all subsets, the SVC* model performs best, or close to best, for most metrics. For the *topical* subset, the RF model without over-sampling and feature scaling achieves better ROC AUC and

**TABLE 3 |** Scores for all relevance prediction models trained and evaluated with data collected based on the g-REL corpus.

| | Model | F1 Score | ROC AUC | Balanced accuracy | TPR | FPR |
|---|---|---|---|---|---|---|
| *agree* | RF | 0.674 | 0.748 | 0.680 | 0.694 | 0.333 |
| | RF* | 0.677 | 0.747 | **0.689** | 0.688 | **0.317** |
| | SVC* | **0.702** | **0.787** | 0.683 | **0.782** | 0.417 |
| *topical* | RF | 0.119 | **0.546** | **0.527** | 0.100 | **0.047** |
| | RF* | 0.247 | 0.528 | 0.518 | 0.250 | 0.213 |
| | SVC* | **0.270** | 0.460 | 0.509 | **0.325** | 0.307 |
| *all* | RF | 0.458 | 0.650 | 0.594 | 0.405 | **0.217** |
| | RF* | 0.495 | 0.652 | 0.594 | 0.505 | 0.317 |
| | SVC* | **0.506** | **0.652** | **0.605** | **0.510** | 0.300 |

**TABLE 4 |** Scores for all relevance prediction models trained and evaluated with data collected based on the Google NQ corpus.

| | Model | F1 Score | ROC AUC | Balanced accuracy | TPR | FPR |
|---|---|---|---|---|---|---|
| *agree* | RF | 0.052 | 0.54 | 0.502 | 0.03 | **0.027** |
| | RF* | 0.246 | 0.543 | **0.543** | 0.278 | 0.229 |
| | SVC* | **0.297** | **0.563** | 0.54 | **0.467** | 0.388 |
| *all* | RF | 0.189 | 0.552 | 0.517 | 0.129 | **0.095** |
| | RF* | 0.331 | 0.552 | 0.527 | 0.343 | 0.289 |
| | SVC* | **0.428** | **0.596** | **0.57** | **0.552** | 0.412 |

FPR scores. However, we observe a very low TPR and F1 score in this case. For the Google NQ data, models trained on the *all* subset rank best compared to their counterpart trained on the *agree* subset. Similar to our experiment on the g-REL data, the SVC* model performs best, or close to best, for both subsets. Also, the RF model achieves the best FPR score, but the worst TPR and F1 scores.

## 5. DISCUSSION

The results of our machine learning experiment for short news articles (g-REL data) are similar to those in Bhattacharya et al. (2020a) (see **Table 3**). Our results indicate that we can effectively predict the perceived relevance for the *agree* subset, i.e., if the user's relevance rating agrees with the actual relevance of a paragraph and if irrelevant articles are not on topic. The *topical* trials are most difficult to classify: our models fail in differentiating between relevant and irrelevant paragraphs if they are on topic. Including *all* samples for training, our models perform better than chance with an F1 score greater than 0.5. The best-performing model pipeline, on average, is SVC*, a support vector classifier with over-sampling and feature scaling. Bhattacharya et al. (2020a) reported results for the RF model based on the original g-REL corpus using the same features for training, but with data from other participants. For the *agree* subset, their best model achieved an F1 score of 0.82, an ROC AUC of 0.92, and a balanced accuracy of 0.84. For the *topical* subset, they observed an F1 score of 0.3, an ROC AUC of 0.77, and a balanced accuracy of 0.59. Using *all* data samples results in an F1 score of 0.65, an ROC AUC of 0.85, and a balanced accuracy of 0.73. Even though we observed worse results per subset, we found the same overall pattern: the best performance is observed for models trained on the *agree* subset, followed by models for the *all* subset, and model for the *topical* subset rank last. This similarity is a good indicator for the validity of our gazeRE dataset and, eventually, it suggests that we may confirm our hypothesis H1. The differences in model performance may have several reasons. For instance, it is likely that the higher amount of training data in Bhattacharya et al. (2020a) yields better models. They used 3355 trials from 48 participants compared to 288 trials from 24 participants in our experiment. Further, our user study was conducted at a University in Germany with participants being, besides one, non-native English speakers, while the studies reported in Bhattacharya et al. (2020a) were conducted at two universities in the Unites States and predominantly included native English speakers. This may lead to a higher degree of variance in eye movements from our study. Another aspect may be that we used another eye tracking device and, hence, the data quality and pre-processing steps likely differ.

Using the Google NQ data in our machine learning experiment, we observe better scores when training on *all* data than when training on the *agree* subset only (see **Table 4**). However, the best-performing model, which is also the SVC* model, achieves F1 scores less than 0.5 in both cases although we have access to a higher number of training samples (see **Table 2**). The area under the ROC curve indicates classification performances better than chance, but we do not see enough evidence to confirm our hypothesis H2. A potential reason for the low performance might be that irrelevant paragraphs in fact belong to the same Wikipedia article than the relevant ones: the *agree* subset is rather a *topical* subset for which all user ratings agree with the system relevance. This would explain why models for the *agree* subset perform worse than models trained on *all* data. Also, the individual paragraphs in the Google NQ corpus are smaller than the ones in the g-REL corpus. This means that we aggregate less information per scanpath, which may deteriorate the model performance. Further, having multiple paragraphs allows the participants to revisit paragraphs. As we decided to encode the longest visit to a paragraph, we may miss indicative gaze patterns from another visit, which would have a negative impact on model training. In addition, the gaze estimation error inherent in eye tracking (Cerrolaza et al., 2012) may lead to a higher number of incorrect gaze-to-paragraph mappings: gaze-based interfaces should be aware of this error and incorporate it in the interaction design (Feit et al., 2017; Barz et al., 2018).

### 5.1. Feature Importance

We use 17 features as input to model the perceived paragraph relevance. In the following, we assess the importance of individual features to our best-performing model, the SVC* model. We use the permutation feature importance[6] method of the scikit-learn package (Pedregosa et al., 2011) to estimate feature importance, because SVCs with an rbf kernel do not

[6]https://scikit-learn.org/stable/modules/permutation_importance.html (accessed on Dec 2nd, 2021).

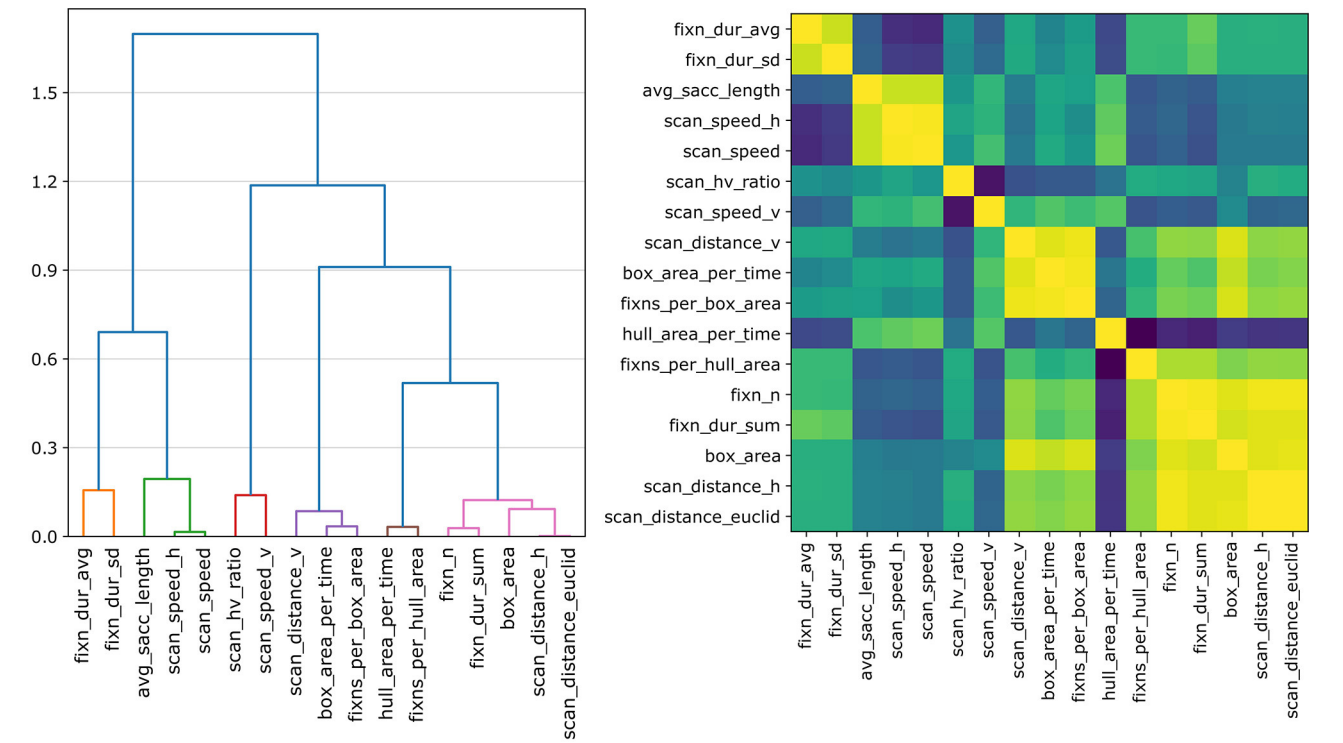


**FIGURE 4 |** The heatmap **(right)** shows the pairwise Spearman rank-order correlations for our features using all samples of the g-REL data. The dendrogram **(left)** shows feature groupings based on their correlation-based distances. Setting the distance threshold to $t = 0.3$ yields six feature clusters (see colored leafs).

allow direct feature analysis. This method randomly shuffles the values of one feature at a time and investigates the impact on the model performance. The loss in model performance reflects the dependency of the model on this feature. We report the mean loss in the f1 score from 30 repetitions per feature as importance measure. We analyze the feature importance for the *all* and *agree* subsets of the g-REL corpus only, because we observed f1 scores lower than 0.5 for all other conditions. The importance is reported on the training and test set of a single train-test split (80/20 split). We include both because features that are important on the training data but not on the test data might cause the model to overfit. The f1 test scores are 0.714 for the *agree* subset and 0.682 for *all* samples. However, this method might return misleading values if two features correlate. A model would still have access to nearly the same amount of information, if one feature was permuted but could be represented by another one. Hence, we perform a hierarchical clustering on the feature's Spearman rank-order correlations and use one feature per cluster to asses its importance.[7] The pairwise correlations and a grouping of our features based on correlation-based distances are visualized in **Figure 4** (*all* samples of the g-REL data). We set the distance threshold to $t = 0.3$ for the feature importance analysis for which we obtain six feature clusters as indicated by the colored leafs of the dendrogram. We obtain the same feature clusters for the *agree* subset and for both subsets of the Google NQ data. Using one feature per cluster to train and evaluate the SVC* model, we observe a drop in f1 scores of 0.015 for the *all* subset and no decline for the *agree* subset. These representative features include `fixn_dur_avg`, `scan_speed_h`, `scan_speed_v`, `scan_distance_v`, `scan_distance_h`, and `hull_area_per_time`. We remain at $t = 0.3$ because higher thresholds lead to substantially lower f1 scores and to differences in the resulting feature clusters between subsets and corpora.

The importance of feature clusters is visualized in **Figure 5**. For the *all* subset, we observe f1 losses ranging from 0.065 for `scan_speed_v` and 0.142 for `scan_distance_h` for the test set. For the train set, we observe slightly lower losses but the same importance ranking. Eventually, the features `scan_distance_h` and `hull_area_per_time` are most important when using *all* samples. For the *agree* subset, `hull_area_per_time` is by far the most important feature with an f1 loss of 0.162 on the train set and 0.137 of the test set. The features `scan_distance_v` and `fixn_dur_avg` are somewhat important with losses of 0.036 and 0.032. For `scan_speed_h`, we observe a higher importance on the train set (0.057) than on the test set (0.01), which may indicate that this feature causes the model to overfit to the training data. Overall, the `hull_area_per_time` feature introduced by

[7]We follow the scikit-learn manual for handling multicollinearity: https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance_multicollinear.html (accessed on Dec 2nd, 2021).

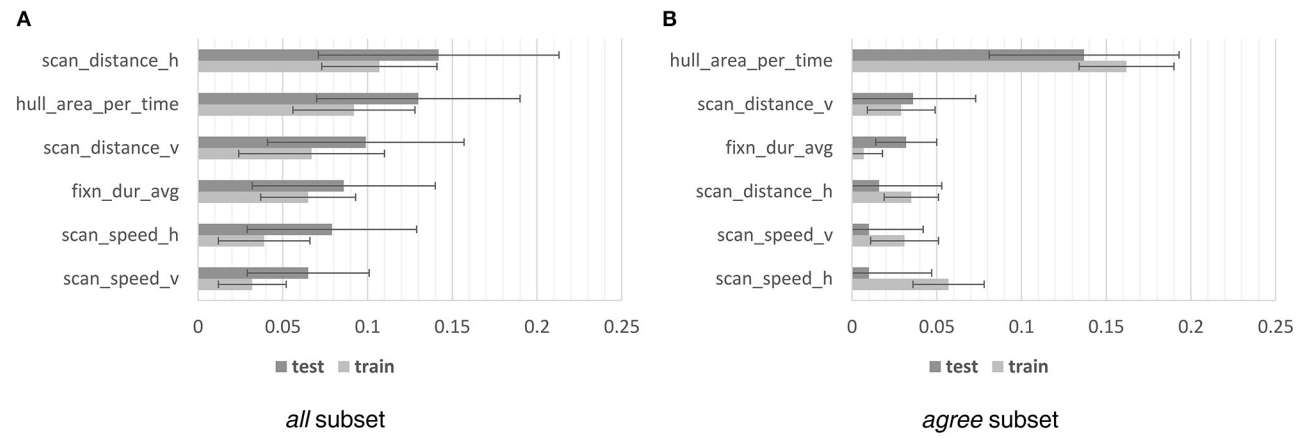


**FIGURE 5 |** Permutation feature importance in terms of the f1 loss for the *all* subset **(A)** and the *agree* subset **(B)** of the g-REL data. We show the mean loss in f1 scores from 30 permutation iterations ± the standard deviation. We include the feature importance estimates for the train and test split.

Bhattacharya et al. (2020a) is of high importance for modeling the perceived paragraph relevance and stable when including *topical* samples and samples for which the user rating disagrees with the ground truth. The remaining five features are important when including all samples, in particular the `scan_distance_h`. This result suggests that, in a first stage, these five features could be used to identify *topical* (irrelevant) samples and, in a second stage, the `hull_area_per_time` can predict paragraphs perceived as relevant among the remaining, non-topical samples.

## 5.2. Application to Adaptive User Interfaces

Our relevance estimation method can enable the development of adaptive user interfaces (UIs) that emphasize relevant contents or suppress irrelevant ones similar to Feit et al. (2020). Over time, their system detects relevant and irrelevant elements of a UI that shows different records of flat advertisements: a certain UI element always shows the same type of information, which depends on the currently viewed flat record. Our use case differs in that we want to highlight relevant text passages of a document or hide irrelevant ones. Adaptations may be based on perceived relevance estimates from recent eye movements and could, e.g., ease revisiting of relevant paragraphs in a document by immediately highlighting them or by hiding irrelevant passages. Alternatively, collecting relevant and irrelevant text passages in the pass of a search session may allow an adaptive UI to properly format text passages of documents hitherto unseen by the user. An adaptation method requires a precise recognition of relevant (true positive) or irrelevant (true negative) paragraphs to emphasize or suppress them, respectively. Misclassifications would lead to incorrect adjustments and subsequently to usability problems. Emphasizing irrelevant content (false positive) or suppressing relevant content (false negative) is likely to have a stronger negative impact on the user interaction than failing to suppress irrelevant content or to highlight a relevant one (Feit et al., 2020). To avoid strong negative impacts, adjustments by accentuation require a relevance model with a low false positive rate (FPR) and adjustments by suppression require a model with a high true positive rate (TPR), i.e., with a low number of false negatives. Depending on the type of adjustment, the TPR and FPR could be traded off against each other by using different decision thresholds. We show possible trade offs for our SVC* models using ROC curves. One model is trained on *all* g-REL data and one on *all* Google NQ data (see **Figure 6**). We do not consider other subsets for realistic application scenarios, because we would not be able to determine whether a user agreed with the actual (system) relevance of a paragraph or whether a text passage was on topic but irrelevant (*topical*). This differentiation, which is aligned to the work in Bhattacharya et al. (2020a), requires prior knowledge about the paragraphs and was meant to identify *topical* samples as being the most challenging cases for classification algorithms. Analogous to Feit et al. (2020), the shaded areas in our ROC plots in **Figure 6** indicate acceptable true and false positive rates for emphasizing or suppressing contents. For g-REL data, the ROC curve of the SVC* model hits the *emphasize* area, which indicates that it could be used to emphasize short news articles that were perceived as relevant, if the decision threshold is tuned accordingly. However, many relevant contents would be missed, as indicated by the low true positive rate (recall). Also, the shaded areas reveal that our models are not suitable for other kinds of UI adjustments.

# 6. CONCLUSION

In this work, we investigated whether we can confirm the findings from Bhattacharya et al. (2020a) that gaze-based features can be used to estimate the perceived relevance of short news articles read by a user. Further, we investigated whether the approach can be applied to multi-paragraph documents that require the user to scroll down to see all text passages. For this, we conducted a user study with $n = 24$ participants who read documents from two corpora, one including short news articles and one including longer Wikipedia articles in English, and rated their relevance at the paragraph-level with respect to a previously shown trigger question. We used

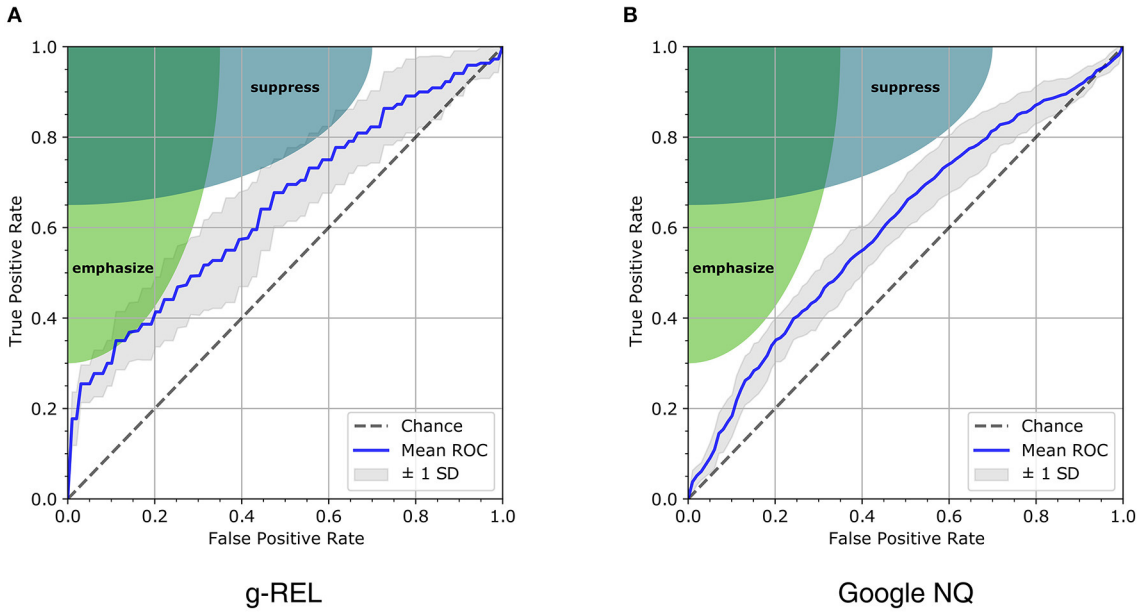


**FIGURE 6 |** Receiver Operator Characteristic curves for the SVC* relevance prediction models trained and tested using *all* samples from the g-REL data **(A)** and the Google NQ data **(B)**. The curves are averaged over the 10 training and test cycles; the gray area indicates the standard deviation (± 1 SD). The elliptic areas indicate acceptable true and false positive for emphasizing or suppressing contents based on Feit et al. (2020).

this data to train and evaluate machine learning models that predict the perceived relevance at the paragraph-level using the user's eye movements as input. Our results showed that, even though we achieved lower model performance scores than Bhattacharya et al. (2020a), we could replicate their findings under the same experiment conditions: eye movements are an effective source for estimating the perceived relevance of short news articles, if we leave out articles that are on topic but irrelevant. However, we could not clearly show that the approach generalizes to multi-paragraph documents. In both cases, the best model performance was observed when using over-sampling and feature scaling on the training data and a support vector classifier with an RBF kernel for classification. Future investigations should aim to overcome the limited estimation performances. A potential solution could be to use higher-level features such as the *thorough reading ratio*, i.e., the ratio of read and skimmed text lengths (Buscher et al., 2012), or the *refixation count*, i.e., the number of re-visits to a certain paragraph (Feit et al., 2020). Another solution could be found in using scanpath encodings based deep learning Castner et al. (2020); Bhattacharya et al. (2020b). We envision the gaze-based relevance detection to be a part of future adaptive UIs that leverage multiple sensors for behavioral signal processing and analysis Oviatt et al. (2018); Barz et al. (2020a,b). We published our new gazeRE dataset and our code for feature extraction under an open source license on Github to enable other researchers to replicate our approach and to implement and evaluate novel methods in the domain of gaze-based implicit relevance feedback.

## 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/DFKI-Interactive-Machine-Learning/gazeRE-dataset.

## 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

MB, OB, and DS contributed to conception and design of the study. MB performed the statistical analysis and the machine learning experiment and wrote the first draft of the manuscript. OB conducted the study and processed the dataset and wrote sections of the manuscript. MB and DS acquired the funding for this research. All authors contributed to manuscript revision, read, and approved the submitted version.

## FUNDING

This work was funded by the German Federal Ministry of Education and Research (BMBF) under grant number 01JD1811C (GeAR) and in the Software Campus project SciBot.

## REFERENCES


Agichtein, E., Brill, E., and Dumais, S. (2006). "Improving web search ranking by incorporating user behavior information," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '06* (New York, NY: Association for Computing Machinery), 19–26. doi: 10.1145/1148170.1148177

Ajanki, A., Hardoon, D. R., Kaski, S., Puolamaki, K., and Shawe-Taylor, J. (2009). Can eyes reveal interest? Implicit queries from gaze patterns. *User Model. User Adapt. Interact.* 19, 307–339. doi: 10.1007/s11257-009-9066-4

Akuma, S., Iqbal, R., Jayne, C., and Doctor, F. (2016). Comparative analysis of relevance feedback methods based on two user studies. *Comput. Hum. Behav.* 60, 138–146. doi: 10.1016/j.chb.2016.02.064

Barz, M., Altmeyer, K., Malone, S., Lauer, L., and Sonntag, D. (2020a). "Digital pen features predict task difficulty and user performance of cognitive tests," in *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2020* (Genoa: ACM), 23–32. doi: 10.1145/3340631.3394839

Barz, M., Bhatti, O. S., Laers, B., Prange, A., and Sonntag, D. (2021). "Multisensor-pipeline: a lightweight, flexible, and extensible framework for building multimodal-multisensor interfaces," in *Companion Publication of the 2021 International Conference on Multimodal Interaction, ICMI '21 Companion* (Montreal, QC: ACM).

Barz, M., Daiber, F., Sonntag, D., and Bulling, A. (2018). "Error-aware gaze-based interfaces for robust mobile gaze interaction," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, ETRA 2018*, eds B. Sharif and K. Krejtz (Warsaw: ACM), 24:1–24:10. doi: 10.1145/3204493.3204536

Barz, M., Stauden, S., and Sonntag, D. (2020b). "Visual search target inference in natural interaction settings with machine learning," in *ACM Symposium on Eye Tracking Research and Applications, ETRA '20*, eds A. Bulling, A. Huckauf, E. Jain, R. Radach, and D. Weiskopf (Stuttgart: Association for Computing Machinery), 1–8. doi: 10.1145/3379155.3391314

Bhattacharya, N., and Gwizdka, J. (2018). "Relating eye-tracking measures with changes in knowledge on search tasks," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, ETRA '18* (New York, NY: Association for Computing Machinery). doi: 10.1145/3204493.3204579

Bhattacharya, N., Rakshit, S., and Gwizdka, J. (2020a). "Towards real-time webpage relevance prediction using convex hull based eye-tracking features," in *ACM Symposium on Eye Tracking Research and Applications, ETRA '20 Adjunct* (New York, NY: Association for Computing Machinery). doi: 10.1145/3379157.3391302

Bhattacharya, N., Rakshit, S., Gwizdka, J., and Kogut, P. (2020b). "Relevance prediction from eye-movements using semi-interpretable convolutional neural networks," in *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval, CHIIR '20* (New York, NY: Association for Computing Machinery), 223–233. doi: 10.1145/3343413.3377960

Buchanan, G., McKay, D., Velloso, E., Moffat, A., Turpin, A., and Scholer, F. (2017). "Only forward? Toward understanding human visual behaviour when examining search results," in *Proceedings of the 29th Australian Conference on Computer-Human Interaction, OZCHI '17* (New York, NY: Association for Computing Machinery), 497–502. doi: 10.1145/3152771.3156165

Buscher, G., Dengel, A., Biedert, R., and Elst, L. V. (2012). "Attentive documents: eye tracking as implicit feedback for information retrieval and beyond," in *ACM Transactions on Interactive Intelligent Systems* (New York, NY: Association for Computing Machinery). doi: 10.1145/2070719.2070722

Buscher, G., Dengel, A., and van Elst, L. (2008a). "Eye movements as implicit relevance feedback," in *CHI '08 Extended Abstracts on Human Factors in Computing Systems, CHI EA '08* (New York, NY: Association for Computing Machinery), 2991–2996. doi: 10.1145/1358628.1358796

Buscher, G., Dengel, A., and van Elst, L. (2008b). "Query expansion using gaze-based feedback on the subdocument level," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08* (New York, NY: Association for Computing Machinery), 387–394. doi: 10.1145/1390334.1390401

Buscher, G., van Elst, L., and Dengel, A. (2009). "Segment-level display time as implicit feedback: a comparison to eye tracking," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09* (New York, NY: Association for Computing Machinery), 67–74. doi: 10.1145/1571941.1571955

Carpineto, C., and Romano, G. (2012). "A survey of automatic query expansion in information retrieval," in *ACM Computing Surveys* (New York, NY: Association for Computing Machinery). doi: 10.1145/2071389.2071390

Castner, N., Kuebler, T. C., Scheiter, K., Richter, J., Eder, T., Huettig, F., et al. (2020). "Deep semantic gaze embedding and scanpath comparison for expertise classification during OPT viewing," in *ACM Symposium on Eye Tracking Research and Applications, ETRA '20* (New York, NY: Association for Computing Machinery). doi: 10.1145/3379155.3391320

Cerrolaza, J. J., Villanueva, A., Villanueva, M., and Cabeza, R. (2012). "Error characterization and compensation in eye tracking systems," in *Proceedings of the Symposium on Eye Tracking Research and Applications, ETRA '12* (New York, NY: Association for Computing Machinery), 205–208. doi: 10.1145/2168556.2168595

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, Y., Zhang, P., Song, D., and Wang, B. (2015). "A real-time eye tracking based query expansion approach via latent topic modeling," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15* (New York, NY: Association for Computing Machinery), 1719–1722. doi: 10.1145/2806416.2806602

Cole, M. J., Gwizdka, J., Liu, C., Belkin, N. J., and Zhang, X. (2013). Inferring user knowledge level from eye movement patterns. *Inform. Process. Manage.* 49, 1075–1091. doi: 10.1016/j.ipm.2012.08.004

Davari, M., Hienert, D., Kern, D., and Dietze, S. (2020). "The role of word-eye-fixations for query term prediction," in *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval, CHIIR '20*, (New York, NY: Association for Computing Machinery), 422–426. doi: 10.1145/3343413.3378010

Eickhoff, C., Dungs, S., and Tran, V. (2015). "An eye-tracking study of query reformulation," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '15* (New York, NY: Association for Computing Machinery), 13–22. doi: 10.1145/2766462.2767703

Feit, A. M., Vordemann, L., Park, S., Berube, C., and Hilliges, O. (2020). "Detecting relevance during decision-making from eye movements for UI adaptation," in *ACM Symposium on Eye Tracking Research and Applications, ETRA '20* (New York, NY: Association for Computing Machinery), 13–22. doi: 10.1145/3379155.3391321

Feit, A. M., Williams, S., Toledo, A., Paradiso, A., Kulkarni, H., Kane, S., et al. (2017). "Toward everyday gaze input: accuracy and precision of eye tracking and implications for design," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery), 1118–1130. doi: 10.1145/3025453.3025599

Graff, D. (2002). *The AQUAINT Corpus of English News Text LDC2002T31*. Philadelphia, PA.

Gwizdka, J. (2014a). "Characterizing relevance with eye-tracking measures," in *Proceedings of the 5th Information Interaction in Context Symposium* (New York, NY: Association for Computing Machinery), 58–67. doi: 10.1145/2637002.2637011

Gwizdka, J. (2014b). "News stories relevance effects on eye-movements," in *Proceedings of the Symposium on Eye Tracking Research and Applications, ETRA '14* (New York, NY: Association for Computing Machinery), 283–286. doi: 10.1145/2578153.2578198

Gwizdka, J. (2017). "Differences in reading between word search and information relevance decisions: evidence from eye-tracking," in *Information Systems and Neuroscience*, eds F. D. Davis, R. Riedl, J. vom Brocke, P. M. Lager, and A. B. Randolph (Cham. Springer International Publishing), 141–147. doi: 10.1007/978-3-319-41402-7_18

Gwizdka, J., and Dillon, A. (2020). "Eye-tracking as a method for enhancing research on information search," in *Understanding and Improving Information Search: A Cognitive Approach*, eds W. T. Fu and H. van Oostendorp (Cham: Springer International Publishing), 161–181. doi: 10.1007/978-3-030-38825-6_9

Hardoon, D. R., Shawe-Taylor, J., Ajanki, A., Puolamaki, K., and Kaski, S. (2007). "Information retrieval by inferring implicit queries from eye movements," in

*Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, eds M. Meila and X. Shen (San Juan; Puerto Rico: PMLR), 179–186.

Hienert, D., Kern, D., Mitsui, M., Shah, C., and Belkin, N. J. (2019). "Reading protocol: understanding what has been read in interactive information retrieval tasks," in *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, CHIIR '19* (New York, NY: Association for Computing Machinery), 73–81. doi: 10.1145/3295750.3298921

Jacob, S., Ishimaru, S., Bukhari, S. S., and Dengel, A. (2018). "Gaze-based interest detection on newspaper articles," in *Proceedings of the 7th Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction, PETMEI '18* (New York, NY: Association for Computing Machinery). doi: 10.1145/3208031.3208034

Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. (2017). *Accurately Interpreting Clickthrough Data as Implicit Feedback*. New York, NY: Association for Computing Machinery. doi: 10.1145/3130332.3130334

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., et al. (2019). Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguist.* 7, 453–466. doi: 10.1162/tacl_a_00276

Lemaitre, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* 18, 1–5. Available online at: http://jmlr.org/papers/v18/16-365.html

Li, X., Liu, Y., Mao, J., He, Z., Zhang, M., and Ma, S. (2018). "Understanding reading attention distribution during relevance judgement," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18* (New York, NY: Association for Computing Machinery), 733–742. doi: 10.1145/3269206.3271764

Loboda, T. D., Brusilovsky, P., and Brunstein, J. (2011). "Inferring word relevance from eye-movements of readers," in *Proceedings of the 16th International Conference on Intelligent User Interfaces, IUI '11* (New York, NY: Association for Computing Machinery), 175–184. doi: 10.1145/1943403.1943431

Miller, T., and Agne, S. (2005). "Attention-based information retrieval using eye tracker data," in *Proceedings of the 3rd International Conference on Knowledge Capture, K-CAP '05* (New York, NY: Association for Computing Machinery), 209–210. doi: 10.1145/1088622.1088672

Olsen, A. (2012). *The Tobii I-VT Fixation Filter: Algorithm Description*, Danderyd: Tobii Technology.

Oviatt, S., Schller, B., Cohen, P. R., Sonntag, D., Potamianos, G., and Kruger, A. (eds.). (2018). *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition*, volume 2. New York, NY: Association for Computing Machinery.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Salojarvi, J., Kojo, I., Simola, J., and Kaski, S. (2003). "Can relevance be inferred from eye movements in information retrieval?" in *Workshop on Self-Organizing Maps (WSOM'03)* (Hibikino), 261–266.

Salojarvi, J., Puolamaki, K., and Kaski, S. (2004). "Relevance feedback from eye movements for proactive information retrieval," in *Workshop on Processing Sensory Information for Proactive Systems (PSIPS 2004)* (Oulu), 14–15.

Salojarvi, J., Puolamaki, K., and Kaski, S. (2005a). "Implicit relevance feedback from eye movements," in *Artificial Neural Networks: Biological Inspirations – ICANN 2005* (Berlin; Heidelberg), 513–518. doi: 10.1007/11550822_80

Salojarvi, J., Puolamaki, K., Simola, J., Kovanen, L., Kojo, I., and Kaski, S. (2005b). *Inferring Relevance from Eye Movements: Feature Extraction*. Helsinki University of Technology.

Salvucci, D. D., and Goldberg, J. H. (2000). "Identifying fixations and saccades in eye-tracking protocols," in *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications, ETRA '00* (New York, NY: Association for Computing Machinery), 71–78. doi: 10.1145/355017.355028

**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.