Predicting Feature-based Similarity in the News Domain Using Human Judgments

ALAIN D. STARKE, Wageningen University & Research, The Netherlands and University of Bergen, Norway
SEBASTIAN ØVERHAUG LARSEN and CHRISTOPH TRATTNER, University of Bergen, Norway

When reading an online news article, users are typically presented 'more like this' recommendations by news websites. In this study, we assessed different similarity functions for news item retrieval, by comparing them to human judgments of similarity. We asked 401 participants to assess the overall similarity of ten pairs of political news articles, which were compared to feature-specific similarity functions (e.g., based on body text or images). We found that users indicated to mostly use text-based features (e.g., title) for their similarity judgments, suggesting that body text similarity was the most representative for their judgment. Moreover, we modeled similarity judgments using different regression techniques. Using data from another study, we contrasted our results across retrieval domains, revealing that similarity functions in news are less representative of user judgments than those in movies and recipes.

Additional Key Words and Phrases: news, similarity, similar-item retrieval, recommender systems, user study, human judgment

ACM Reference Format:
Alain D. Starke, Sebastian Øverhaug Larsen, and Christoph Trattner. 2021. Predicting Feature-based Similarity in the News Domain Using Human Judgments. In *INRA 2021: full name, September, 2021, Amsterdam, NL.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Similarity functions are central to recommender systems and information retrieval systems [13]. They assess the similarity between a reference article and a set of possible recommendations [27]. Using a dataset with political news articles, 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.

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 or may be updated later [6]. Consequently, user preferences may strongly depend on contextual factors, such as a user's time of day or location [7, 9].

News websites typically present content-based recommendations [13]. 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.

Whether two news articles are alike can be computed using similarity functions [7, 13]. Features (e.g., title) considered by such functions should to a large extent reflect a user's similarity assessment [8], while not being too similar to what a user is currently reading, for it may lead to redundancy [27]. However, research on feature-based similarity is limited

Manuscript submitted to ACM

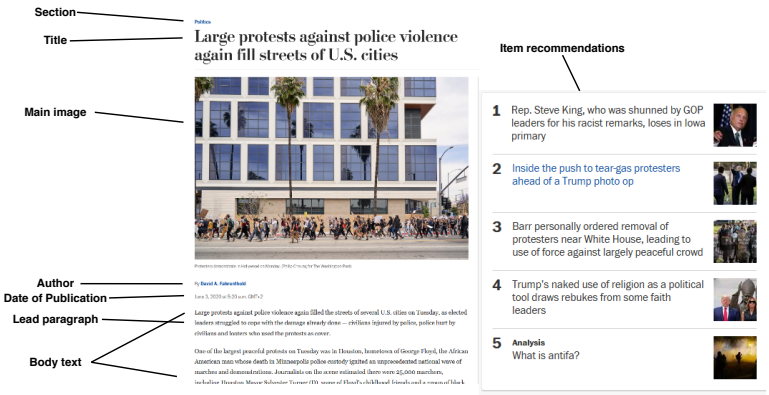

Fig. 1. Different features in a news article, which may be used by a news recommender system to recommend items to a user.

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 [28]. 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.

Hence, the current study assesses a set of similarity functions for news article retrieval, particularly for the task of similar-item recommendation. We ask users of an online news system to judge the similarity between pairs of news articles, which is used to develop a model to predict news similarity. Subsequently, we perform cross-domain comparisons, comparing which features are used for human similarity judgments in news, movies, and recipes, using data from [28]. We posit the following research questions:

- **RQ1:** Which news article features are used by humans to judge similarity and to what extent are different feature-specific similarity functions related to human similarity judgments?
- **RQ2:** Which combination of news article features is best suited to predict user similarity judgments?
- **RQ3:** How does the use of news features and their similarity functions compare to those used in the recipe and movie domains?

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 juxtapose our news study with findings from the movie and recipe domains, using data from [28], showing that feature-specific similarity functions in the news domains are less representative of human judgment than functions in the movie and recipe domains.
- We present a reproducible data processing pipeline, available on Github[1], and add a benchmarking dataset for the publicly available Washington Post Corpus news article database.

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

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.

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 [13]. In the recommender domain, this is referred to as a similar-item recommendation problem. A fundamental question is how to compute similarity between concepts [21, 30], which is examined in studies on semantic similarity [23], a field of research that usually not only captures the similarity between two concepts, but also how different they are [16]. 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 [26, 27]. For example, latent semantic analysis derives meaning and similarity from the text context itself, by examining *how* and *how often* words are used [27].

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 [19], *Term Frequency-Inverse Document Frequency* remains one of the most commonly used IR methods to create similarity vectors [2]. It uses the term frequency per document and the inverse appearance frequency across all documents [11], while similarity between the vectors of liked and unseen items can be computed using cosine similarity [4].

A much simpler approach is to derive a set of keywords from each item [11]. 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).

2.2 Similarity Representations in the News Domain

News recommender systems primarily focus on textual representations of news articles [13]. Most approaches utilize the main text or title, ignoring most other textual features, such as the author [2]. A straightforward, but more uncommon approach in academic studies [18], 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 [24], such as food [28].

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 [10]. 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 [1].

Besides the aforementioned methods, a common approach is to derive latent topics from texts. Although recent work uses Word2Vec and BERT [5, 17], this work considers Latent Dirichlet Allocation (LDA) and Probabilistic Latent Semantic Indexing (PLSI) [15]. LDA and PLSI can cluster topically-similar news articles based on tags and named entities. News recommendations can be refined afterwards based on recency scores.

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 [25].

2.2.2 Other News Features. A news article's date-time feature is also leveraged in the context of similar-item news recommendation, either through pre-filtering, recency modeling, or post-filtering [13]. 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. [22] 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, as also used in similar-item recommender systems, are typically validated using human judgments [26]. An important question is to what extent 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 [30]. This has been studied in the movie and recipe domains: [28] 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, [31] 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 [27, 29], none have done so across multiple features. For example, users in [27] successfully judged the similarity between news articles, but only based on their headlines.

2.4 Key differences with previous work

Novel to our approach is the use of feature-specific similarity representations and functions in news, as well as grounding them in human similarity judgments. Most relevant to our approach is the work of [28] and [31], for they explore how computational functions for similarity compare to users' perception of similarity. In particular, [28] serves as an example for our approach, 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 goal of the current study is to assess their performance in news.

3 METHOD

We assess the utility of different feature-specific similarity functions by collecting human judgments of similarity for pairs of news articles. 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.

3.1 Dataset and Feature Engineering

3.1.1 News Database. 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 [20]. 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.

Table 1. 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

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 2400 'Politics' news articles, 400 from each year between 2012 and 2017, for the descriptive statistics are reported in Table 1.

3.2 Modeling Similarity with Feature-Based Similarity Functions

To model the similarity between two news articles, 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 – based on findings from the movie and recipe domains [28].

Table 2 describes the developed similarity functions. For each pair of news articles, we computed similarity scores based on 7 main features: 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.

3.2.1 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) [32], the Jaro-Winkler method (JW) [12], the longest common subsequence, and the bi-gram distance method (BI) [14]. As in [28], Latent Dirichlet Allocation (LDA) topic-modeling was set to 100 topics.

3.2.2 Image Features. In line with the current state-of-the-art [28], 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.

3.2.3 Body Text. Body 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 [25]). 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 [28], and similarity based on the entire body text.

Table 2. Similarity functions employed in the current study, each comprised of a feature 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

3.3 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.

3.3.1 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 [28]. We computed the pairwise similarity across all 2400 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 2000 similar pairs from decile Q10.

3.3.2 Procedure and Measures. The resulting 6000 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.[2] Much like [27], users were asked to 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.

3.3.3 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

[2]Users were asked for this pair 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.

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 [28]. 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.

4 RESULTS

For our analyses, we first examined the use of different news features, assessing different similarity functions through human judgments (RQ1). Furthermore, we predicted human similarity judgments using model-based approaches (RQ2). In addition, we compared our results for RQ1-RQ2 with the news and recipe domains (RQ3).

4.1 News Features Usage

We examined to what extent participants used different features to assess similarity between news articles (RQ1). 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.

With regard to [RQ3], most findings were compatible with the movie and recipes domains. 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 movies [28]. The use of the genre cue in movies was also more frequent than the use of a news article's subcategory.

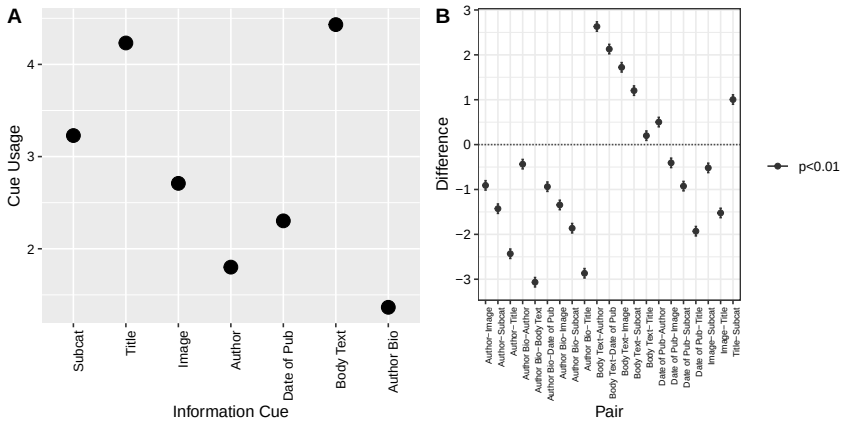

Fig. 3. **A**: Mean reported cue usage for news articles, scaled 1-5; **B**: Tukey's HSD post hoc tests (means and S.E.) that examine differences in cue usage.

4.2 Grounding Similarity Functions in Human Similarity Judgments

4.2.1 Descriptive Statistics. To address [RQ1], 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.

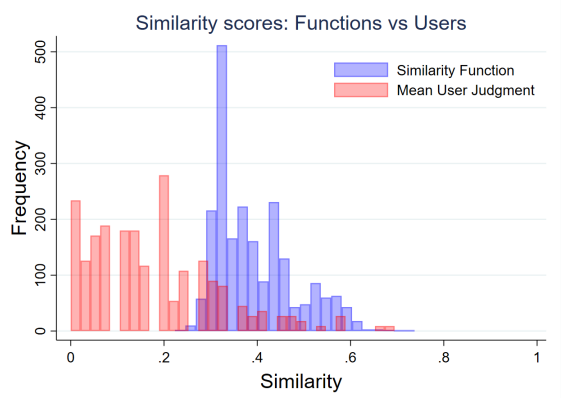

Fig. 4. Frequency of similarity scores (scaled 0-1). Similarity functions depict the average score per news article pair, user judgments show the mean given similarity judgment per user.

4.2.2 Feature-specific Comparison in News. Table 3 outlines the Spearman correlations between similarity functions and the similarity judgments given by users. It differentiates between the results of our own user study (i.e., 'News Articles'), and that of [28] for the movie and recipe domains, allowing for cross-domain comparisons (discussed later).

Table 3. Spearman correlations between similarity functions and human similarity judgments, for news (current study), and recipes and movies (obtained from [28]). 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			**Recipes**			**Movies**	
Similarity Function	ρ_{pass}	ρ_{all}	Sim. Function	ρ_{pass}	ρ_{all}	Sim. Function	ρ_{pass}	ρ_{all}
Subcat:Jacc	0.14^{***}	0.11				Genre:Jacc	0.56^{***}	0.53^{***}
Title:LV	0.06^{**}	0.04^{*}	Title:LV	0.48^{***}	0.38^{***}	Title:LV	0.19^{***}	0.18^{***}
Title:JW	0.05^{*}	0.03	Title:JW	0.46^{***}	0.35^{***}	Title:JW	0.16^{***}	0.16^{***}
Title:LCS	0.07^{***}	0.05^{**}	Title:LCS	0.50^{***}	0.40^{***}	Title:LCS	0.20^{***}	0.19^{***}
Title:BI	0.08^{***}	0.07^{***}	Title:BI	0.48^{***}	0.38^{***}	Title:BI	0.17^{***}	0.17^{***}
Title:LDA	0.02	0.00	Title:LDA	0.22^{***}	0.19^{***}	Title:LDA	0.01	0.01
Image:BR	0.10^{***}	0.07^{***}	Image:BR	0.18^{**}	0.14^{*}	Image:BR	0.22^{***}	0.20^{***}
Image:SH	0.06^{**}	0.03	Image:SH	0.16^{*}	0.11^{*}	Image:SH	0.10^{***}	0.08^{***}
Image:CO	0.05^{*}	0.05^{**}	Image:CO	0.29^{***}	0.20^{***}	Image:CO	0.03	0.03
Image:COL	0.05^{*}	0.03^{*}	Image:COL	0.09^{*}	0.07^{*}	Image:COL	0.15^{***}	0.14^{***}
Image:EN	0.07^{**}	0.05^{**}	Image:EN	0.34^{***}	0.28^{***}	Image:EN	0.15^{***}	0.09^{***}
Image:EMB	0.17^{***}	0.13^{***}	Image:EMB	0.44^{***}	0.34^{***}	Image:EMB	0.18^{***}	0.16^{***}
Author:Jacc	0.13^{***}	0.10^{***}				Dir:Jacc	0.10^{***}	0.07^{***}
Date:ND	0.09^{***}	0.08^{***}				Date:MD	0.37^{***}	0.35^{***}
BodyText:TFIDF	0.29^{***}	0.23^{***}						
BodyText:50TFIDF	0.14^{***}	0.12^{***}	Dir:TFIDF	0.50^{***}	0.40^{***}	Plot:TFIDF	0.25^{***}	0.20^{***}
BodyText:LDA	0.03	0.01	Dir:LDA	0.54^{***}	0.43^{***}	Plot:LDA	0.37^{***}	0.34^{***}
BodyText:Sent	-0.02	-0.02						
AuthorBio:TFIDF	0.15^{***}	0.12^{***}						
AuthorBio:LDA	0.11^{***}	0.09^{***}						

We first discuss the results for the news domain and focus on users who passed the attention check. Table 3 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 [13]. 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*, *AuthorBio:TFIDF*, and *Subcat:Jacc*, seemed to best represent user similarity judgments in news.

Table 3 highlights that other functions did not represent a user's similarity judgment in news, 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 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 3. 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).

4.2.3 Cross-domain Comparison. Using data from [28], we compared the results in Table 3 across the news, recipe, and movie domains. Correlations between human judgments and similarity functions in the news domain were shown to be much weaker than in the recipe domain and, to a lesser extent, the movie domain. This applied to most features, including title, image, and body text.

Two notable differences lie in title and image-based functions. Whereas the reported correlations for title features were weak in news ($\rho < 0.1$), the distance-based title metrics showed strong correlations with user judgments for recipes ($rho \approx 0.5$). With regard to image-specific similarity, functions in news were only weakly correlated to human judgments ($\rho_{max} = 0.17$), while they were more representative for recipes ($\rho_{max} = 0.44$) and movies ($\rho_{max} = 0.22$).

4.3 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 (RQ2), as used in recommender systems research. This helped us to understand each feature's importance, beyond the feature-specific correlations in Table 3.

4.3.1 Model Evaluation and Cross-Domain Comparison. To determine model performance, standard metrics such as Root Mean Square Error (RMSE), R^2, and Mean Absolute Error (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 4. In part (i), a Wilcoxon Rank-Sum test on RMSE pointed out that all models except GB performed significantly better than a random baseline ($p_{all} < 0.05$). Table 4 (i) also compares our results to findings from the recipe and movie domains (RQ3), adapted from [28]. Most notably, we found that Lasso is the best performing model, while Ridge was the winner in the Recipe and Movie domains. Moreover, the news model (i.e., $R^2 = 0.33$) was less accurate than the recipe model (i.e., $R^2 = 0.51$), while its accuracy was comparable to that of the movie model (i.e., $R^2 = 0.36$). This suggested that the similarity functions adapted from [28] were less representative for user similarity judgments in the news domain.

Table 4. 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, recipe, and movie domains (using data from [28]), (ii) describe the accuracy of feature-specific models in news, and include (iii) user characteristics. The best performing models per domain are denoted in bold.

	News Articles ($N = 2,169$)			**Recipes** ($N = 1,539$)			**Movies** ($N = 1,395$)		
Method	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE
(i) Model performance (All features)									
All (Random Forest (RF))	0.9219	0.2982	0.7643	0.8958	0.4734	0.6787	0.8807	0.3543	0.7007
All (Gradient Boosting (GB))	0.9177	0.3123	0.7520	0.8805	0.4921	0.6672	0.8844	0.3489	0.7029
All (Ridge Regression)	0.9141	0.3257	0.7459	**0.8654**	**0.5063**	**0.6651**	**0.8745**	**0.3628**	**0.6926**
All (Linear Regression)	0.9120	0.3289	0.7453	0.8700	0.5022	0.6668	0.8752	0.3616	0.6929
All (Lasso Regression)	**0.9101**	**0.3339**	**0.7480**	0.8873	0.3574	0.7286	0.8873	0.3574	0.7286
Mean	0.9652	0.0000	0.8122	1.2292	0.4995	1.0433	1.0942	0.5001	0.9140
Random	0.9659	-0.0226	0.8125	1.2290	0.0010	1.0435	1.0948	0.0061	0.9140
(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						

4.3.2 Feature-specific Models and User Characteristics. To further explore [RQ2], Table 4 (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 3), 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 4 (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.

5 DISCUSSION

This work contributes to the literature on similarity estimates, with a particular focus on the news domain, which is a central theme in the recommender systems literature. It is among the first to study news similarity representations in detail, making the following contributions:

(1) Determining which features are considered by users when judging similarity between news articles.
(2) Assessing how feature-specific similarity functions relate to similarity judgments.
(3) Predicting similarity judgments of users through machine learning models.
(4) Comparing our results to findings from the movie and recipe domains.

We have set a first step towards designing representative feature-specific similarity functions for news, going beyond other studies that focused on overall similarity or just a single feature [27, 29].

5.1 Feature-specific Similarity

We have assessed the value of feature-specific similarity functions in the news domain, adapted from recommender literature in the news, movie, and recipe domains [28]. We find that most feature-specific similarity functions only partially reflect 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 is grounded in the reported feature use, as well as consistent with previous studies on news retrieval [13]. 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 (cf. Table 1), compared to the other domains in scope. At the 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 have used machine learning to determine model accuracy and feature importance, and to examine the predictive value of additional user characteristics. We find 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.

5.2 Cross-domain Comparisons

We have also explored cross-domain differences. In line with [28], we have found further evidence that different domains call for different similarity functions. For one, the ridge regression model for news is found to be somewhat less accurate than for news and recipes, although a R^2 of 0.33 is reasonable. However, the MAE of 0.75 for a measure that is scaled from 1 to 5 suggests that there is room for improvement, which could be attributed to the low given similarity scores.

It seems that text-based similarity (i.e., movie plot, recipe directions, news' body text) is useful in most domains in scope, given an appropriate similarity function. BodyText features are listed among the strongest correlations, as well as among the strongest predictors. In contrast, the title and image features are less representative of similarity judgments in news and movies, compared to the recipe domain. Whereas only image embeddings seem to be somewhat representative of news similarity assessments, images features are more useful in determining recipe similarity.

We have observed that the model accuracy reported in Table 4 is comparable to findings from the movie domain (cf. [28]). This is despite the differences in given similarity scores across domains (which is much lower for news; see Figure 4), and the weaker correlations reported in Table 3. All in all, the news domain seems to require similarity functions that are less 'taste-related' than movies or recipes, but further research is needed to develop more accurate ones, possibly by also using psychological theories on similarity [30].

5.3 Limitations & Future Work

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'. Furthermore, the news articles shown to users were a few years old, which might have reduced familiarity levels and, in turn, decreased similarity ratings.

Another shortcoming is that it is not entirely clear on what grounds users have made their similarity judgments. We have asked them a single question on similarity, while some other studies have also used multiple questionnaire items [27]. 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.

For future studies, we suggest to develop and assess feature-specific similarity functions that unambiguously apply to the news domain. For example, similarity functions that leverage named entities (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 [28] and [31].

Above all, we like to emphasize that the current study serves as a first step. Based on these findings, future studies can further develop feature-specific similarity functions for the news domains, for this paper provides insight in what types of functions and features are successful, and which ones are not.

ACKNOWLEDGEMENTS

This work was supported by industry partners and the Research Council of Norway with funding to MediaFutures: Research Centre for Responsible Media Technology and Innovation, through the Centres for Research-based Innovation scheme, project number 309339.

REFERENCES

[1] Daniel Billsus and Michael J. Pazzani. 1999. Personal news agent that talks, learns and explains. In *Proceedings of the International Conference on Autonomous Agents*.

[2] Daniel Billsus and Michael J. Pazzani. 2000. User modeling for adaptive news access. *User Modelling and User-Adapted Interaction* (2000).

[3] Toine Bogers and Antal Van Den Bosch. 2007. Comparing and evaluating information retrieval algorithms for news recommendation. In *RecSys'07: Proceedings of the 2007 ACM Conference on Recommender Systems*. https://doi.org/10.1145/1297231.1297256

[4] Iván Cantador and Pablo Castells. 2009. Semantic contextualisation in a news recommender system. In *Workshop on Context-Aware Recommender Systems at the RecSys 2009: ACM Conference on Recommender Systems*. ACM, New York.

[5] Benjamin P Chamberlain, Emanuele Rossi, Dan Shiebler, Suvash Sedhain, and Michael M Bronstein. 2020. Tuning Word2vec for Large Scale Recommendation Systems. In *Fourteenth ACM Conference on Recommender Systems*. 732–737.

[6] Abhinandan S Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. 2007. Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th international conference on World Wide Web*. 271–280.

[7] Toon De Pessemier, Cédric Courtois, Kris Vanhecke, Kristin Van Damme, Luc Martens, and Lieven De Marez. 2016. A user-centric evaluation of context-aware recommendations for a mobile news service. *Multimedia Tools and Applications* 75, 6 (2016), 3323–3351.

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

[9] Blaž Fortuna, Carolina Fortuna, and Dunja Mladenić. 2010. Real-time news recommender system. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 583–586.

[10] Frank Goossen, Wouter Ijntema, Flavius Frasincar, Frederik Hogenboom, and Uzay Kaymak. 2011. News personalization using the CF-IDF semantic recommender. In *ACM International Conference Proceeding Series*. https://doi.org/10.1145/1988688.1988701

[11] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. 2010. *Recommender systems: an introduction*. Cambridge University Press.

[12] Matthew A. Jaro. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *J. Amer. Statist. Assoc.* (1989). https://doi.org/10.1080/01621459.1989.10478785

[13] Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. 2018. News recommender systems–Survey and roads ahead. *Information Processing & Management* 54, 6 (2018), 1203–1227.

[14] Grzegorz Kondrak. 2005. N-gram similarity and distance. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/11575832{_}13

[15] Lei Li, Dingding Wang, Tao Li, Daniel Knox, and Balaji Padmanabhan. 2011. SCENE: A scalable two-stage personalized news recommendation system. In *SIGIR'11 - Proceedings of the 34th International ACM SIGIR Conference*.

[16] Dekang Lin. 1998. An information-theoretic definition of similarity.. In *ICML*, Vol. 98. 296–304.

[17] Jingang Liu, Chunhe Xia, Xiaojian Li, Haihua Yan, and Tengteng Liu. 2020. A BERT-based Ensemble Model for Chinese News Topic Prediction. In *Proceedings of the 2020 2nd International Conference on Big Data Engineering*. 18–23.

[18] Andreas Lommatzsch, Benjamin Kille, Frank Hopfgartner, and Leif Ramming. 2018. NewsREEL multimedia at MediaEval 2018: News recommendation with image and text content. In *CEUR Workshop Proceedings*.

[19] Yuanhua Lv, Taesup Moon, Pranam Kolari, Zhaohui Zheng, Xuanhui Wang, and Yi Chang. 2011. Learning to model relatedness for news recommendation. In *Proceedings of the 20th International Conference on World Wide Web*. 57–66.

[20] NIST. 2019. TREC Washington Post Corpus. data retrieved from, https://trec.nist.gov/data/wapost/.

[21] Özlem Özgöbek, Jon Atle Gulla, and R Cenk Erdur. 2014. A Survey on Challenges and Methods in News Recommendation. In *International Conference on Web Information Systems and Technologies*, Vol. 2. SCITEPRESS, 278–285.

[22] Raymond K. Pon, Alfonso F. Cardenas, David Buttler, and Terence Critchlow. 2007. Tracking multiple topics for finding interesting articles. In *Proceedings of the ACM SIGKDD International Conference*. https://doi.org/10.1145/1281192.1281253

[23] Ray Richardson, A Smeaton, and John Murphy. 1994. *Using WordNet as a knowledge base for measuring semantic similarity between words*. Technical Report Working Paper CA-1294.

[24] Mark Rorvig. 1999. 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 (1999), 639–651.

[25] Stuart Soroka, Lori Young, and Meital Balmas. 2015. 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

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

[27] Nava Tintarev and Judith Masthoff. 2006. Similarity for news recommender systems. In *In Proceedings of the AH'06 Workshop on Recommender Systems and Intelligent User Interfaces*. Citeseer.

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

[29] Carolyn Watters and Hong Wang. 2000. Rating news documents for similarity. *Journal of the American Society for Information Science* 51, 9 (2000), 793–804.

[30] Amy A Winecoff, Florin Brasoveanu, Bryce Casavant, Pearce Washabaugh, and Matthew Graham. 2019. Users in the loop: a psychologically-informed approach to similar item retrieval. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 52–59.

[31] Yuan Yao and F Maxwell Harper. 2018. Judging similarity: a user-centric study of related item recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 288–296.

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