Examining the Merits of Feature-specific Similarity Functions in the News Domain using Human Judgments

Alain D. Starke[1,2*], Vegard R. Solberg[2], Sebastian Øverhaug[2] and Christoph Trattner[2]

[1*]Amsterdam School of Communication Research, University of Amsterdam, P.O. Box 15791 Amsterdam, 1001 NG, The Netherlands.

[2]MediaFutures, Department of Information Science and Media Studies, University of Bergen, Lars Hilles gate 30, Bergen, 5008, Vestland, Norway.

*Corresponding author(s). E-mail(s): a.d.starke@uva.nl;
Contributing authors: vegard.solberg@uib.no;
overhaug15@gmail.com; christoph.trattner@uib.no;

Abstract

Online news article recommendations are typically of the 'more like this' type, generated by similarity functions. Across three studies, we examined the representativeness of different similarity functions for news item retrieval, by comparing them to human judgments of similarity. In Study 1 ($\boldsymbol{N} = \mathbf{401}$), participants assessed the overall similarity of ten randomly paired news articles on politics, and compared their judgments to different feature-specific similarity functions (e.g., based on body text or images). In Study 2, we checked for domain differences in a mixed-methods survey ($\boldsymbol{N} = \mathbf{45}$), surfacing evidence that the effectiveness of similarity functions differs across different news categories ('Recent Events', 'Sport'). In Study 3 ($\boldsymbol{N} = \mathbf{173}$), we improved the design of Study 1, by controlling for how news articles were matched, differentiating between dissimilar news articles and articles that were matched on a shared topic, named entities, and/or date of publication, across 'Recent Events' and 'Sport' categories. Across all studies, we found that users mostly used text-based features (e.g., body text, title) for their similarity judgments, while

BodyText:TF-IDF was found to be the most representative for their judgments. Moreover, the strength of similarity judgments by humans and similarity scores by feature-specific functions was strongly affected by how news article pairs were matched. We show that humans and similarity functions are better aligned when two news articles are more alike, such as in a news recommendation scenario.

Keywords: news, similarity, similar-item retrieval, recommender systems, human judgment

1 Introduction

Similarity functions are central to recommender systems and information retrieval systems [1]. They assess the similarity between a reference article and a set of possible recommendations [2]. While there has been a lot of research into algorithmic optimization of news retrieval and recommendation [1, 3], less is known about how users evaluate presented recommendations. Specifically, similarity detected by retrieval approaches is rarely tested for human judgment. This paper employs a semantic similarity approach to assess the utility of different feature-based similarity functions in the news domain, grounding them in human judgments of similarity. We report the results of three studies, also examining to what extent both similarity functions and human judgments are affected by straightforward methods of topical matching.

1.1 Problem Outline

News retrieval faces several domain-specific challenges. Compared to leisure domains (e.g., movies), news articles are volatile, in the sense that they become obsolete quickly, may be updated later, or are superseded by other breaking news events [4–6]. Moreover, user preferences for news may also strongly depend on contextual factors, such as the time of day or a user's location [7, 8], and may change rapidly due to major events that may impact a user's life.

Many news websites employ, in part due to cold-start problems [1, 7, 8], content-based recommender systems [9]. A common setup is to present a list of articles that are similar to the story the user is currently reading, such as depicted in Figure 1. These are often labeled 'More on this Story' (e.g., at BBC News), showcasing similar articles in terms of their publication time or specific keywords. Similar-item or related-item recommendations as these are also found in other domains, helping users to explore commodities (e.g., photo cameras, videos, etc.) that are similar to, but also slightly different from an item that is currently being inspected [10–12]. These systems are often implemented at e-commerce platforms (e.g., Zalando) to keep users engaged with the service, particularly supporting those users who have a specific product goal in mind (e.g., a red dress) [13]. When it comes to news recommendation

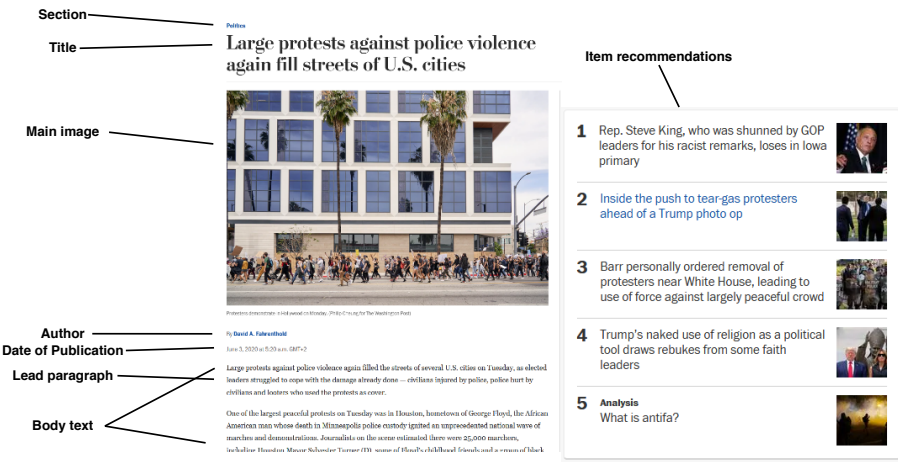

Fig. 1 Different features in a news article, which may be used by a news recommender system to recommend items to a user.

scenarios, it seems that users prefer to be presented similar content-based recommendations, compared to diverse content-based recommendation and news articles generated by collaborative filtering [14].

Whether two news articles are alike can be computed using similarity functions [1, 8]. Features (e.g., title) considered by such functions should to a large extent reflect a user's similarity assessment [15], while not being too similar to what a user is currently reading, for it may lead to redundancy [2]. However, research on feature-based similarity is limited and rather domain-dependent. For example, users browsing on recipe websites tend to use titles and header photos to assess similarity between recipes, while users of movie recommenders use plot descriptions and genre [16]. As a result, there is no consensus on which news article features best represent a user's similarity judgment. This may be problematic, as similarity functions in recommender systems may be more effective if they reflect user perceptions. We, thus, consider it an open question which news features are actually being paid attention to by users, assuming that this For this study, it is an open question which parts and features of a news

In this paper, we examine how similar news recommendation should be generated. While similar-item recommendation is common [1], the concept of similarity in relation to algorithmic similarity functions is not well understood and might be disconnected from a more human-based understanding of similarity. To alleviate this, we assess a set of similarity functions for news article retrieval, particularly for the task of similar-item recommendation. We ask users of online news systems to judge the similarity between pairs of news articles, which is used to develop a model to predict news similarity. We build upon work from Trattner and Jannach [16], that used feature-specific similarity functions to assess similarity in the movie and recipe recommendation domains. In addition to examining the representativeness of these functions in

the news domain, we examine more specifically what affects both similarity computations (by similarity functions) and judgments (by humans), based on the underlying news category and different matching characteristics.

We present three studies. In Study 1 (i.e., an extension from a conference workshop paper published earlier [17]), we use a dataset of political news articles to examine which news article features are used by humans to judge similarity. Besides inquiring on feature or cue usage, we investigate this by presenting news article pairs to users. For each pair, we ask users to judge the similarity between the two articles, while also computing the similarity between them based on different feature-specific functions, which is then compared. We find that even the best performing similarity functions (i.e., based on body text) are only modestly representative of human similarity judgments, particularly when compared to the much stronger correlations found in the movie and recipe domains [16]. As this could be attributed to the news domain involved and the random pairing of news articles, we performed two additional studies.

In Study 2, we inquire on what users perceive as important factors in a news recommendation scenario, in mixed methods survey. We specifically examine differences in terms of the type of news articles involved, comparing traditional 'recent events' news articles and 'sport' news articles. Similarity functions may be more representative of human judgment depending on the goal of the news article (e.g., to only inform or also to entertain [1]), or whether the news article is centered around a specific person or a news event. We find that 'recent events' news similarity relies mostly on topic-based similarity, while 'sport' news articles could leverage the use of named entities (cf. [18]).

Based on the findings of Study 2, we determine that news articles could be matched on three different characteristics in news recommendation: Topic, Named Entities, and Date-time. In Study 3, we address the limitation of Study 1 that news articles were paired randomly, presenting a more realistic similar-item retrieval scenario in which news article pairs are similar. We compare similarity scores by functions and similarity judgments by humans across different matching characteristics, based on a dataset from British news website The Guardian. We find that both similarity judgments and scores increase when news articles are matched on topic and named entities. Moreover, we find a stronger relation between human judgment and the most representative feature-specific similarity functions, based on title and body text, with some small differences across 'Recent Events' and 'Sport' news categories.

The different studies are covered by the following research questions:

- **RQ1.1:** To what extent are news article features and similarity functions based on those features related to human similarity judgments?
- **RQ1.2:** Which combination of news article features is best suited to predict user similarity judgments?
- **RQ2:** To what extent do similarity computations and judgments depend on the news category domain?

- **RQ3:** To what extent are both the computed and perceived similarity between two news articles affected by how these two are matched?

1.2 Contributions

This paper makes the following contributions:

- We advance the understanding of how readers perceive similarity between news articles, in terms of (i) which article cues or features are reported as important, and (ii) how features correlate with similarity ratings provided by users, (iii) that user-reported feature importance is not always consistent with the computed correlations.
- We show which news information features can predict a user's similarity judgment.
- We present a qualitative study in which we provide evidence for domain differences within news retrieval and recommendation, highlighting different user expectations across 'Recent Events' and 'Sport' news articles
- We show that both similarity judgments and scores strongly depend on how well two news articles match in terms of topic and named entities, confounding possible results from earlier semantic similarity studies in news.
- We present a reproducible data processing pipeline for Study 1, available on Github[1], and add a benchmarking dataset for the publicly available Washington Post Corpus news article database.

2 Related Work

We highlight work from the domains of Similar-item Retrieval and Semantic Similarity to craft similarity functions. Moreover, we discuss specific challenges in news recommendation, and explain how similarity functions are assessed by using human similarity judgments as ground truth. Finally, we discuss the influence of news categories in news recommendation, noting that what we refer to typical 'News' articles as a 'Recent Events' category, complementary to other categories, such as 'Sport'.

2.1 Similar Item Retrieval

Similar item retrieval seeks to identify *unseen* or *novel* items that are similar to what a user has elicited preferences for [1]. In the recommender domain, this is referred to as a similar-item recommendation problem. A fundamental question is how to compute similarity between concepts [13, 19], which is examined in studies on semantic similarity [20], a field of research that usually not only captures the similarity between two concepts, but also how different they are [21]. This can be based on ontological relations, based on human knowledge, or on co-occurrence metrics that stem from a hierarchical or annotated corpus of words [2, 22]. For example, latent semantic analysis derives meaning and

[1] https://github.com/Overhaug/HuJuRecSys

similarity from the text context itself, by examining *how* and *how often* words are used [2].

A traditional method is to compute similarity between items by deriving *vectors* from text items. Although TF-IDF has been outperformed by other metrics, such as BM25 [23], *Term Frequency-Inverse Document Frequency* remains one of the most commonly used IR methods to create similarity vectors [24]. It uses the term frequency per document and the inverse appearance frequency across all documents [10], while similarity between the vectors of liked and unseen items can be computed using cosine similarity [25]. Such metrics are popular because they can be applied to various datasets, for news always offers text-based content for retrieval.

A much simpler approach is to derive a set of keywords from each item [10]. For example, a book recommender could compute the similarity between $book1 = fantasy, epic, bloody$, and $book2 = fantasy, young, dragons$, through the *Jaccard coefficient*: $J(A, B) = \frac{|book1 \cap book2|}{|book1 \cup book2|}$. There are various other similarity metrics available, such as the Levenshtein distance (i.e., 'edit distance'), and LDA (Latent Dirichlet Allocation). For news, keywords could be useful when a news outlet takes care in curating and selecting them.

2.2 Similarity Representations in the News Domain

News recommender systems primarily focus on textual representations of news articles [1]. Most approaches utilize the main text or title, ignoring most other textual features, such as the author [24]. A straightforward, but more uncommon approach in academic studies [26], is to retrieve articles based on date-time, such as those that are published on the same day as the article that is currently inspected. Other approaches include the use of (sub)categories, while image-based similarity is more common in other domains [27], such as food [16]. An overview of features and similarity functions used in the context of news recommendation can be found in Table 1.

2.2.1 Text-based approaches

Most similarity functions relevant in news retrieval are text-based. TF-IDF is traditionally combined with Cosine similarity and used as a news recommendation benchmark [28]. In some cases, its effectiveness can be improved by constraining it on a maximum number of words [3]. TF-IDF can also be combined with a K-Nearest Neighbor algorithm to recommend short-term interest news articles [29].

Besides the aforementioned methods, a common approach is to derive latent topics from texts. Although recent work uses Word2Vec and BERT [30, 31], this work considers Latent Dirichlet Allocation (LDA) and Probabilistic Latent Semantic Indexing (PLSI) [32, 33]. LDA and PLSI can cluster topically-similar news articles based on tags and named entities.

A combination of Cosine similarity and K-Nearest Neighbors can be used to measure the similarity between two TF-IDF vectors and, subsequently, to

identify articles that belong to the same *thread of events* – that a user already knows. In addition, [29] point out how to identify long-term interests using a Naive Bayesian classifier, which is shown to perform competitively with more complex algorithms.

A final interesting text-based method is based on sentiment analysis. Sentiment analysis mines a text's opinions in terms of the underlying attitude, judgments, and beliefs. It has been suggested that negativity in news has a large impact, triggering more vivid recall of news story details among users [43].

2.2.2 Other News Features

A news article's date-time feature is also leveraged in the context of similar-item news recommendation [1]. Date-time tends to be popular when a news platforms aims to show recent stories, or when it wishes to retrieve stories from a specific time periods [44]. Date-time can be applied either through pre-filtering, recency modeling, or post-filtering [1]. Pre-filtering involves omitting outdated news articles before computation starts, while the more uncommon post-filtering removes all non-recent articles from a Top-N set. Recency modeling is the most common, which incorporates recency as one of the factors in an algorithm's similarity computation (e.g., by giving it a higher weight). Pon et al. [42] describe an approach that targets users with multiple interests, by considering recency in conjunction with a 'multiple topic tracking' technique.

2.3 Assessing Similarity Functions Using Human Judgments

Similar-item retrieval approaches, commonly utilized in recommender systems, are typically validated using human judgments [22]. For example, if a database is used to compute the similarity between various news headlines [2], a group of participants are then invited to judge the similarity between these headlines,

Table 1 Methods used in similar-item retrieval and recommendation news scenarios for specific features. Sets of citations apply to the methods before a semicolon.

Feature	Method & Literature
Title	Okapi BM25, Language model Jelinek-Mercer (LM-JM), Language model Dirichlet prior (LM-DIR), Cosine similarity *[23]*; TF-IDF *[34]*; Dependency structure language model (DSLM) *[35]*;
Body text	Okapi BM25, Language model Jelinek-Mercer (LM-JM), Language model Dirichlet prior (LM-DIR), Cosine similarity *[23]*;.
Abstract	Okapi BM25, Language model Jelinek-Mercer (LM-JM), Language model Dirichlet prior (LM-DIR) *[23]*;.
All text	TF-IDF & K-Nearest Neighbor *[24, 29, 36]*); Cosine Similarity, Naïve Bayesian classifier *[24]*; Overlap Coefficient *[37]*; Probabilistic Latent Semantic Indexing (PLSI) *[33]*; Latent Dirichlet Allocation *[4, 33, 38]*; Fisher Kernel Function (PLSA) *[39]*; Dependency structure language model (DSLM) *[35]*;
Imagery	Image-label overlap similarity *[26]*;
Date-time	Pre-filtering *[40, 41]*; Recency modeling *[7, 38, 42]*;

after which the two outcomes are compared. This principle has been applied in a number of domains and methods, also referred to as 'semantic similarity' [2], including web documents and graph-based approaches [22].

Similarity functions can be the starting point for a recommendation method. The underlying assumption is that a high degree of similarity (i.e., in terms accuracy, with distance metrics) for specific features would represent user preferences for additional news content. An important question is to what extent these similarity functions reflect a user's similarity assessment of item pairs. This could lead to problems if a user either ignores or overvalues different item features, compared to what is being computed [13]. For example, the click-through rate is a popular metric for analyzing the success of news recommender systems' recommendations [45]. In the context of similar news recommendation, however, it is not particularly helpful if our goal is to determine whether or not people agree that the recommended items are truly similar.

Trattner and Jannach [16] have studied the representativeness of similarity functions in the movie and recipe domains. They contrast user similarity assessments to a set of similarity functions, pointing out that specific features (e.g., a recipe's title or a movie's genre) strongly correlate with user similarity judgments. In a similar vein, Yao and Harper [12] assess to what extent different algorithms for related item recommendations in music are consistent with user similarity judgments.

However, assessing similarity between news articles might be harder than between movies. Whereas similarity between movie pairs is usually attributed to the annotated metadata (e.g., genre), two news articles could be similar because they are recent, address a common topic, or because a person appears in both stories. Although a few studies let humans assess the overall similarity between news headlines [2, 46], none have done so across multiple features. For example, users in the work of Tintarev and Masthoff [2] successfully judged the similarity between news articles, but only based on their headlines. Finally, Winecoff et al. [13] present similar item retrieval functions that are 'psychologically-aware'. They use the contrast model from [47] to better predict human judgments of similarity in fashion recommendation. Although this research is promising, our current work focuses on the representativeness of more traditional, 'psychology-naive' information retrieval functions, as found in [16].

2.4 Matching News Articles

The strength of detected similarity may be predicted by the available metadata of a reference news article and novel recommendations. In this work, we capture them in specific 'matching characteristics'.

The profiles created to capture a user's reading interests in news recommender systems is a reasonable starting point when searching for factors that determine similarity. Keyword profiles used to be the most prevalent user profile type [48], which are comprised of a list of keywords that indicate topics

of interest, with each term being assigned a numerical value that reflects its significance to the profile. This argues for topic-based similar-item retrieval, which is later extended by Li et al. [33] towards profiles that also capture a user's interests in named entities. They found that recommender systems that incorporate preferred named entities perform better than those that do not.

A different factor, separate from user profiles, is the proximity in publication date between two news articles. Recency is a popular approach in news recommendation, particularly to alleviate cold-start problems [1], showing the most recently published news articles. However, date-time can also be leveraged to locate news articles that are published on the same (part of the) day [49, 50], to match two news articles.

We propose three matching characteristics for news articles that may be relevant in a similar-item retrieval scenario. Matching news articles on topic and named entities seems most sensible from the news recommendation literature. Topic-based retrieval allows users to learn more about one specific news event or subject, supporting news values related to education [51, 52]. Named entity-bases retrieval works similarly, In addition to topic and named entities, people might like to read more news articles from a specific period, we propose to also leverage date to match news articles.

2.5 Key Differences with Previous Work

Novel to our approach is the use of feature-specific similarity representations in news, as well as grounding them in human similarity judgments. Comparing the representativeness of specific functions in terms of human judgment that apply to specific features is novel for news recommender systems [1]. Various studies test the accuracy of different algorithmic approaches without understanding what aspects or features of a news article a user specifically values.

Most relevant to our approach are the works of Trattner and Jannach [16], and Yao and Harper [12], for they explore how computational functions for similarity compare to users' perception of similarity. In particular, Trattner and Jannach [16] serve as an example for our approach in Study 1, for they also present an online study on similarity perceptions. However, these studies concerned retrieval in music, movies, and recipes. Since the merit of feature-specific similarity functions in other domains is unknown for news, the current study aims to assess their performance in news.

Another key difference is the use of different methods in the current paper, across three studies. Whereas Study 1 and Study 3 are computationally-driven experiments, Study 2 comprises survey with also qualitative methods. Because of the findings in Study 2, we have been able to address the shortcomings of the research design in Study 1, taking heed of the different matching characteristics: topic, named entities, and date.

3 General Methods: Modeling Similarity with Feature-Based Similarity Functions

To model the similarity between two news articles in Study 1 and Study 3, we used twenty similarity functions and representations across seven dataset features. We designed functions in line with the field's current state of the art, by exploiting specific cues that people may use to assess similarity between two items. The similarity functions used were based on findings from the movie and recipe domains [16]. Therefore, they also reflect the most important metrics from around 2019.

Table 2 describes the developed similarity functions. These were based on a variety of Python-based libraries, while Java was used for image attributes (OpenIMAJ). For stop word removal and stemming, we NLTK.

For each pair of news articles, we computed similarity scores based on seven main features. These included subcategory, title, presented images, author (including bio), publication dates, and body text (first 50 words and full text). For text-based features, the similarity functions were either based on word mappings or distance methods, while similarity based on subcategories and authors was computed using a Jaccard coefficient. Moreover, we computed date-time similarity (i.e. recency modeling) through a linear function that computed how many days apart two articles were published.

Table 2 Similarity functions employed in the current study, each comprised of a feature (i.e., name) and a metric.

Name	Metric	Explanation
Subcat:JACC	$sim(n_i, n_j) = \frac{subcat(n_i) \bigcap subcat(n_j)}{subcat(n_i) \bigcup subcat(n_j)}$	Subcategory Jaccard-based similarity
Title:LV	$sim(n_i, n_j) = 1 - \lvert dist_{LV}(n_i, n_j)\rvert$	Title Levenshtein distance-based similarity
Title:JW	$sim(n_i, n_j) = 1 - \lvert dist_{JW}(n_i, n_j)\rvert$	Title Jaro-Winkler distance-based similarity
Title:LCS	$sim(n_i, n_j) = 1 - \lvert dist_{LCS}(n_i, n_j)\rvert$	Title longest common subsequence distance-based similarity
Title:BI	$sim(n_i, n_j) = 1 - \lvert dist_{BI}(n_i, n_j)\rvert$	Title bi-gram distance-based similarity
Title:LDA	$sim(n_i, n_j) = \frac{LDA(Title(n_i)) * LDA(Title(n_j))}{\lvert\lvert LDA(Title(n_i))\rvert\rvert \lvert\lvert LDA(Title(n_j))\rvert\rvert}$	Title LDA cosine-based similarity
Image:BR	$sim(n_i, n_j) = 1 - \lvert BR(n_i) - BR(n_j)\rvert$	Image brightness distance-based similarity
Image:SH	$sim(n_i, n_j) = 1 - \lvert SH(n_i) - SH(n_j)\rvert$	Image sharpness distance-based similarity
Image:CO	$sim(n_i, n_j) = 1 - \lvert CO(n_i) - CO(n_j)\rvert$	Image contrast distance-based similarity
Image:COL	$sim(n_i, n_j) = 1 - \lvert COL(n_i) - COL(n_j)\rvert$	Image colorfulness distance-based similarity
Image:EN	$sim(n_i, n_j) = 1 - \lvert EN(n_i) - EN(n_j)\rvert$	Image entropy distance-based similarity
Image:EMB	$sim(n_i, n_j) = \frac{EMB(n_i) * EMB(n_j)}{\lvert\lvert EMB(n_i)\rvert\rvert \lvert\lvert EMB(n_j)\rvert\rvert}$	Image embedding cosine-based similarity
Author:JACC	$sim(n_i, n_j) = \frac{author(n_i) \bigcap author(n_j)}{author(n_i) \bigcup author(n_j)}$	Author Jaccard-based similarity
Date:ND	$sim(n_i, n_j) = 1 - \lvert dist_{days}(n_i, n_j)\rvert$	Date published distance-based similarity (unit = days)
BodyText:TFIDF	$sim(n_i, n_j) = \frac{TFIDF(Text(n_i)) * TFIDF(Text(n_j))}{\lvert\lvert TFIDF(Text(n_i))\rvert\rvert \lvert\lvert TFIDF(Text(n_j))\rvert\rvert}$	All article body text cosine-based similarity
BodyText:50TFIDF	$sim(n_i, n_j) = \frac{TFIDF(Text(n_i)) * TFIDF(Text(n_j))}{\lvert\lvert TFIDF(Text(n_i))\rvert\rvert \lvert\lvert TFIDF(Text(n_j))\rvert\rvert}$	First 50 words in article body text cosine-based similarity
BodyText:LDA	$sim(n_i, n_j) = \frac{LDA(Text(n_i)) * LDA(Text(n_j))}{\lvert\lvert LDA(Text(n_i))\rvert\rvert \lvert\lvert LDA(Text(n_j))\rvert\rvert}$	All article body text LDA cosine-based similarity
BodyText:Senti	$sim(n_i, n_j) = 1 - \lvert SENTI(n_i) - SENTI(n_j)\rvert$	Article body text sentiment distance-based similarity
AuthorBio:TFIDF	$sim(n_i, n_j) = \frac{TFIDF(Text(n_i)) * TFIDF(Text(n_j))}{\lvert\lvert TFIDF(Text(n_i))\rvert\rvert \lvert\lvert TFIDF(Text(n_j))\rvert\rvert}$	Author bio cosine-based similarity
AuthorBio:LDA	$sim(n_i, n_j) = \frac{LDA(Title(n_i)) * LDA(Title(n_j))}{\lvert\lvert LDA(Title(n_i))\rvert\rvert \lvert\lvert LDA(Title(n_j))\rvert\rvert}$	Author bio LDA cosine-based similarity

3.1 Title

Title-based similarity was computed using four string similarity functions and a topic-based one. The string-based functions were based on distance metrics: the Levenshtein distance (LV) [53], the Jaro-Winkler method (JW) [54], the longest common subsequence, and the bi-gram distance method (BI) [55]. Similar to Trattner and Jannach [16], Latent Dirichlet Allocation (LDA) topic-modeling was set to 100 topics.

3.2 Image Features

In line with the current state-of-the-art [16], we computed image-based similarity using six different functions. These were an image's brightness, sharpness (i.e., based on a pixel's intensity), contrast, colorfulness (i.e., based on the sRGB color space), entropy (i.e., amount of information captured per image dot), and image embeddings. Mathematical details are available in our Github repository[2].

3.3 Body Text

Body text similarity was computed for two string-based functions (i.e., TF-IDF), a topic-based function (i.e., LDA), and a text sentiment-based metric (based on research of [43]). TF-IDF encodings were paired with cosine similarity, for which we discerned between similarity based on an article's first 50 words (i.e., an article's first paragraph), which could be compared to the average movie plot length in [16], and similarity based on the entire body text.

3.4 Other features

We have additionally examined a news article's date-time, subcategory, its author(s) and the author's biography. The latter is specific to the Washington Post corpus and might not be available for news article datasets.

Author-based and subcategory-based metrics consisted of a single keyword metric, the Jaccard coefficient. They follow Trattner and Jannach's use of the Jaccard coefficient [16], in that study for a movie's director and genre. Date-time consisted of a linear function which computed the similarity based on how many days apart two articles were published.

4 Study 1: Assessing Feature-specific Similarity

We assess the utility of different feature-specific similarity functions by collecting human judgments of similarity for pairs of news articles. In doing so, we examine which news article features or cues are used by humans and to what extent the different features and similarity functions are representative of human judgment [RQ1.1]. In addition, we model human judgments, predicting which features can most accurately predict them [RQ1.2].

[2] https://github.com/Overhaug/HuJuRecSys

Table 3 Descriptive statistics and contents of the dataset employed for the user study.

Feature	Mean	Median	Min	Max
Number of words in title	9.78	10	2	25
Number of characters in title	60.16	61	11	195
Article image brightness	0.37	0.35	0.04	0.98
Article image sharpness	0.24	0.2	0.03	1.27
Article image contrast	0.18	0.18	0.01	0.64
Article image colorfulness	0.17	0.16	0	0.73
Article image entropy	7.05	7.33	0.75	7.95
Number of words in article body text	768.44	637	6	10640
Number of characters in article body text	4676.99	3895.5	38	65641
Article body text sentiment	0.54	0.54	0.05	0.89
Date of publication	2015-01-04	2014-12-31	2012-01-10	2017-08-22
Number of words in author biographies	21.63	17	4	306
Number of characters in author biographies	140.32	115	33	1989
Number of authors	1.05	1	1	8

4.1 Method

In this section, we describe (1) the dataset and its specific features, (2) the engineered similarity functions, and (2) the design of our user study to determine the effectiveness of these functions.

4.1.1 Dataset and Feature Engineering

We employed a publicly available news article database. We focused on a scenario of a single news source, as the use of multiple news websites could lead to 'duplicate' articles on the same news event. To ensure reproducibility, we obtained news articles from the open Washington Post Corpus [56]. The news items in the dataset comprised title, author (including a bio), date of publication, section headers, and the main body text. In addition, we retrieved the images associated with the news articles, 655,533 in total. After removing duplicates from the original source, our remaining dataset contained 238,082 articles, which were originally published between Jan'12 and Aug'18.

For our user study, we selected news articles categorized in 'Politics', as they were on (inter)nationally relevant topics. Other categories were neglected as they focused more on local events and may have an effect on similarity estimates, as these events may not be familiar to the user. We sampled a total of 2,400 'Politics' news articles, 400 from each year between 2012 and 2017, for the descriptive statistics are reported in Table 3.

4.1.2 User Study

The similarity functions in Table 2 were assessed by computing similarity scores per news article pair and comparing them to human judgments. We explain our sampling strategy and how we collected human judgments of similarity.

Sampling News Article Pairs on Similarity

We compiled a set of news article pairs that were either strongly similar, dissimilar or in-between. To ensure a good distribution, we employed a stratified sampling strategy that was in line with previous work [16]. We computed the pairwise similarity across all 2,400 news articles, averaging the similarity values of all functions in Table 2. Pairs were ordered on their similarity levels and divided into ten deciles, groups D1-D10 of equal size. We sampled a total of 6,000 news article pairs: 2,000 dissimilar pairs between decile D1, 2,000 pairs from deciles D2-D9, and 2,000 similar pairs from decile D10.

Procedure and Measures

The resulting 6,000 news article pairs were used to collect human judgments on similarity. Figure 2 depicts a mock-up of the main application, showing from top to bottom different news article features (Note: an author bio could also be inspected). Users could read all text if they clicked 'read more'.

Users were presented ten news article pairs, of which one was an attention check.[3] Much like in the study by Tintarev and Masthoff [2], users were asked to

[3]Instead of being presented regular news content, the body text of the attention check news articles asked the user to only answer '5' on all answer scales.

Fig. 2 Mock-up of a pair-wise similarity assessment in our web application. Users were asked to assess the similarity of two presented news articles, as well as how familiar they were with the articles and the confidence level of their judgment.

assess the similarity of each news article pair on a 5-point scale (cf., Figure 2). As an extension to other studies, users also indicated their familiarity with each article and the level of confidence in their assessment (all 5-point scales). Moreover, we asked users to what extent they employed different features in their similarity judgments (5-point scales). Finally, we inquired on a user's frequency of news consumption and their demographics.

Participants

Participants were recruited from Amazon MTurk. Since we used a database of news articles that concerned American politics, we only recruited U.S.-based participants. They had at least an average hit acceptance rate of 98% and 500 completed HITs. A total of 401 participants completed our study, with a median time of 6 minutes and 35 seconds, who were compensated with 0.5 USD.

Only 241 participants (60.01%) passed our attention check, which was slightly higher than in [16]. This resulted in usable 2,169 similarity judgments; only 21 pairs were presented twice, to different users. This final sample (53% males) mostly consisted of age groups 25-34 (33.2%) and 35-44 (30.3%), of which 66% reported to visit news websites at least once a week (24.9% did so daily), while 50 participants rarely read online news. We performed t-tests on the available demographic characteristics between the full and the reduced sample. In terms of significant differences, we only found that a difference in terms of gender, where the full sample was significant more skewed towards men (58.1%) than the reduced sample (53.1%).

4.2 Results

For our analyses, we first examined the use of different news features and assessed different similarity functions through human judgments [RQ1.1]. Furthermore, we predicted human similarity judgments using model-based approaches [RQ1.2].

4.2.1 Information Cue Usage

We examined to what extent participants used different cues or features to assess similarity between news articles [RQ1.1]. Figure 3A summarizes the results for participants who passed the attention check. On average, an article's title (M=4.2) and body text (M=4.4) were considered most often, while sentiment (M=3.7) and an article's subcategory (M=3.2) saw above average use. In contrast, author features, publication date, an article's image were rarely used to assess similarity. Figure 3B shows that all differences between features were significant (all: $p < 0.01$), based on a one-way ANOVA on feature usage and a Tukey's HSD post-hoc analysis.

Most findings were in line with earlier findings from the movie and recipe domains [16]. The use of title and body text was also observed for recipes (i.e., ingredients and directions), while plot and genre features were used in

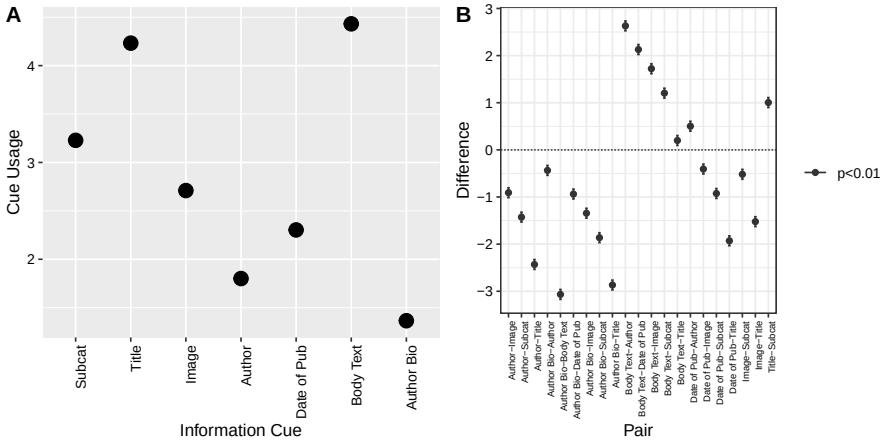

Fig. 3 **A**: Mean reported cue usage for news articles, scaled 1-5; **B**: Tukey's HSD post hoc tests (means and S.E.) that examine differences in cue usage.

movies. In that sense, the use of these features in news as a starting points for similar-item retrieval was supported.

4.2.2 Grounding Similarity Functions in Human Similarity Judgments

To further address [RQ1.1], we compared feature-specific similarity scores of presented news article pairs to similarity ratings given by users. Figure 4 contrasts the similarity scores, averaged across all similarity functions, with the users' similarity judgments, averaged per user. As shown, there was a discrepancy between the similarity inferred by the similarity functions, which was distributed around the mean value of 0.39 ($SD = 0.085$), and the similarity judgments of users, which was lower ($M = 0.18$, $SD = 0.24$). This suggested that users were less likely to judge two news articles to be similar, compared to our similarity functions.

Feature-specific Comparison in News. Table 4 outlines the Spearman correlations between similarity functions and the similarity judgments given by users. We focus on users who passed the attention check. Table 4 shows that most correlations were modest (all $\rho < 0.3$), suggesting that the news similarity functions did not fully reflect a user's judgment. Among all features, we found that full body text similarity (*BodyText:TFIDF*) correlated most strongly to user judgments: $\rho = 0.29$, $p < 0.001$, which was also the most commonly used feature in earlier news recommendation scenarios [1]. Although some users might have only inspected an article's first 50 words (cf., the text visible in Figure 2; on average 15% of the full body text), the *BodyText:50TFIDF* metric had a much lower correlation: $\rho = 0.14$, $p < 0.001$.

Among all image similarity metrics, embeddings (*Image:EMB*) had the highest correlation with user judgments: $\rho = 0.17^{***}$, which was modest nonetheless. This function, along with *BodyText:TFIDF*, *Author:Jacc*,

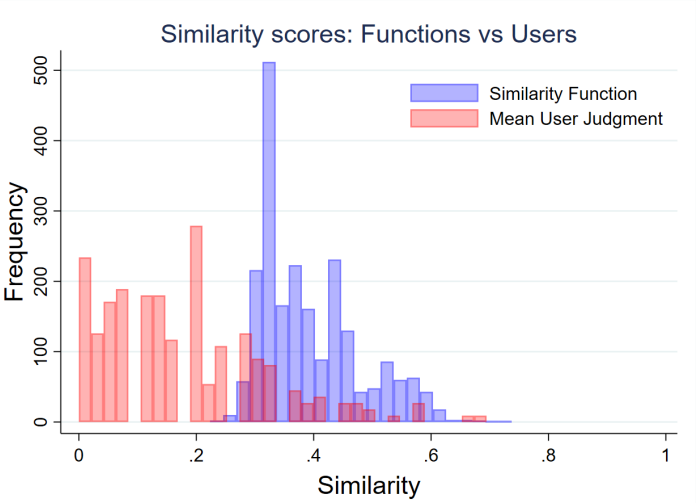

Fig. 4 Frequency of similarity scores (scaled 0-1). Similarity functions depict the average score per news article pair, user judgments show the mean given similarity judgment per user.

AuthorBio:TFIDF, and *Subcat:Jacc*, seemed to best represent user similarity judgments in news.

Table 4 highlights that other functions did not represent a user's similarity judgment, such as sentiment (*BodyText:Sent*): $\rho = -0.02$. Surprisingly, although most users considered titles to assess similarity, their judgments were hardly similar to each distance-based title similarity function (all $\rho < 0.1$). Note that the *Title:LDA* and *BodyText:LDA* might have suffered from insufficient latent topic information, as their correlations were close to zero.

Finally, because similarity ratings correlated positively with familiarity scores ($\rho = 0.27^{***}$), we tested whether only including judgments for familiar news article pairs (i.e., with scores of 4 or higher) affected the results in Table 4. Although this would increase correlations with 1 to 4 percentage points for most features, most changes were statistically significant (e.g., *TFIDF:BodyText* would increase from 0.29 to 0.33).

When inspecting the strength of the correlations, they were found to be lower than correlations observed in studies on the movie and recipe domains [16]. This applied to most features, including title, image, and body text.

4.2.3 Predicting Human Similarity Judgments

Going beyond simple correlation analyses, we also sought to predict similarities with these functions using state-of-the-art machine learning methods [RQ1.2], as used in recommender systems research. This helped us to understand each feature's importance, beyond the feature-specific correlations in Table 4.

Model Evaluation. To determine model performance, standard metrics such as Root Mean Square Error (RMSE), R^2, and Mean Absolute Error

Table 4 Spearman correlations between similarity functions and human similarity judgments in Study, for news articles in politics. ρ_{pass} denotes correlations with users who passed the attention check, ρ_{all} denotes those with all users. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

	News Articles	
Similarity Function	ρ_{pass}	ρ_{all}
Subcat:Jacc	0.14***	0.11
Title:LV	0.06**	0.04*
Title:JW	0.05*	0.03
Title:LCS	0.07***	0.05**
Title:BI	0.08***	0.07***
Title:LDA	0.02	0.00
Image:BR	0.10***	0.07***
Image:SH	0.06**	0.03
Image:CO	0.05*	0.05**
Image:COL	0.05*	0.03*
Image:EN	0.07**	0.05**
Image:EMB	0.17***	0.13***
Author:Jacc	0.13***	0.10***
Date:ND	0.09***	0.08***
BodyText:TFIDF	0.29***	0.23***
BodyText:50TFIDF	0.14***	0.12***
BodyText:LDA	0.03	0.01
BodyText:Sent	-0.02	-0.02
AuthorBio:TFIDF	0.15***	0.12***
AuthorBio:LDA	0.11***	0.09***

(MAE) were used. Five-fold cross-validation was used as an evaluation protocol. Furthermore, by applying grid search on a validation set from the training data, the optimal hyper-parameters for each model were found.

The performance of the models on News Articles is described in Table 5. In part (i), a Wilcoxon Rank-Sum test on RMSE pointed out that all models except GB performed significantly better than a random baseline ($p_{all} < 0.05$). We found that Lasso is the best performing model, while the $R^2 = 0.33$. This was at odds with a comparable model in the recipe and movie domains from Trattner et al. [16], where Ridge Regression performed best, while the model accuracy was somewhat higher in the recipe domain ($R^2 = 0.51$). This suggested that the similarity functions adapted from [16] were less representative for user similarity judgments in the news domain.

Feature-specific Models and User Characteristics. To further explore [RQ1.2], Table 5 (ii) describes the performance of feature-specific models. To compare our findings to other domains, Ridge regression was used to combine multiple similarity functions per feature, while linear regression was used for features with a single function. Although the representativeness of the

different BodyText similarity functions varied (cf., Table 4), it was the best predicting feature, even outperforming the *All features* model.

Finally, we included user characteristics and demographics in our Ridge model. We tested the impact of each additional feature separately, as well as simultaneously. Table 5 (iii) outlines that the addition of user characteristics (e.g., news consumption frequency) hardly affected the model's predictive quality. A model that included the user's *age* reported the lowest RMSE, but this decrease (from 0.9141 in (i) to 0.9081 in (iii)) was not statistically significant different according to a Wilcoxon Rank-Sum test.

4.3 Conclusion

We assessed the representativeness of feature-specific similarity functions in the news domain. The functions used were adapted from recommender literature

Table 5 Model accuracy of different learning approaches, predicting a user's similarity judgment in the news domain. We compare (i) models averaged across all features in the news domain, (ii) describe the accuracy of feature-specific models, and include (iii) user characteristics. The best performing models are denoted in bold.

	News Articles ($N = 2,169$)		
Method	RMSE	R^2	MAE
(i) Model performance (All features)			
All (Random Forest (RF))	0.9219	0.2982	0.7643
All (Gradient Boosting (GB))	0.9177	0.3123	0.7520
All (Ridge Regression)	0.9141	0.3257	0.7459
All (Linear Regression)	0.9120	0.3289	0.7453
All (Lasso Regression)	**0.9101**	**0.3339**	**0.7480**
Mean	0.9652	0.0000	0.8122
Random	0.9659	-0.0226	0.8125
(ii) Regression model per news article feature			
Subcat (Linear)	0.9554	0.1406	0.7943
Title (Ridge)	0.9618	0.0889	0.8071
Image (Ridge)	0.9548	0.1495	0.7913
Author (Linear)	0.9568	0.1333	0.7991
Date (Linear)	0.9616	0.0911	0.8070
BodyText (Ridge)	**0.9141**	**0.3244**	**0.7514**
AuthorBio (Ridge)	0.9561	0.1414	0.7991
(iii) All (Ridge) + Additional User Characteristics			
News website visits	0.9164	0.3207	0.7463
Num. days reads news	0.9186	0.3215	0.7476
Gender	0.9125	0.3314	0.7456
Age	**0.9081**	**0.3435**	**0.7338**
All additional features	0.9099	0.3412	0.7358

in the movie and recipe domains [16]. We found that most feature-specific similarity functions only partially reflected a user's similarity judgment, yielding modest correlations. To best reflect user perceptions, we suggest that content-based news recommender systems should exploit the body text, supported by image embeddings, article categories, and the author.

The representativeness of body text was grounded in the reported feature use, as well as consistent with previous studies on news retrieval [1]. In contrast, although users used a news article's title in their similarity judgments, we have found title-based similarity functions to be hardly representative for these judgments. The weak correlations could be attributed to the relatively 'wordy' titles of news articles. At the similarity function level, it is possible that the string-based functions do not capture more subtle similarities between news articles, for example if two headlines describe an identical news event, but from a different news angle. Moreover, the insignificant correlation between *Title:LDA* and a user's similarity judgment suggests that word-based similarity is unrelated to how users perceive a pair of news articles.

In terms of *predicting* similarity judgment, we used machine learning to determine model accuracy and feature importance. In doing so, we examined the predictive value of additional user characteristics. We found that the addition of user characteristics and demographics in our models does not significantly improve the accuracy indicators, indicating there is little variance across users. In terms of similarity modeling, these findings suggest that the main focus should be on leveraging a news article's *BodyText*, while other features should only be used if the similarity functions would be more accustomed to the news domain.

4.3.1 Improvements for subsequent studies

A notable limitation of our approach is the use of a single dataset, which only comprises political articles. It is possible that the relation between similarity judgments and feature-specific similarity functions would be affected when employing additional main categories. For example, 'name-dropping' sports teams in a news article title might result in a higher feature importance for news article titles, compared to 'political judgments' [33]. Furthermore, the news articles shown to users were a few years old, which might have reduced familiarity levels and, in turn, decreased similarity ratings. Hence, in Study 2 and 3, we considered domain or category differences within news as a factor, addressing [RQ2].

Another limiting factor in Study 1 was the random pairing of news articles. Our setup was not necessarily representative of a recommender system scenario, as some of the presented news article pairs were likely to be completely unrelated to each other. Hence, it could have been difficult for users to judge the similarity for some pairs. In Study 2 and 3, we examined multiple factors on how news articles could be paired, and compared similarity judgments given to a news articles that were paired in such way to news articles that were deemed dissimilar, addressing [RQ3]. However, we would first need

to confirm what the best approaches to such matching would be, which was explored in Study 2. This was based on previous work and the intuition that news articles matched on topic, named entities and date would provide a more representative similarity judgment scenario.

5 Study 2: Examining Domain Differences in News

Study 2 had two goals. First, similarity functions might have different correlational strengths with human judgments depending on the category of the article. For example, some features or similarity functions might be more important or more effective in news articles about 'Sports' than in news articles about 'Politics' [RQ2]. However, related work on this topic revealed little evidence that could support this assumption. Second, we wished to either confirm that the similarity factors described in the related work section, topic, named entities, and date, were indeed the most important ones, while also exploring possibly overlooked factors [RQ3].

We describe Study 2 based on the order in which the survey was administered. We first discuss the methodology, detailing the three key questions posed, which concerned a user's expectations regarding similar-item retrieval, domain differences, and main similarity factors.

5.1 Methodology

5.1.1 Procedure

We set up a survey that was distributed among crowd workers. They were initially questioned about their gender, age, and the frequency with which they read online news articles per week. Following this, participants were asked three open-ended questions on news recommender system methods and similarity.

Similar Item Retrieval

The first question inquired on the type of recommendations that people would like to receive. It was formulated as follows: *'News recommenders are encountered on news websites, where they suggest articles to you that you might be interested in reading next, after you have a finished reading a news article. We want to know more about your thoughts on what information should be used for such news recommendations. Imagine that you have just finished reading an article, and you reach the list of potentially interesting articles for you to read next. What do you think should be the criteria for an article to appear on this list?'*. Based on this question, we aimed to gain insight on what users perceived as good recommendation methods, regarding similar-item retrieval.

Domain Differences

The second question examined to what extent users had different expectations regarding The writing styles of articles in the 'Sport' category and the 'Recent

Events' categories are typically distinct [57]. As its primary purpose is to inform readers, 'Recent Events' articles are often written in a straightforward syle, while 'Sport' articles tend to both inform and entertain. The latter is typically reflected by use of more colorful and lively language.

The second question was formulated as follows: *'Given the news article above, give us a short description of a either made up or real news article you would consider to be very similar.'*. Participants were presented one of the two news articles depicted in Figure 5, thus either belonging to the 'Recent Events' or 'Sports', which were tags given to these articles by the Guardian.

Fig. 5 Illustration of the two BBC news articles used for Question 2 in Study 2. Participants were only shown one of these two news articles.

News Similarity Factor

The objective of the third question was to determine which characteristics the readers consider most important when evaluating article similarity: *'When comparing two news articles, what is to you the single biggest factor that determines whether they are similar?'*. This would provide evidence as to which factor seemed to be the most important when assessing similarity.

5.1.2 Research Design

The second question in the survey was subject to a single-arm between-subjects design. It showed either a news article about Boris Johnson from the 'Recent Events' category, or a news article about Sadio Mané from the 'Sport' category (cf. Figure 5). The study was run in two batches, with the first half of the participants being presented the Recent Events article and the second half the Sports article.

5.1.3 Participants

A total of 45 participants (60% men) were recruited through Amazon Mechanical Turk. Participants were required to have a high approval rate across multiple tasks. The average completion time fell just below 5 minutes, for which they were compensated with $1 each.

Figure 6 reveals that the age distribution is skewed toward the younger end, as only six participants were aged 45 or older. Figure 7 shows that while 13 people read online newspapers daily, as many as 8 participants read online articles just one day a week or less. This was arguably surprising, as our sample was relatively young, assuming that the digital literacy of younger people was much higher [58].

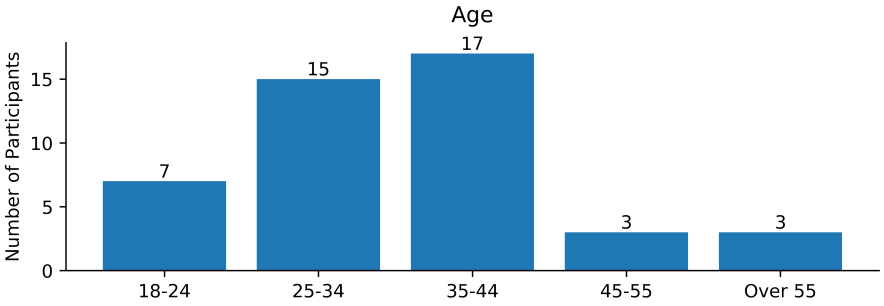

Fig. 6 A bar graph displaying the age distribution among the participants.

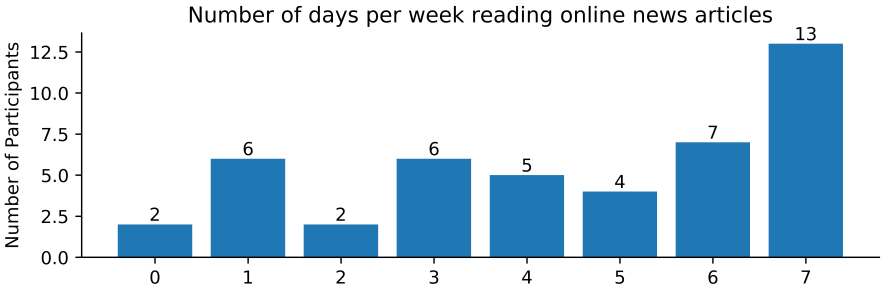

Fig. 7 A bar graph displaying how often the participants read online news articles per week on average.

5.2 Results

5.2.1 Criteria for news recommendation [Q1

We asked users to write down which criteria they found important in news recommendation. We categorized the qualitative responses, which is depicted in

Figure 8. It shows that a majority of participants stated that a recommended news article should be related to the article they just read, which was in line with general principles of news recommendation [1]. This was followed by 9 participants who indicated that selecting trending would be a suitable criterion. Seven individuals said that relevance to previously read articles would be a fitting criterion, while other responses were less frequent. Participants argued to consider the diversity among the suggested articles ($n = 5$), to recommend news that was geographically relevant ($n = 5$), to consider news that was of significant importance the public ($n = 5$), and to consider recency ($n = 4$).

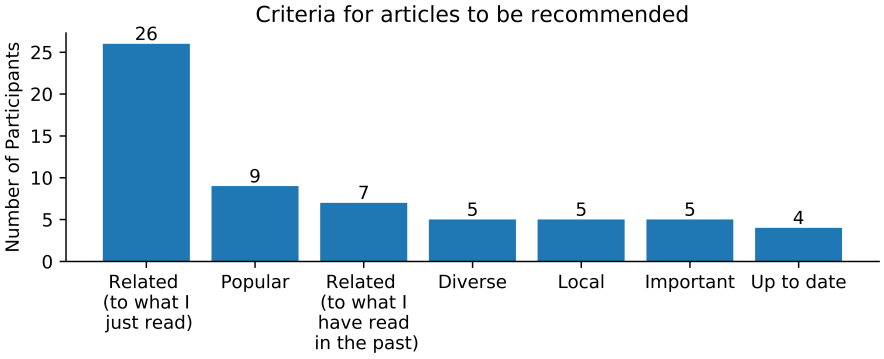

Fig. 8 A bar graph displaying the criteria that individuals believe should be considered by news recommendation systems. Note: People were not confined to a single response; they could list as many factors as they wanted.

The results suggested that similar-item retrieval was suitable for news recommendation. It was also notable that participants argued for factors highlighted in the related work section, such as diversity and recency. Overall, the findings here were in line with the literature and did not affect the design of Study 3.

5.2.2 Sport vs. Recent Events [Q2]

A main difference regarding the use of named entities was revealed across the two news categories. Participants who responded to the 'Recent Events' about Covid and Boris Johnson, all proposed to present Covid-related news articles as follow-up. In contrast, Boris Johnson was not mentioned by any of the participants. Participants who commented on the 'Sport' news report regarding Sadio Mane's injury responded differently. The majority of responses were either about football player Mané or an entity directly related to him, such as his teammates or manager. A few participants even suggested that a news article about other football players would be similar, whereas others wrote about different sports, such as rugby.

This illustrated a difference in terms of what the main topic was. Suggestions for follow-up articles in the 'Recent Events' domains all concerned Covid,

while the article also featured Boris Johnson, the British prime minister at the time of the research, discussing how he addressed the issue. Yet, none of the participants deemed that to be sufficient related and focused on the topic or news event at hand. In contrast, the suggested fictitious similar news articles concerned persons, such as Sadio Mané himself, straying further away from the original article in terms of the suggested article. None of the respondents mentioned injures or other medical conditions; they all focused on other named entities.

Considering these articles to be representative of the 'Recent Events' and 'Sports' categories, it could be argued that people perceived similarity differently across both categories. Regarding the design of Study 3, we expected users to perceive similarity different across different news article categories, a factor that would hardly be considered by the used similarity functions. This would be consistent with the different follow-up similar news articles mentioned by the participants.

5.2.3 Main Similarity Factors [Q3]

The coded versions of the responses are depicted in Figure 9. Valid responses were cited by 29 of 30 participants as indicating that a news article's topic is the most influential aspect when assessing similarity between news articles. According to five respondents, the title was the most critical feature, while two participants cited the keywords. Two others indicated named entities as the most significant indicator of similarity, while one individual identified journalistic quality as the most significant factor.

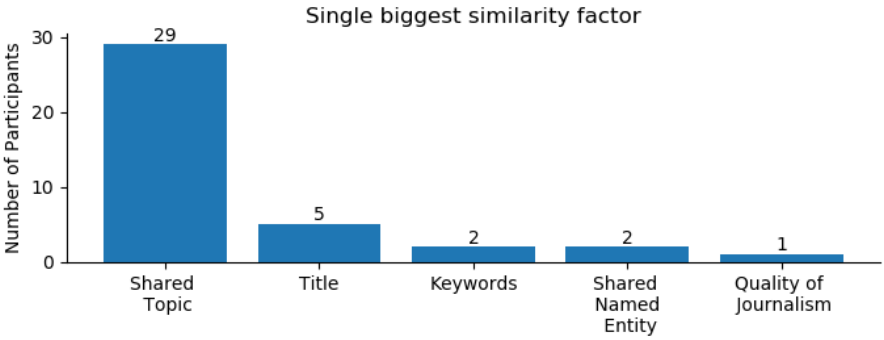

Fig. 9 A bar graph displaying what people considered to be the single biggest factor that determines similarity between articles. Note: Six responses were omitted because it was impossible to discern with a high degree of certainty exactly what the respondents meant.

The findings were consistent with the related work. Out of the three similarity factors chosen based on the literature, two of them - shared topic and shared named entities - were mentioned as the single most important factor by participants. Articles with similar publication dates was, however, not mentioned.

5.3 Conclusion

Study 2 explored which factors contributed to similarity, and whether people have different perceptions of what similarity would entail. Although none of the results indicated that 'Shared publication' would an important matching characteristics, we will consider it in Study 3, alongside 'Shared Topic' and 'Shared Named Entity'.

Study 2 also supported the notion that there could be a difference in between what people think similarity entails, depending on whether the reference article belongs to Sport or Recent Events. Because of this, it was decided to not just use articles from 'Recent Events' in Study 3, but to also use news articles from the 'Sport' category. The comparison between the two categories would then be examined further.

6 Study 3: Assessing Similarity based on different Matching Characteristics

The design of Study 3 is consistent with Study 1, but improved based on the insights from Study 2. Again, participants are asked to assess the similarity of news article pairs. In contrast to Study 1, news articles in Study 3 are not paired randomly, but based on similarity across three different characteristics or criteria: topic, date-time, and named entities, examining [RQ3]. In addition, following up on the findings in Study 2, we also explore differences across two news category domains: 'Recent Events' and 'Sport' [RQ]).

6.1 Methods

6.1.1 Obtaining Data

We obtained a dataset of 385 articles from the British newspaper The Guardian[4]. These were published between 2019 and 2021. The selected news articles were not selected randomly, but were required to attain to a number of criteria.

Familiarity

To assure the quality of responses when individuals are asked to rate the similarity between news articles, it was important that they would be familiar with the news articles. Hence we avoided niche articles that covered events and topics unknown to most people. For instance, articles about Formula 1 were fine as this was a popular sport, while articles about the sport of orienteering were not.

Recency

The vast majority of news articles that are read online are rather recent. To create a naturalistic environment, it was determined that the dataset should

[4] https://www.theguardian.com/

not contain old articles. To this end, 2019 was chosen as the cutoff point going back in time. Regarding recency, we opted to avoid news articles that were too recent, to avoid any impact because of that factor. As a result, no articles from 2022 were included, noting that data was collecting in September 2022. Note that although Covid-19 was part of Study 2, news articles about this topic were omitted from the dataset, as it could form a category of its own.

News Article Selection Procedure

News articles were selected from two main categories of the Guardian: 'Recent Events' and 'Sport'. Each time, a subcategory of one of the two main categories was selected at random, such as 'UK Politics' as a subcategory of Recent Events. Then, a random date between 2019 and 2021 was chosen, after which a reference article that met all of the aforementioned requirements would be selected. Then, five to ten articles that were within two weeks of the reference article in terms of publication date would be picked along with it, for the purpose of similar-item retrieval. Afterwards, the procedure would be repeated with picking a new date each time, up to five times.

After completing this for a single subcategory, the entire process would be repeated. The number of articles from each subcategory fluctuated, since some featured a wide variety of diverse topics, resulting in more articles, whilst those with less diversity ended up with fewer articles, ensuring that no single topic was over-represented. Once a sufficient number of articles had been obtained, the process was concluded.

6.1.2 Extraction of Named Entities

Due to the possible inconsistencies in writing styles, it was determined that named entities would be extracted manually. For example, we wanted to avoid that 'Trump' and 'Donald Trump' would be considered different named entities. The following were the selection criteria for extracting named entities:

- Named entities were defined as a real-world object that can be identified by a proper name, such as a person, place, organization, or product.
- Named entities were extracted from the headline and subheading of articles.
- Any sequential capitalized words will be part of the same entity (Monaco Grand Prix is a single entity)
- If countries were used as an adjective (e.g., 'A Russian man'), the name of the country was extracted instead (Russia).
- The full name of people would be used, even if only the surname was used in a specific news article.

Data Structure. Information was stored for a limited number of features. This included: *ID, Category* (Recent Events or Sports), *Topic, Title, Subheading, Main Image, Image Caption, Body Text, Date, Time, Day of the Week, Author, Author Bio, Article URL, and Named Entities.* This structured approach ensures precise querying and enhances data analytics, thereby increasing the usability of the stored information.

Table 6 The 12 conditions from the 2x2x3 factorial design. Note: Dissimilar refers to the scenario where an article pair is matched on neither Date, Topic, or Named Entity.

Conditions
Sport: Dissimilar
Sport: Date
Sport: Topic
Sport: Topic + Named Entity
Sport: Topic + Date
Sport: Topic + Date + Named Entity
Recent Events: Dissimilar
Recent Events: Date
Recent Events: Topic
Recent Events: Topic + Named Entity
Recent Events: Topic + Date
Recent Events: Topic + Date + Named Entity

6.1.3 Similarity Functions

Most of the similarity functions mentioned in the General Method section were also used in this study. This allowed us to compare the results of Study 3 to the results of Study 1. However, as there were some computational issues, we only used Image embeddings as an image-specific similarity function. Since this was the best performing function in Study 1, we felt this was justified.

6.1.4 Research Design

How news articles were paired and whether they matched was subject to four factors: Category, Date, and Topic and Named Entity. This was operationalized in a 2x2x3-Within Subjects design, which is described in Table 6.

News articles were either similar in terms of the published date or not. The cutoff was set at 14 days, indicating that articles published within 14 days of one another were similar in terms of data. In contrast, any pair of news articles separated by more than 28 days were deemed dissimilar, while those in-between were considered neutral.

Topic similarity was assessed by the labels available in the Guardian dataset. For example, news articles would labelled with 'Brexit' or 'Formula 1' to indicate its topic. If the pair of articles cover the same topic, they are regarded similar. A comparable rationale was used for Named Entities and category: if two articles shared a named entity or were from the same main category (i.e., 'Recent Events' or 'Sport'), they were considered similar. Note that Named Entities were considered to be tight to topic similarity, as we could not think of cases where Named Entities would intentionally match, while the topic was dissimilar. Hence, topic similarity comprised three levels: baseline, topic similarity, topic similarity and named entity similarity.

6.1.5 Dataset

Eventually, a total of 385 news articles were obtained. The process obtaining news articles had stopped once it was possible to divide the set of news articles into 60 groups of 12 pairs, with one unique pair from each of Table 6's 12 conditions in every group. Any given article could appear in a maximum of two different article pairs, but never in two article pairs from the same group.

6.1.6 Procedure and Measures

Participants were invited to use our web-based experiment[5], which was built from scratch. They would be randomly assigned to any of the news article pair sequences. After having read the consent form, they would proceed to the second phase, depicted in Figure 10. This involved rating article pairs based on their degree of similarity. Users were asked to evaluate 13 different news article pairs, one for each of the twelve conditions listed in Table 6, plus one pair that served as an attention check. For all article pairs, participants were asked to rate their similarity on a 5-point Likert scale. On top of that, they were asked to indicate how confident they were in making this similarity judgments, as well as to indicate their familiarity with either news article.

After evaluating 13 news article pairs, users were directed to the third phase of Study 3. Participants were shown a picture of a news article, in which its features were pointed out: Category, Title, Subheading, Image, Author, Date of Publication, and Body of Text. Participants were then asked to indicate for each of these features, how important they were for making their similarity judgments, measured on 5-point scales. Finally, users were asked to answer a number of final questions. This included questions on the frequency of their online news reading behavior, their level of education, and other basic demographic questions.

6.1.7 Participants

Participants were recruited on the crowdsourcing platform Prolific. The quality of responses was expected to be higher than those obtained from Amazon MTurk, in terms of attentiveness, comprehension, and reliability [59]. Only workers with an approval rating of 99% were invited. Only participants based in the United Kingdom were invited, as we used news articles from the English news website The Guardian. The median completion time was 14 minutes and 20 seconds, while the participants received 2.25 pounds for their work.

Eventually, we recruited 173 individuals, resulting in 2,076 evaluated news article pairings. The attention check was passed by 65 percent of the participants. When excluding individuals who failed the attention check, we were left with 1356 news article pair ratings. Since the attention check was perhaps difficult to spot and we eventually found only minor differences between the two groups, we used the sample of all participants for most analyses. Note that

[5]The study's web application's source code is available in a Github repository containing all pertinent code for this paper https://github.com/VRS-MT.

Fig. 10 Illustration of the second phase of the user study.

t-tests on the demographics of the participants did not reveal any significant differences between those who did and did not pass the attention check.

Among the 173 participants (38% Male), the age distribution was fairly diverse, as depicted in Figure 11. Just over half of participants had attained at least a Bachelor's Degree. The participants' self-reported assessments of news consumption patterns is depicted in Figure 12. The most frequent response was that a person would read online publications seven days per week, which accounted for around 35 percent of responses. However, nearly 30 percent of users reported reading articles on two or fewer days per week.

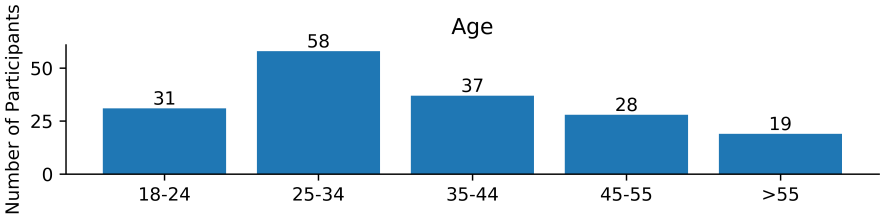

Fig. 11 A bar graph displaying the age distribution among the participants.

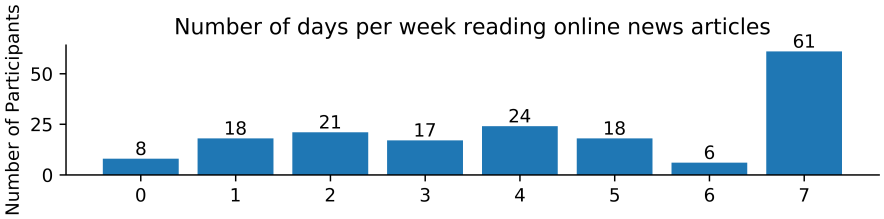

Fig. 12 A bar graph displaying how often the participants read online news articles per week on average.

6.2 Results

We present four sets of analyses. We first confirmed our findings from Study 1, again examining cue usage and the representative of feature-specific similarity functions for human judgment [RQ1.1]. Then, we examined differences in our findings across the 'Recent Events' and 'Sport' news categories [RQ2], continuing with out analysis on matching characteristics [RQ3], examining to what extent both similarity judgments and scores are affected by matching characteristics on topic, named entities, and date.

6.2.1 Information Cue Usage [RQ1.1]

Participants were asked to rank the importance of different news article cues or features when determining similarity between articles. On a 5-point scale, we

found that Body of Text had the highest importance score (M=4.17), which was followed by Title (M=4.04). These findings were consistent with Study 1 (cf. Figure 3). To assess whether the differences in cue use were significant, we performed an one-way ANOVA on all of the research's conditions, including a Tukey's HSD post hoc test. The results are depicted in Figure 13, illustrating that the difference between Body of Text and Title were not significant different. However, their use was significantly higher than use of the subheading (M=3.50) and Topic (M=3.37). We further found that the reported use of Image (M=2.92), Date (M=2.42), and Author (M=1.76) was much lower, indicating that even if similarity functions for these functions would reflect user judgment, users pay little attention to these features.

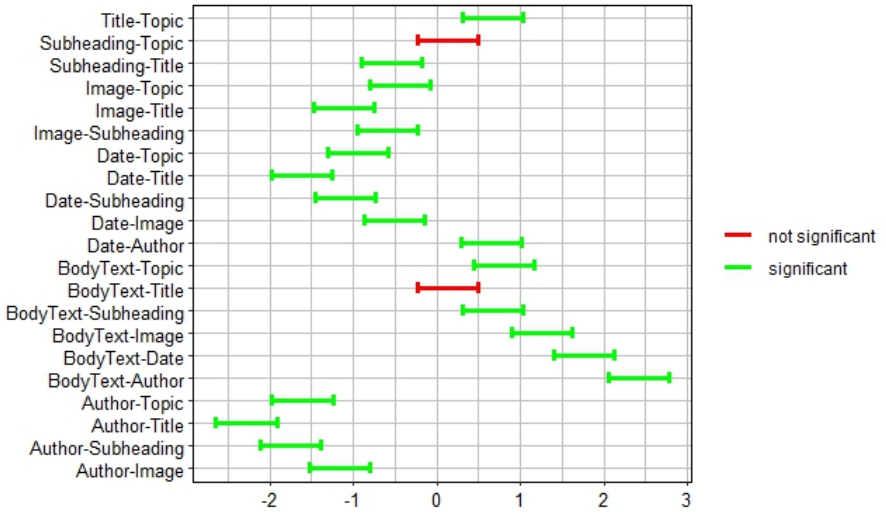

Fig. 13 Tukey-HSD post hoc test result for information cue usage.

6.2.2 Similarity judgment and Function Correlation [RQ1.1]

Similar to our analysis in Study 1, we examined to what extent different feature-specific similarity functions were representative of human judgment. We first did so by disregarding the similarity matching conditions. Table 7 outlines the results for the different functions used, differentiating between a sample of all participants and those who passed the attention check. It also outlined the correlational strengths for judgments that were given with a high level of confidence. Overall, it seemed that BodyText:TF-IDF ($\rho = 0.52$, $p < 0.001$) had the strongest correlation, which was in line with the findings of Study 1, but somewhat stronger. Although Body Text was an appropriate feature to use, the similarity function that leveraged a text's sentiment was less representative ($\rho = 0.07$, $p < 0.01$). The second strongest correlation was found in Topic:Jaccard ($\rho = 0.45$, $p < 0.001$), which signalled a potentially significant

role for shared topics in article recommendation. Title:BI ($\rho = 0.30$, $p < 0.001$) had the third strongest correlation overall, deeming it the best title-based function. In contrast, other functions were much less representative, although almost all correlations were stronger than those found in Study 1 (cf. Table 4).

Table 7 Spearman Rank Correlations (ρ) between two similarity representations: Feature-specific similarity judgments and human judgment. The 'All' column takes all ratings and similarity scores and compares them, testing for significance. The 'Pass: Diff.' column denotes the correlational strength for participants who passed the attention check, along with the difference with the 'All' column. Similarly, the 'HiConf: Diff.' column does so for observations that were made with a confidence score of five, where significance levels indicate a significant change from the strength in the 'All' column. Differences in correlations were tested using Fisher r-to-z transformation to produce a z-value: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Function	All	Pass: Diff.	HiConf: Diff.
Topic:Jacc	0.45***	0.44: -0.01	0.56: 0.11***
Title:LV	0.10***	0.10	0.13: 0.03
Title:JW	0.15***	0.13: -0.02	0.14: -0.01
Title:LCS	0.19***	0.18: -0.01	0.20: 0.01
Title:BI	0.30***	0.30	0.40: 0.10**
Title:LDA	-0.031	-0.05: 0.02	-0.06: -0.03
Subheading:BI	0.12***	0.14: 0.02	0.10: -0.02
Subheading:LCS	0.14***	0.13: -0.01	0.16: 0.02
Subheading:TF-IDF	0.21***	0.21	0.30: 0.09*
Image:EMB	0.11***	0.12: 0.01	0.10: -0.01
Date:ND	0.046*	0.03: -0.02	0.11: 0.06
BodyText:TF-IDF	0.52***	0.53: 0.01	0.65: 0.13***
BodyText:LDA	0.17***	0.16: 0.01	0.26: 0.09*
BodyText:Senti	0.071**	0.08: 0.01	0.09: 0.02
Author:Jacc	0.26***	0.25: -0.01	0.35: 0.09 *
AuthorBio:TF-IDF	0.22***	0.21: -0.01	0.30: 0.08*
AuthorBio:LDA	0.21***	0.18: -0.03	0.28: 0.07

Table 7 further shows little changes between the sample of 'All' participants and the 'Passed attention check' sample. Hence, we continued using the 'All' sample for subsequent analyses. Noticeable was, however, that judgments made with a self-reported confidence level of 5 out 5, significantly improved the representativeness of a few similarity functions. In particular, the correlation between the best performing function BodyText:TF-IDF and human similarity judgments increased from $\rho = 0.52$ to $\rho = 0.65$ ($p < 0.001$). Such improvements were only observed for functions that already had a relatively strong correlation. This suggested that, on the one hand, it mattered whether participants could confidently rate article pairs, but also that, on the other hand, functions would also need to be representative across all individuals.

6.2.3 The Influence of News Categories [RQ2]

We evaluated the representativeness of feature-specific functions across different news categories. We computed Spearman correlation for Sport and Recent Events separately, performing Fisher r-to-z transformations to examine whether the respective correlation coefficients differed significantly. Table 8 describes the results, showing that five functions performed better in the Recent Events domain, while two functions did so in the Sport domain.

The largest difference was observed for Title:LV ($\rho = 0.00$ vs $\rho = 0.19$, $p < 0.001$), suggesting that this title-method was not representative for Sports news articles. Two other notable differences were found for BodyText:TF-IDF ($\rho = 0.48$ vs. $\rho = 0.56$, $p < 0.01$), and Topic:Jacc ($\rho = 0.41$ vs. $\rho = 0.49$, $p < 0.05$), which were the best performing functions. They seemed to be performing better in the Recent Events domain than for Sports articles, suggesting that these methods may depend on the type of news article content. The two functions that performed better in Sport, Image:EMBand Subheading:TF-IDF, reported rather weak correlations to start weak with, making their improvements significant, but not as representative as some other functions.

We also tested our results when only including ratings that were given with a high level of confidence. Again, we observed improvements in the correlational strength across most functions, in line with the non-category specific findings reported in Table 7.

6.2.4 The Role of Matching Characteristics [RQ3]

ANOVA and Post-Hoc Tukey

We examined whether matching news article pairs based on specific matching characteristics would affect similarity scores and judgments. We first examined differences in human similarity judgment across all conditions (Topic, Date, Topic+Entity), performing two one-way ANOVAs per news category (Sports and Recent Events) on human similarity judgments. Figure 14 reports the results of Tukey's HSD post-hoc tests between each of the matching characteristic conditions.

Our findings suggested that most of the matching characteristics positively affected human similarity judgments. Starting at the top-2, matching news articles on topic significantly affect similarity judgments. The difference between 'Dissimilar' and 'Topic' was significant across both domains, with larger differences for Recent Events. This was in line with our finding that many similarity functions were more representative of human judgments in the Recent Events domain, such as Jacc:Topic. Adding a match in 'Named Entities' to a topically matched news article pairs did not significantly increase human similarity judgments in either domain. While 'Named Entity' and matching on date did not affect similarity judgments on their own, there was a significant difference between 'Topic' and 'Topic + Date + Named Entity' across

Table 8 Spearman correlations between two similarity representations: Feature-specific similarity functions and human similarity judgments. All correlations are divided across news categories (Sport, Recent Events). All participants were included in this analysis. The Difference column denotes the difference between the two categories. Note that significance is only reported for the difference, not the correlational values per category. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Function	Sport (All)	Rec. Events (All)	Diff.
Topic:Jacc	0.41	0.49	0.08*
Title:LV	0.002	0.19	0.19***
Title:JW	0.15	0.15	
Title:LCS	0.15	0.21	0.06
Title:BI	0.30	0.31	0.01
Title:LDA	-0.04	-0.02	0.02
Subheading:BI	0.14	0.11	-0.03
Subheading:LCS	0.10	0.18	0.08*
Subheading:TF-IDF	0.24	0.17	-0.07*
Image:EMB	0.15	0.07	-0.08*
Date:ND	0.04	0.04	
BodyText:TF-IDF	0.48	0.56	0.08**
BodyText:LDA	0.12	0.21	0.09*
BodyText:Senti	0.05	0.10	0.05
Author:Jacc	0.27	0.27	
AuthorBio:TF-IDF	0.23	0.22	-0.01
AuthorBio:LDA	0.22	0.20	-0.02

both domains. As shown in Figure 14, this difference larger than what would be obtained by merely stacking the individual main effects.

The findings also suggested that matching news articles only on 'Date' had little influence on similarity judgments. When adding that characteristic to either 'Dissimilar' or 'Topic', it did not significantly increase similarity judgments. In other words, the proximity of two articles' publication dates did not appear to have much influence how similar readers perceived them to be.

We performed a similar examination of the similarity score for the best-performing similarity function. We examined changes in the similarity scores of BodyText:TF-IDF using two one-way ANOVAs, depicting the results of the Tukey's HSD post-hoc tests in Figure 15. It shows that this similarity function was more sensitive to any matching factor than our sample of humans. While, again, we observed no significant difference between 'Dissimilar' and 'Date' and mixed results for the use of Named Entities, all other differences between condition pairs were significantly. Similarly, almost all of the matching factors led to higher similarity scores, while named entities yielded higher scores than date similarity. Although the exact results may depend on the use

of this specific similarity function, the results were consistent with other representative similarity functions (i.e., those that had moderate to high correlation with human judgment).

Multiple Linear Regression Analyses

We further examined the influence of different matching characteristics, by comparing the model fits of different linear regression models. Table 9 describes the results of our analyses, in which each line represented a regression model with three dichotomous predictors: Match in Topic, Named Entity, and Date. We first predicted human judgments (denoted by 'Human:'), after which we predicted the similarity scores of feature-specific functions, separating Title:Bigram and BodyText:TF-IDF across the two news categories. different news categories, as well as humans and similarity functions, first for human judgments, then for all similarity functions across all news categories, where we differentiated between Recent Events and Sport for the two best performing functions: Title:Bigram and BodyText:TF-IDF.

Regarding similarity judgments made by humans, we observed that the R^2 of the model with all observations was 0.205. This model showed that all matching characteristics positively affected the strength of the similarity judgment (All: $p < 0.01$). Based on the reported standardized β-coefficients

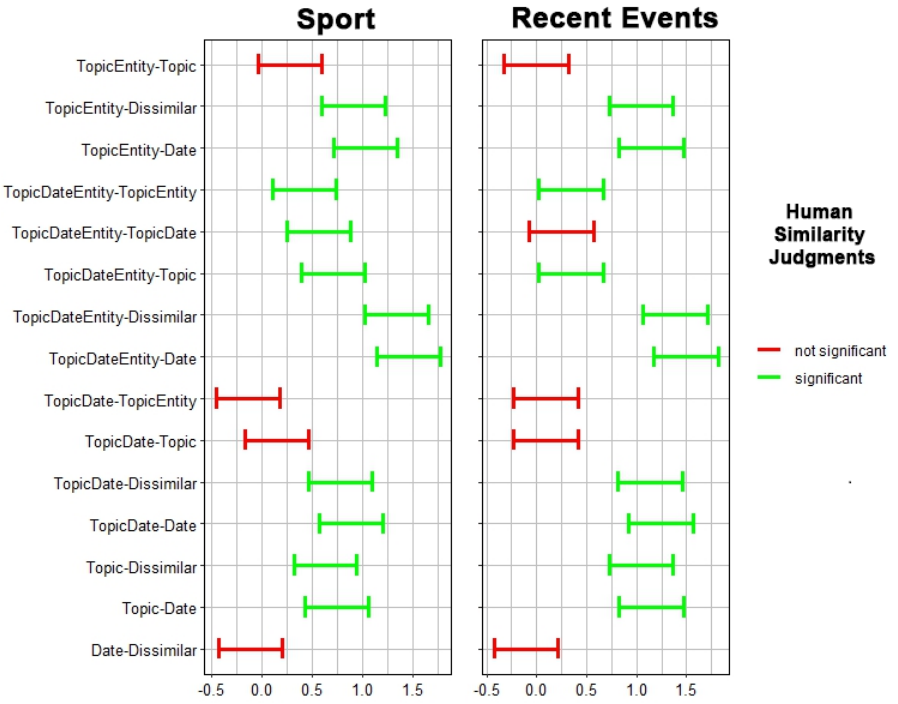

Fig. 14 Tukey-HSD post hoc test result for human similarity judgment. News = Recent Events.

in Table 9, it was apparent that news articles matched on topic led to the largest increase in similarity judgments, followed by named entity and date. In line with our previous findings in Study 3, we found the model to be more accurate when only considering judgments with a high confidence level (R^2=0.316)[6]. When examining differences across news categories, we found the Recent Events model (R^2=0.226) to be more accurate than the sport model (R^2=0.189) due to the large impact of topic matching, while the sport similarity judgment was positively affected by all characteristics.

Table 9 further reports how the three matching characteristics impacted the similarity scores different feature-specific similarity functions. For two functions that were most representative of human judgment (cf. Table 7), Title:Bigram and BodyText:TF-IDF, we also differentiated findings across two news categories, in addition to using all observations.

In general, all similarity scores were positively and significantly affected by matching news articles on Topic and Topic plus Named Entity, Except for AuthorBio, BodyText:LDA and BodyText:Senti, compared to dissimilar pairs. In contrast, the impact of matching news articles on date showed mixed results, as it only had a positive impact on similarity score for most Title-based and BodyText-based functions. As the matching characteristics had positive and

[6]We analyzed this for all models reported in Table 9 and found this to apply to nearly all of them. For the sake of brevity, we only reported the Human:HiConf.

Fig. 15 Tukey-HSD post hoc test result for BodyText:TF-IDF.

significant impacts for both human judgment and similarity functions, this indicated that both humans and retrieval algorithms were sensitive to news articles being matched topically; a factor not accounted for in Study 1.

In terms of model fit, Title:BI and BodyText:TF-IDF performed best, which was also in line with their strong correlation with human judgment. For these functions, we observed that the Sport-based models (R^2 of 0.310 and 0.429, respectively) were more accurate than the Recent Events models (R^2 of 0.171 and 0.308, respectively). For these specific functions, it was apparent that matching Sports article pairs on Named Entities had a larger impact on similarity scores, than matching on Topic only. While this contrasted with the models for Recent Events for these two functions, in which Topic matching had the largest impact, it was consistent with the findings for most of the other feature-specific similarity functions reported in Table 9. For, among others, Title:LV, Subheading:LCS, and BodyText:LDA, news articles matched on named entities led to a larger increase in similarity scores than matching on topic only.

Table 9 MLR table for human judgment and all functions, using the matching characteristics as independent variables. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

	Standardized β			
	Topic	Named Entity	Date	R^2
Human:All	0.383***	0.111***	0.056**	0.205
Human:RecentEvents	0.451***	0.049	0.046	0.226
Human:Sport	0.313***	0.176***	0.068*	0.189
Human:HiConf	0.410***	0.208***	0.095**	0.316
Title:LV	0.084***	0.212***	0.031	0.070
Title:JW	0.201***	0.088***	0.059**	0.070
Title:LCS	0.184***	0.209***	0.055**	0.120
Title:LDA	0.079**	0.073**	-0.029	0.018
Title:BI (All)	0.238***	0.303***	0.064***	0.226
Title:BI (Rec.Ev)	0.309***	0.156***	0.057*	0.171
Title:BI (Sport)	0.172***	0.445***	0.072**	0.310
Subheading:BI	0.102***	0.084***	-0.017	0.026
Subheading:LCS	0.143***	0.204***	0.039	0.092
Subheading:TF-IDF	0.163***	0.174***	0.029	0.085
Image:EMB	-0.081**	0.201***	0.022	0.031
BodyText:TF-IDF (All)	0.396***	0.256***	0.153***	0.348
BodyText:TF-IDF (Rec.Ev)	0.456***	0.132***	0.145***	0.308
BodyText:TF-IDF (Sport)	0.330***	0.401***	0.163***	0.429
BodyText:LDA	0.042	0.218***	0.093***	0.066
BodyText:Senti	0.020	0.053*	-0.042	0.004
AuthorBio:TF-IDF	0.370***	-0.043	0.027	0.122
AuthorBio:LDA	0.348***	-0.026	0.041*	0.114

Comparing Correlations across Dissimilar and Similar Pairs. Finally, we examined the influence of matching characteristics on the correlational strength between human judgment and different feature-specific similarity functions. To do so, we compared correlations across dissimilar and similar news article pairs, where 'Dissimilar' included pairs without any matching characteristics or those matched on date, as the latter was not significantly different from the former. In contrast, similar news article pairs included Topic, Topic + Named Entity, and Topic + Date.

The results are described in Table 10. Significance levels denoted differences across dissimilar and similar news article pairs within a news category. The correlational strengths for the dissimilar category were in line with those reported in Study 1 in Table 4, where the Washington Post Corpus was used to randomly generate news article pairs[7]. This suggested that the findings across Study 1 and Study 3 could be compared, despite the different datasets.

Regarding Dissimilar and Similar news articles, there were a few differences. Most similarity functions reported stronger correlations with human judgments if news article pairs shared a Topic. The differences were significant for a few functions across one or both news categories, including the representative functions Title:BI and BodyText:TF-IDF. This indicated that these functions detected similarity much like a human would, if there was any topic similarity to be involved. Across the board, this suggested that similarity functions performed better in contexts were a recommendation would make sense, for example in case of a topical match. This was also illustrated in Study 1 through Figure 4, which showed that users provided low similarity ratings for news article pairs that were predominantly not matched topically.

6.2.5 Conclusion

We examined whether similarity judgments made by human similarity scores of feature-specific functions was affected across different news categories and matching characteristics. In doing so, we replicated findings from Study 1 and examined whether they could be explained by our improved research design.

We confirmed our findings from Study 1, as cue usage for similarity judgment was found to be mostly title and body text-based. Moreover, the correlational strengths between feature-specific similarity functions and human judgments were comparable to Study 1 for dissimilar news articles, being relatively modest at best (ρ ¡ 0.3).

Regarding our extensions [RQ2, RQ3], however, we found that the strength of these correlations depended on two of the three examined matching characteristics. Both topic and named entities positively affected similarity judgments and the similarity scores of functions, indicating that such keyword-based similarity was detected by retrieval algorithms and humans alike. The correlations between similarity judgments and scores tended to be stronger when examining news articles that were matched on topic or named entities. This

[7]We also checked the results when only including judgments with a high level of confidence. This lead to stronger correlations across the board.

Table 10 Spearman correlations between two similarity representations: Feature-specific similarity functions and human similarity judgments (using all observations). The *z*-test column denotes whether differences between correlation strengths across dissimilar and similar were significant; this was tested using Fisher r-to-z transformation to produce z-values: $^{***}p < 0.001$, $^{**}p < 0.01$ $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

	Sport			—	Recent Events		
Function	Dissimilar	Similar	*z*-test	—	Dissimilar	Similar	*z*-test
Title:LV	-0.07	0.02		—	0.05	0.10	
Title:JW	0.06	0.09		—	0.07	0.01	
Title:LCS	0.02	0.09		—	0.00	0.12	*
Title:BI	0.03	0.21	**	—	0.04	0.19	*
Title:LDA	-0.04	0.02		—	-0.09	0.08	
Subheading:BI	0.05	0.10		—	-0.10	0.12	
Subheading:LCS	-0.02	0.05		—	0.00	0.10	
Subheading:TF-IDF	0.05	0.19	*	—	-0.05	0.08	
Image:EMB	0.11	0.15		—	0.20	0.06	*
Date:ND	-0.05	0.13		—	-0.09	0.09	
BodyText:TF-IDF	0.17	0.36	***	—	0.32	0.37	
BodyText:LDA	0.06	0.04		—	0.02	0.22	**
BodyText:Senti	0.00	0.06		—	0.06	0.06	
Author:Jacc	-0.08	0.13		—	0.14	0.23	
AuthorBio:TF-IDF	-0.04	0.13		—	0.13	0.18	
AuthorBio:LDA	0.11	0.08		—	0.06	0.20	*

suggested that similarity functions were more representative of human judgment in news retrieval scenarios, topically rather than randomly matched news articles would be more common on news websites. In contrast, we observed little impact of matching on date, which suggested that using such a strategy for news recommendation would not resonate with end users.

Regarding specific similarity functions [RQ1.1], we observed that BodyText: TF-IDF performed relatively well. This was consistent with the findings in Study 1, even though that two-thirds of the news articles were matched on either topic, named entities, or both. In addition, Title:BI and Topic:JACC were also found to be representative and tended to be positively influenced by the matching characteristics. Particularly Title:BI yielded low correlations in Study 1, but improved significantly for news article pairs that were matched topically, which would be common for a recommendation scenario.

7 Discussion

We have investigated to what extent similarity functions used in other similar-item retrieval domains reflect human judgments of similarity for news articles. We have specifically compared similarity functions used in Trattner and Jannach [16] for different movie and recipe features, to analogous features that

are relevant in the news domain, such as title-based or image-based similarity. In doing so, we have employed a semantic similarity approach to validate different feature-specific functions, while also examining which exogenous factors might affect the performance of these functions, in the form of news categories and matching characteristics.

7.1 Key Findings

The studies presented in this paper can be considered a chronological investigation into this topic. While Study 1 presents a first attempt at developing similarity functions for the news domain using human judgments, Study 2 and Study 3 extend the findings of Study 1 by addressing it shortcomings. Overall, we have found that most cosine-based and distance-based metrics only partially reflect a user's similarity judgment, particularly when considering news article pairs that are rather dissimilar. While these values are found to be modest at best in Study 1, they have improved in Study 3 when controlling for topically matched news articles. What stands out across all studies, however, is that user similarity perceptions are best reflected in the news domain by relying on a news article's body text, supported to a lesser extent by title and subcategories if news articles are matched on topic or named entities.

We have also, directly and indirectly, examined what specific aspects in the body text of a news article determine a similarity judgment. The findings in Study 2 and Study 3 show that of these differentiating factors is a news article's category, which we have examined across 'Recent Events' and 'Sport' categories. Whereas 'Recent Events' reflect general news articles that are, indeed, oriented around a news event, and seem to be most related to the underlying topic(s) involved, 'Sport' articles have been found to be more person or entity-driven. The recommendation scenario provided in Study 2 on 'Recent Events', i.e., Boris Johnson talking about covid-19 policies, has prompted users to expect additional news articles about covid-19 instead of Johnson. In contrast, our 'Sport' article on an injured Liverpool football player prompted users to expected follow-up articles about the player himself or his team, instead of other injuries. This is also reflected in the similarity scores and judgments: topical matching has had a stronger influence on 'Recent Events' compared to named entities, while this is found to be the opposite for 'Sport' news articles. It could, however, be that the way in which named entities have been extracted put 'Recent Events' articles at a disadvantage, because entities such as 'UK' and 'England' were also included, which was arguably irrelevant for the similarity judgments of UK-based users for British news articles.

The findings of Study 2 and Study 3 are also largely consistent. While Named Entity is shown to have a greater influence on similarity in the 'Sport' category, this does not apply for Topic. Despite the fact that Topic has a significant and positive influence, its impact is found to be larger in the 'Recent Events' category. This appears to be consistent with the findings in Study 2, in which participants tasked with describing a similar article to a reference article were considerably more ready to stray from the topic when the reference article

concerned sports. The minimal impact of matching news articles on date is also consistent across both studies, for this factor was hardly mentioned in Study 2 and a minimal impact on similarity judgments and scores in Study 3.

When addressing [RQ1.1] and [RQ3], it seems that the representativeness of how similarity functions are affected by matching characteristics is function-dependent. While BodyText:TF-IDF, which is rather representative for human judgments, is more strongly affected by topical matches, other functions seem to be more strongly affected by the presence of named entities. Our multiple linear regression analyses reveal that matching characteristics account for approximately 20% of the variance in human similarity judgment, while this is roughly 40% for BodyText:TF-IDF. In addition, the relative importance of named entities is larger in BodyText:TF-IDF.

Overall, while we have not optimized our similarity-based functions [RQ1.1] or models [RQ1.2], this is not the focal point of this paper. Instead, we have aimed to show how existing metrics would perform, as well as how they compare across domains. The news domain seems to require metrics that are less 'taste-related' than movies or recipes, but further research is needed to develop accurate ones, possibly by also using psychologically-grounded approach as done in other studies [13].

We have been unable to use the same dataset across Study 1 and Study 3. While this has made the comparison more difficult, particularly because also different crowdsourcing user bases have been used, we do have observed comparable results. When comparing the results in Study 3 for dissimilar news articles with Study 1 (e.g., on the correlations between human judgments and feature-specific functions), the correlation strengths are well aligned. Thus, it actually seems that our findings can be generalized beyond a single dataset.

7.2 Domain Differences and Confounding Factors

In line with [16], we have found further evidence that different domains call for different similarity functions [RQ2]. Whereas images are very important in recipe recommendations, their role seems to be negligible in news similarity assessments of humans. However, the promising results using text-based similarity metrics might also be applicable to other recommender domains. Although what specific text a user is attentive to might vary (e.g., an item's name, its description, etc.), recommender domains that do not solely rely on images or bullet point features (e.g., electronics) can to a large extent represent the similarity judgments of its user using text-based methods. Moreover, our findings in Study 2 (regarding [RQ2]) show that even within a single recommendation domain (i.e., news), subcategories can affect the appropriateness of recommendation approaches, for named entities are found to be more appropriate for retrieval in 'Sport' articles.

We have further shown that confounding variables can affect the outcomes of semantic similarity studies. A user's level of familiarity can improve a model's accuracy when predicting similarity judgments, while users making similarity assessments with a high level of confidence seems to be more in line

with similarity functions. As this particularly applies to the news domain and not to other recommendation domains [16], this implies that the current similarity functions are surpassed by other variables with regard to representing human similarity judgments.

7.3 Similarity as a Metric for News Recommender Systems

The main assumption of this study is that similarity is a representative criterion for the evaluation of news recommender systems. Even though our studies 'second guess' similarity by inquiring on human judgments, it does posit that inter-article similarity is important.

Our assumption stems from various news platforms and builds upon multiple news recommender studies [1]. It particularly falls in line with 'more like this' sections on news platforms that are content-based driven. Due to the lack of users logging in and relatively fast churn of items [5], we believe that content-based recommendation will be an inherent part of the future of news recommendation.

Additional factors may complement this similarity. For one, user satisfaction and retention may be perceived as more important for news platform with a high degree of subscribed traffic. Platforms are also expected to present or facilitate the reporting and promotion of recent news events, in addition to more topically related content. Moreover, news platforms that seek to promote other specific values beyond 'breaking news', such as amplifying marginal voices [60], may supersede similarity metrics in terms of importance.

7.4 Limitations

The current study relies on correlations and linear regression analyses to support its conclusions. This prevents us from establishing causal relations, such as whether specific item features positively affect or diminish a user's similarity assessment. However, the goal of this paper is to create a broader understanding of whether similarity functions from one domain can be applied seamlessly in another, for which correlations are sufficient. Nonetheless, it would be valuable to examine a more longer-term impact of user evaluation on news evaluation, expanding on the session-based work done in Trattner and Jannach [16]. A session-based recommender scenario as a follow-up to this study would be a good start, to compare our findings to [16]. One of the future objectives should be to investigate whether news articles retrieved using similarity functions, that have been found to be most representative of human judgment, are also perceived as satisfactory to use.

A main shortcoming for Study 1 and Study 3 is that it is not entirely clear on what grounds users have made their similarity judgments. Although Study 2 provides some suggestions as to what users expect when it comes to news recommendation, similarity judgments on a 5-point scale, asked using a single question, might not sufficiently capture individual differences and

interpretations of that questions. Some other studies have also used multiple questionnaire items to circumvent this issue [2]. However, our inquiry on reported feature use by participants [RQ1] reveals a part of the underlying cognitive process, and suggests what are good features to optimize for. In fact, this is also a new finding. Moreover, it seems that differences in news categories should be examined further, as little is known about this in related work.

Contrasting with Study 1, Study 3 has been performing in a rather controlled setup, using a small dataset. Moreover, due to the extensive research design, we have only been able to use 173 news article pair ratings for each condition. This has arguably prevented us to examine confounding personal characteristics as is done in Study 1. However, instead, we have focused on comparing similarity functions to human judgments across all data or larger subsets, while predicting the impact of the news article matching (e.g., based on Date + Topic + Named Entity) in different ways.

Furthermore, the current study has assessed existing similarity functions. However, we suggest to develop and assess feature-specific similarity functions that unambiguously apply to the news domain. For example, similarity functions that leverage names (e.g., 'Donald Trump' or 'France') could help to manage user expectations about inter-article similarity. Furthermore, it would be most useful to test our assertions in an online study where news article recommendations are evaluated, much like the work of [16] and [12].

We have also collected assessments of a user's familiarity with each presented news article. What familiarity means exact in the news context is arguably less clear than for, for example, movies and recipes [16]. Some movies are even watched multiple times, just as recipes can be prepared on multiple occasions. In contrast, news articles are rarely read more than once. Moreover, it is possible that users have already heard of a specific movie or recipe, while they could only be familiar with a certain news event. For example, if the trailer of an unseen movie has been watched by a user, she could indicate to be familiar with that particular movie. In contrast, news articles have a fast turnover rate, and are rarely read a few days after publication [1]. Hence, most news articles can only be assessed as familiar due to other cues, such as a known person of interest or a familiar topic.

7.5 Future Work

We aim to further improve our content-based news recommender studies. For one, it may be worthwhile to examine the impact of additional factors. This can include tone, style, or quality of journalism, which are also mentioned in Study 2. Additionally, utilizing different and superior functions may also support the representativeness of similarity functions. For example, there are newer and improved versions of TF-IDF [61, 62], which are sensible to test next. Moreover, similarity functions based on named entities might be worthwhile to test, for instance in a Jaccard-based function.

We also propose to additional methods for similarity functions. Recent methods such as Word2Vec and BERT [30, 31] could be used for topic modelling and possibly gain improvements in performance. Another possibility would be to use algorithms based on large language models, which could be to be effective in body text-based metrics, due to the larger input.

Finally, we would like to endorse work in which computer science methods and social science methods are combined. The starting point is this paper is rather computationally-driven, but examining whether our methods actually resonate with user perceptions is an important step to take. We would like to advocate for such research, as also performed by Winecoff et al. [13], to fuse these two scientific domains further, such as recommender systems and psychology. Moreover, we feel that computer-mediated communication or digital journalism could play an important role in topic, by also considering how news article recommendations are communicated to end users and whether algorithms take heed of democratic values (cf. [60]). Since similarity functions have the potential to be explainable, they could also be employed in studies were news recommendation can have an impact beyond a single session.

Acknowledgments

We thank Gloria A.B. Kasangu for her help in check the manuscript. This work is in part funded by MediaFutures partners and the Research Council of Norway (grant number 309339).

References

[1] Karimi, M., Jannach, D., Jugovac, M.: News recommender systems–survey and roads ahead. Information Processing & Management **54**(6), 1203–1227 (2018)

[2] Tintarev, N., Masthoff, J.: Similarity for news recommender systems. In: In Proceedings of the AH'06 Workshop on Recommender Systems and Intelligent User Interfaces (2006). Citeseer

[3] Bogers, T., Van Den Bosch, A.: Comparing and evaluating information retrieval algorithms for news recommendation. In: RecSys'07: Proceedings of the 2007 ACM Conference on Recommender Systems (2007). https://doi.org/10.1145/1297231.1297256

[4] Luostarinen, T., Kohonen, O.: Using topic models in content-based news recommender systems. In: Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013), pp. 239–251 (2013)

[5] Das, A.S., Datar, M., Garg, A., Rajaram, S.: Google news personalization: scalable online collaborative filtering. In: Proceedings of the 16th International Conference on World Wide Web, pp. 271–280 (2007)

[6] Chu, W., Park, S.-T., Beaupre, T., Motgi, N., Phadke, A., Chakraborty, S., Zachariah, J.: A case study of behavior-driven conjoint analysis on Yahoo! front page today module. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1097–1104 (2009)

[7] Fortuna, B., Fortuna, C., Mladenić, D.: Real-time news recommender system. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 583–586 (2010). Springer

[8] De Pessemier, T., Courtois, C., Vanhecke, K., Van Damme, K., Martens, L., De Marez, L.: A user-centric evaluation of context-aware recommendations for a mobile news service. Multimedia Tools and Applications **75**(6), 3323–3351 (2016)

[9] Lops, P., De Gemmis, M., Semeraro, G.: Content-based recommender systems: State of the art and trends. In: Recommender Systems Handbook, pp. 73–105. Springer, New York, NY, USA (2011)

[10] Jannach, D., Zanker, M., Felfernig, A., Friedrich, G.: Recommender Systems: an Introduction. Cambridge University Press, Cambridge, UK (2010)

[11] Jannach, D., Adomavicius, G.: Recommendations with a purpose. In: Proceedings of the 10th ACM Conference on Recommender Systems, pp. 7–10 (2016)

[12] Yao, Y., Harper, F.M.: Judging similarity: a user-centric study of related item recommendations. In: Proceedings of the 12th ACM Conference on Recommender Systems, pp. 288–296 (2018)

[13] Winecoff, A.A., Brasoveanu, F., Casavant, B., Washabaugh, P., Graham, M.: Users in the loop: a psychologically-informed approach to similar item retrieval. In: Proceedings of the 13th ACM Conference on Recommender Systems, pp. 52–59 (2019)

[14] Joris, G., Grove, F.D., Van Damme, K., De Marez, L.: Appreciating news algorithms: Examining audiences' perceptions to different news selection mechanisms. Digital Journalism **9**(5), 589–618 (2021)

[15] Elbadrawy, A., Karypis, G.: User-specific feature-based similarity models for top-n recommendation of new items. ACM Transactions on Intelligent Systems and Technology (TIST) **6**(3), 1–20 (2015)

[16] Trattner, C., Jannach, D.: Learning to recommend similar items from human judgments. User Modeling and User-Adapted Interaction **30**(1), 1–49 (2020)

[17] Starke, A.D., Øverhaug, S., Trattner, C.: Predicting feature-based similarity in the news domain using human judgments. In: 15th ACM Conference on Recommender Systems, RecSys 2021 (2021)

[18] Capelle, M., Hogenboom, F., Hogenboom, A., Frasincar, F.: Semantic news recommendation using WordNet and Bing similarities. In: Proceedings of the 28th Annual ACM Symposium on Applied Computing, pp. 296–302 (2013)

[19] Özgöbek, Ö., Gulla, J.A., Erdur, R.C.: A survey on challenges and methods in news recommendation. In: International Conference on Web Information Systems and Technologies, vol. 2, pp. 278–285 (2014). SCITEPRESS

[20] Richardson, R., Smeaton, A., Murphy, J.: Using WordNet as a knowledge base for measuring semantic similarity between words (1994)

[21] Lin, D.: An information-theoretic definition of similarity. In: ICML, vol. 98, pp. 296–304 (1998)

[22] Takale, S.A., Nandgaonkar, S.S.: Measuring semantic similarity between words using web documents. International Journal of Advanced Computer Science and Applications (IJACSA) **1**(4) (2010)

[23] Lv, Y., Moon, T., Kolari, P., Zheng, Z., Wang, X., Chang, Y.: Learning to model relatedness for news recommendation. In: Proceedings of the 20th International Conference on World Wide Web, pp. 57–66 (2011)

[24] Billsus, D., Pazzani, M.J.: User modeling for adaptive news access. User Modelling and User-Adapted Interaction (2000)

[25] Cantador, I., Castells, P.: Semantic contextualisation in a news recommender system. In: Workshop on Context-Aware Recommender Systems at the RecSys 2009: ACM Conference on Recommender Systems, p. 5. ACM, New York (2009)

[26] Lommatzsch, A., Kille, B., Hopfgartner, F., Ramming, L.: NewsREEL multimedia at MediaEval 2018: News recommendation with image and text content. In: CEUR Workshop Proceedings (2018)

[27] Rorvig, M.: Images of similarity: A visual exploration of optimal similarity metrics and scaling properties of trec topic-document sets. Journal of the American Society for Information Science **50**(8), 639–651 (1999)

[28] Goossen, F., IJntema, W., Frasincar, F., Hogenboom, F., Kaymak, U.: News personalization using the CF-IDF semantic recommender. In: ACM International Conference Proceeding Series (2011). https://doi.org/10.

1145/1988688.1988701

[29] Billsus, D., Pazzani, M.J.: Personal news agent that talks, learns and explains. In: Proceedings of the International Conference on Autonomous Agents (1999)

[30] Chamberlain, B.P., Rossi, E., Shiebler, D., Sedhain, S., Bronstein, M.M.: Tuning word2vec for large scale recommendation systems. In: Fourteenth ACM Conference on Recommender Systems, pp. 732–737 (2020)

[31] Liu, J., Xia, C., Li, X., Yan, H., Liu, T.: A BERT-based ensemble model for chinese news topic prediction. In: Proceedings of the 2020 2nd International Conference on Big Data Engineering, pp. 18–23. Association for Computing Machinery, New York, NY, USA (2020)

[32] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research (2003). https://doi.org/10.1016/b978-0-12-411519-4.00006-9

[33] Li, L., Wang, D., Li, T., Knox, D., Padmanabhan, B.: Scene: a scalable two-stage personalized news recommendation system. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 125–134 (2011)

[34] Yeung, K.F., Yang, Y.: A proactive personalized mobile news recommendation system. In: Proceedings - 3rd International Conference on Developments in eSystems Engineering, DeSE 2010 (2010)

[35] Qiu, J., Liao, L., Li, P.: News recommender system based on topic detection and tracking. In: International Conference on Rough Sets and Knowledge Technology, pp. 690–697 (2009). Springer

[36] IJntema, W., Goossen, F., Frasincar, F., Hogenboom, F.: Ontology-based news recommendation. In: ACM International Conference Proceeding Series (2010). https://doi.org/10.1145/1754239.1754257

[37] Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., Sartin, M.: Combining content-based and collaborative filters in an online newspaper. In: Proceedings of the ACM SIGIR '99 Workshop on Recommender Systems: Algorithms and Evaluation (1999)

[38] Garcin, F., Faltings, B.: PEN recsys: A personalized news recommender systems framework. In: ACM International Conference Proceeding Series (2013). https://doi.org/10.1145/2516641.2516642

[39] Lu, Z., Dou, Z., Lian, J., Xie, X., Yang, Q.: Content-based collaborative filtering for news topic recommendation. Proceedings of the National

Conference on Artificial Intelligence **1**, 217–223 (2015)

[40] Desarkar, M.S., Shinde, N.: Diversification in news recommendation for privacy concerned users. In: DSAA 2014 - Proceedings of the 2014 IEEE International Conference on Data Science and Advanced Analytics (2014). https://doi.org/10.1109/DSAA.2014.7058064

[41] Lenhart, P., Herzog, D.: Combining content-based and collaborative filtering for personalized sports news recommendations. In: CEUR Workshop Proceedings (2016)

[42] Pon, R.K., Cardenas, A.F., Buttler, D., Critchlow, T.: Tracking multiple topics for finding interesting articles. In: Proceedings of the ACM SIGKDD International Conference (2007). https://doi.org/10.1145/1281192.1281253

[43] Soroka, S., Young, L., Balmas, M.: Bad News or Mad News? Sentiment Scoring of Negativity, Fear, and Anger in News Content. Annals of the American Academy of Political and Social Science (2015). https://doi.org/10.1177/0002716215569217

[44] Carbone, P., Vlassov, V.: Auto-scoring of personalised news in the real-time web: Challenges, overview and evaluation of the state-of-the-art solutions. In: 2015 International Conference on Cloud and Autonomic Computing, pp. 169–180 (2015). IEEE

[45] Garcin, F., Faltings, B., Donatsch, O., Alazzawi, A., Bruttin, C., Huber, A.: Offline and online evaluation of news recommender systems at swissinfo.ch. In: Proceedings of the 8th ACM Conference on Recommender Systems, pp. 169–176 (2014)

[46] Watters, C., Wang, H.: Rating news documents for similarity. Journal of the American Society for Information Science **51**(9), 793–804 (2000)

[47] Tversky, A.: Features of similarity. Psychological review **84**(4), 327 (1977)

[48] Gauch, S., Speretta, M., Chandramouli, A., Micarelli, A.: User profiles for personalized information access. The Adaptive Web, 54–89 (2007)

[49] Li, L., Wang, D.-D., Zhu, S.-Z., Li, T.: Personalized news recommendation: a review and an experimental investigation. Journal of Computer Science and Technology **26**(5), 754–766 (2011)

[50] Pon, R.K., Cardenas, A.F., Buttler, D., Critchlow, T.: Tracking multiple topics for finding interesting articles. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 560–569 (2007)

[51] Bauer, C., Bagchi, C., Hundogan, O.A., van Es, K.: Where are the values? a systematic literature review on news recommender systems. ACM Transactions on Recommender Systems (2024)

[52] Kruse, J., Michiels, L., Starke, A., Tintarev, N., Vrijenhoek, S.: Normalize: A tutorial on the normative design and evaluation of information access systems. In: Proceedings of the 2024 Conference on Human Information Interaction and Retrieval, pp. 422–424 (2024)

[53] Yujian, L., Bo, L.: A normalized Levenshtein distance metric. IEEE Transactions on Pattern Analysis and Machine Intelligence (2007). https://doi.org/10.1109/TPAMI.2007.1078

[54] Jaro, M.A.: Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. Journal of the American Statistical Association (1989). https://doi.org/10.1080/01621459.1989.10478785

[55] Kondrak, G.: N-gram similarity and distance. In: Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2005). https://doi.org/10.1007/11575832

[56] NIST: TREC Washington Post Corpus (2019). https://trec.nist.gov/data/wapost/

[57] Flaounas, I., Ali, O., Lansdall-Welfare, T., De Bie, T., Mosdell, N., Lewis, J., Cristianini, N.: Research methods in the age of digital journalism: Massive-scale automated analysis of news-content—topics, style and gender. Digital journalism **1**(1), 102–116 (2013)

[58] Gaillard, S., Oláh, Z.A., Venmans, S., Burke, M.: Countering the cognitive, linguistic, and psychological underpinnings behind susceptibility to fake news: A review of current literature with special focus on the role of age and digital literacy. Frontiers in Communication **6**, 661801 (2021)

[59] Peer, E., Rothschild, D., Gordon, A., Evernden, Z., Damer, E.: Data quality of platforms and panels for online behavioral research. Behavior Research Methods **54**(4), 1643–1662 (2022)

[60] Vrijenhoek, S., Bénédict, G., Gutierrez Granada, M., Odijk, D., De Rijke, M.: Radio–rank-aware divergence metrics to measure normative diversity in news recommendations. In: Proceedings of the 16th ACM Conference on Recommender Systems, pp. 208–219 (2022)

[61] Capelle, M., Frasincar, F., Moerland, M., Hogenboom, F.: Semantics-based news recommendation. In: Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics, pp. 1–9 (2012)

[62] Goossen, F., IJntema, W., Frasincar, F., Hogenboom, F., Kaymak, U.: News personalization using the CF-IDF semantic recommender. In: Proceedings of the International Conference on Web Intelligence, Mining and Semantics, pp. 1–12 (2011)