Shaping the Future of Content-based News Recommenders: Insights from Evaluating Feature-Specific Similarity Metrics

DANIEL ROSNES, SFI MediaFutures, University of Bergen, Norway

ALAIN D. STARKE, Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Netherlands and MediaFutures, University of Bergen, Norway

CHRISTOPH TRATTNER, SFI MediaFutures, University of Bergen, Norway

In news media, recommender system technology faces several domain-specific challenges. The continuous stream of new content and users deems content-based recommendation strategies, based on similar-item retrieval, to remain popular. However, a persistent challenge is to select relevant features and corresponding similarity functions, and whether this depends on the specific context. We evaluated feature-specific similarity metrics using human similarity judgments across national and local news domains. We performed an online experiment ($N = 141$) where we asked participants to judge the similarity between pairs of randomly sampled news articles. We had three contributions: (1) comparing novel metrics based on large language models to ones traditionally used in news recommendations, (2) exploring differences in similarity judgments across national and local news domains, and (3) examining which content-based strategies were perceived as appropriate in the news domain. Our results showed that one of the novel large language model based metrics (SBERT) was highly correlated with human judgments, while there were only small, most non-significant differences across national and local news domains. Finally, we found that while it may be possible to automatically recommend similar news using feature-specific metrics, their representativeness and appropriateness varied. We explain how our findings can guide the design of future content-based and hybrid recommender strategies in the news domain.

CCS Concepts: • **General and reference → Metrics**; **Evaluation**; • **Information systems → Content ranking**; *Collaborative filtering*; *Personalization*; *Language models*; **Similarity measures**; **Novelty in information retrieval**; **Recommender systems**; **Relevance assessment**; • **Computing methodologies** → *Natural language processing*.

Additional Key Words and Phrases: News Recommender, Content-based Recommendation, Similarity Metrics, Human Similarity Judgements, Recommender Appropriateness

ACM Reference Format:

Daniel Rosnes, Alain D. Starke, and Christoph Trattner. 2024. Shaping the Future of Content-based News Recommenders: Insights from Evaluating Feature-Specific Similarity Metrics. In *Proceedings of the 32th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '24), July 1–4, 2024, Cagliari, Sardinia, Italy*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

1.1 Motivation

The abundance of information in today's digital landscape, particularly in news dissemination, underscores the need for tools that can effectively sift through vast content repositories and guide users toward relevant and engaging materials.

To this end, recommender systems have emerged as crucial instruments, helping to streamline information discovery, optimize content delivery, and enhance the overall user experience [11].

The news domain faces several domain-specific challenges that make the introductions of common recommender system strategies difficult [7, 12]. Similar-item recommenders are able to circumvent many of these challenges [12]. While such recommenders are popular with news websites, there is limited knowledge surrounding whether the recommendations they represent what users consider similarity between items [28]. While there are studies exploring this [28, 29], the studies are generally done with limited data, such as using single outlets, a limited number of categories within outlets, and/or a limited amount of news articles.

In this study, we attempt to explore these issues by investigating how feature-specific similarity metrics represent human similarity judgments in four different Norwegian news outlets that span the local and national domains. The primary objective is the analysis of human similarity judgment representations by feature-specific similarity metrics across local and national levels of Norwegian news outlets. Additionally, the study aims to assess the efficacy of a set of feature-specific similarity metrics, derived from recent advancements in language technologies, in comparison to traditional measures of similarity for news articles. Finally we also evaluate how well similarity by itself represent the users' desired recommendation.

- **RQ1:** To what extent do feature-specific similarity metrics represent human similarity judgments in the Norwegian news domain?
- **RQ2:** To what extent does the correlational strength between human similarity judgments and feature-specific similarity functions differ across local and national news media outlets?
- **RQ3:** To what extent are human similarity judgements reflected in perceived recommendation appropriateness?

1.2 Contributions

The goal of this study is to explore and evaluate feature-specific similarity metrics and whether they represent human similarity judgments in the Norwegian news domain. By doing this we do the following contributions:

- An extension of metrics used in [27, 28, 31], examining to what extent current state-of-the-art NLP methods represent human similarity judgments.
- A novel comparison of metric performance across pairs of national and local news outlets.
- The inclusion of a user evaluation study, examining the appropriateness of different strategies.

2 BACKGROUND

2.1 News Recommender Systems

Many news recommender system use 'more like this' recommendations. Such *Similar Item Retrieval* aims to provide an *unseen* or *novel* item that is similar to a specific reference item [28]. A key question is how to compute the similarity between the base item and candidate items to be retrieved. [20, 33].

Similar Item Retrieval is typically performed through content-based recommendation (CB) methods [12]. While collaborative filtering (CF) and knowledge-based recommenders are common in other domains [11, 22], they are typically not used in the news domain. One of the main reasons is the *permanent cold-start problem* [12], which arises from the lack of historic information from users. In news, this is due to the large number of one-time and first-time users that do not log in. Further compounding the problem is the high frequency of novel items, along with the high volatility of a news article's relevance and contextual factors, such as the time of day and the user's location [12]. It seems that such issues are avoided by using CB algorithms: In their survey, Karimi et al. [12] show that 104 out of 112 reviewed articles on news recommenders use CB algorithms or hybrid algorithms with a CB component.

Similarity-based approaches can leverage *feature-specific similarity metrics*. Among NRS features, these usually involve evaluating the article's text or title, while other features are ignored [12]. The assumption here is that these features are paid most attention to and should therefore determine similarity scores, which is, however, typically not validated [30, 35]. A traditional method to compute the similarity between text items is by deriving vectors from the text [28]. *Term Frequency-Inverse Document Frequency* (TF-IDF) remains one of the most commonly used IR methods to create similarity vectors from text [2][28].

While TF-IDF is still popular, it has been outperformed by other metrics, such as BM25 [19][28]. In recent years approaches using transformer models and Word2Vec also show better performance than TF-IDF on text similarity tasks [4, 17]. Since the introduction of transformer models with the Bidirectional Encoder Representations from Transformers (BERT) model in [32], the use of such models has received immense popularity. In recommender systems there are several approaches utilizing the embeddings provided by various transformer models [10, 13, 36], and combining transformer models with topic modeling techniques [18, 34, 37]. These have, however, not been used in recent studies on similar-item retrieval and feature-specific similarity [28].

Recommender systems are typically evaluated through offline experimentation and simulation based on historical data, through laboratory studies, or through A/B (field) tests on real-world websites [12]. In their survey Karimi et al. [12] found that a large majority of studies relied on traditional IR measures like precision and recall, rank-based measures like *Mean Reciprocal Rank* or *Normalized Discounted Cumulative Gain*, or prediction measures like the *Root Mean Square Error*. These methods all rely on a dataset annotated based on the task the recommender is meant to solve. However, such datasets are not readily available in the news domain [12].

While only 19 of the 112 papers surveyed by Karimi et al. [12] utilize it, *click-through-rate* (CTR) is a popular way to evaluate the performance of news recommenders [8]. However, CTR is not helpful in determining if the items are similar, as the user may click on the item for other reasons than similarity [23].

2.2 Related Work

In order to validate the performance of similar-item recommenders, *human judgments* are typically used [3]. A critical question is to what degree similarity functions mirror a user's judgment of the similarity between pairs of items. Problems could arise if a user undervalues or overemphasizes specific item features compared to which is calculated, and how the similarity is being calculated [28, 33].

Yao and Harper [35] collected human similarity judgments using movie pairs collected from the MovieLens[1] dataset. As part of their study, users are asked to what extent the movies are similar, and whether they would recommend the second movie to someone who likes the first. Their goal was to explore whether CF or CB algorithms provide similar item recommendations that are closer to human similarity judgments. Yao and Harper [35] suggest that CB algorithms perform better in matching human similarity judgments. Another key observation in Yao and Harper [35] is that similarity is not everything in a similar item recommender: Over 60% of the users in their survey choose a compromise over being recommended the most similar item.

Other studies where human judgments have been collected in order to evaluate similar item recommenders include Trattner and Jannach [31], Starke et al. [28], and Solberg [27]. This study builds directly on the work done in these

[1] https://grouplens.org/datasets/movielens/

studies. The main methodology of calculating feature-specific similarity metrics and comparing them with human similarity judgments used in this study is introduced by Trattner and Jannach [31]. Starke et al. [28] then applies the same methodology to the news domain. Similar to Yao and Harper [35], Solberg [27] attempts to discover *news recommender criteria*, before he uses a similar methodology to that of Trattner and Jannach [31] and Starke et al. [28] to examine differences between categories in the news domain.

In the initial work by Trattner and Jannach [31] two main studies are performed across the movie and recipe domains. The studies follow a novel approach where the goal is not to evaluate existing algorithms, but to develop new similarity functions from human similarity judgments. The human similarity judgments are used as baselines for how similar the items are, and what makes the two items similar. Trattner and Jannach [31] also asks the users which *similarity cues* the users used while evaluating the similarity. These similarity cues represent the features the feature-specific metrics are based on.

In Starke et al. [28], a similar approach to Trattner and Jannach [31] is employed, but this time in the news domain. They use a total of 2400 articles are included, with 400 articles from the 'Politics' category are randomly sampled from each year between 2012 and 2017 TREC Washington Post dataset[2]. Following the method put forward by Trattner and Jannach [31], a survey was conducted to collect human similarity judgments. The obtained similarity judgments exhibited low correlations with the metrics across all aspects, with an average Spearman correlation coefficient of 0.092. Among the metrics, the highest correlating one was TF-IDF when applied to body-text, demonstrating a correlation coefficient of 0.29. Several prediction models were then trained based on the data from the survey to create a specific news recommender algorithm.

In his thesis, Solberg [27] builds upon this by addressing two primary problems. The first problem focuses on defining the criteria for news recommendation, while the second problem aims to explore the differences between specific news categories, namely *Sports* and *Recent Events*. His thesis is divided into two separate studies, each addressing one of these questions. Similar to Yao and Harper [35], he shows that only 26 of the 45 participants in the study selected *item similarity* as a factor. While this was the most common response, it does show that similarity may not be the primary goal of a news recommender [27]. He then used insights from the pre-study, particularly regarding categories, to conduct a similar study as in Starke et al. [28]. The study shows some minor differences in how feature-specific similarity metrics perform across categories.

2.3 Key Differences

The use of similar-item retrieval can overcome recommender problems in the news domain related cold start and item [12] Past studies in this context have examined the use of feature-specific similarity functions on news articles from specific corpora in the USA, such as the Washington Post [28]. These employ the method of semantic similarity [30], where users are asked to judge the similarity between two items and to compare this to a computational approach of similarity. Previous work faced a number of limitations. Beyond the use of a limited number of news content, the metrics tend to be relatively simple (e.g., TF-IDF) not reflecting the state of the art. Moreover, there has been little attention for the context of news article, be it whether they are part of a local or national outlet. For example, local news might be geared towards links with specific communities (cf. [26]), using various named entities to emphasize these links. Finally, although the method of semantic similarity is a form of 'user validation', previous studies have not evaluated the recommendation appropriateness using quantitative methods [27, 28].

[2]https://trec.nist.gov/data/wapost/

Uniquely, this study investigates feature-specific similarity functions using human judgments for Norwegian language news. This is a first in this domain where previous investigations have been conducted primarily for English language news. This detailed analysis includes not only national-level news, as previous studies have done, but also local-level news, allowing for a more nuanced view of different outlet levels. In terms of metrics, this study applies recent developments in Natural Language Processing (NLP) to evaluate their effectiveness in representing human similarity judgments. This provides novel insights into the capabilities of current state-of-the-art NLP methods, an aspect overlooked in previous work.

3 METHODS

3.1 Dataset

The dataset used for this study is a combination of data from four separate outlets from two separate media organizations[3]. The datasets were obtained through the MediaFutures research center[4] and consist of outlets from two of the MediaFutures industry partners, Amedia[5] and Schibsted[6]. The datasets followed the following criteria:

- **Contain Local and National news.** The main research question of this study is to find any differences between Human Similarity Judgments between the National and Local news domains. Available large-scale datasets were considered, but none were found to have the sufficient geographical granularity to isolate a clear *local* news domain. Because of this, it was decided that a specific dataset would have to be obtained or created.
- **Participant availability.** One challenge identified early on was the potential struggle of obtaining participants for the Human Similarity Judgment survey. Considering that a local news domain would also require local participants for the survey, overly restricting the definition of *local*, or restricting it to an area where potential participants are difficult to contact, could create unwanted challenges. Because of this, the local domain was chosen to be the Bergen area. As a result of this, the national domain is Norway.
- **Recency.** In the news domain time is a very important factor. The lifespan of breaking news is generally very short, down to a few hours [5, 6]. To avoid the problem of recency affecting the similarity ratings, we avoided recent news but avoided news older than one year. Because of this, we collected news articles from 2022.
- **Comparable Features.** Since this study builds upon previous studies [27, 28, 31], we performed comparative analyses. The features selected are therefore either aligned with previous work or novel (cf. Section 3.2).

3.1.1 Outlets. The dataset includes articles from different Norwegian news sources. These stem from two different news organizations, from which we selected both a local and one national newspaper. For Amedia the datasets include the outlets *Bergensavisen (BA)* and *Nettavisen*. BA is the most local newspaper across the dataset, with its main audience in Bergen and surrounding areas. Nettavisen functions as the national newspaper in the Amedia context of the dataset. Its audience is all of Norway, and ranks 7th in daily online readership.

The Schibsted outlets included are *Bergens Tidende (BT)* and *Verdens Gang (VG)*. BT is the largest newspaper of Western Norway, with its base in Bergen. Its audience is all of Vestland county. In the dataset BT is the local newspaper for the Schibsted context. VG is Norway's largest online newspaper by readership, and its audience is all of Norway. It functions as the national newspaper in the Schibsted context. Figures for the outlets can be seen in Table 1.

[3]The judgments given to the news article dataset will be shared in a repository upon acceptance.
[4]https://mediafutures.no/about/
[5]https://www.amedia.no/english
[6]https://schibsted.com/about/we-are-schibsted/news-media/
[7]https://www.medietall.no/index.php?liste=persontall&r=PERSONTALL

Table 1. Statistics of the outlets in the dataset: Q4 2022 Norwegian readership ranks and daily readership[7] for online versions, the raw and cleaned amount of articles, number of sections, average number of tags, average amount of tokens in the body text and titles.

Outlet	Rank	Readers	Raw Articles	Articles	Sections	$\overline{\text{Tags}}$	$\overline{\text{Text}}$	$\overline{\text{Title}}$
VG	# 1	1 957 961	17 686	11 587	33	4.05	701.02	9.16
Nettavisen	# 7	529 582	20 051	5 468	20	3.46	720.18	10.33
BT	# 16	184 514	17 444	13 808	26	4.51	654.99	9.67
BA	# 22	97 658	8 653	5 865	20	3.99	662.29	10.75

Table 2. News article features used in study.

Feature	Description
Date	The UNIX-time of the publication date
Section	List containing Section or Sections
Tags	List of manually added tags
Title	Title text
Text	Main body text
Image	The main image

3.1.2 Dataset Cleaning. The final dataset contained 36,768 articles which were all published in 2022. The 'raw' dataset was larger (cf. Table 1); to increase the dataset's similar pair diversity, we removed articles on dominant topics like Covid-19, the War in Ukraine, and the Power crisis, based on insights from [27, 28]. Articles were filtered using available journalist tags, with manual review to ensure effectiveness. This approach also helped eliminate periodical articles and those with high similarity within certain tag groups. In addition, we removed incomplete articles, such as those without images and key features as listed in Table 2. Short and long articles were also omitted, removing those with body texts shorter than 1000 characters or longer than 10000 characters were excluded, amounting to the 3% shortest and longest in the dataset. Finally, within each outlet, articles with duplicate titles and text were also removed. Key figures of the datasets after cleaning can be seen in Table 1.

3.2 News Article Features

The selection of features was based on earlier work [27, 28, 31], of which a list is presented in Table 2. A main difference with earlier work was the section feature. In Starke et al. [28] the category feature was used to represent a *subcategory*, while in Solberg [27] a feature named *topic* had similar properties. Where in both studies articles were limited to a single parent category, the current study included multiple categories across across entire outlets. In the Schibsted datasets, this was called *section*, while Amedia utilized a feature named *predicted category*. The *section* feature in this study had a higher granularity than simple categories, which could usually be mapped to a parent category. Another difference is the tag feature, which were added in both the Amedia and Schibsted datasets manually added by the newsrooms, and represented the news content.

3.3 Metrics

As our work builds directly on top of the work done in [31] [28] and [27], several of the metrics used are shared with them. A full list of the similarity metrics and the features they are used on can be seen in Table 3.

When calculating the similarity of the *Image* metrics we used used a similar approach as [31], [28] and [27]. Similarity is compared based on *Brightness*, *Sharpness*, *Contrast*, *Colorfulness* and *Entropy*. To compute similarity, the individual low-level feature was calculated and then compared using Manhattan distance. As in [28, 31], the low-level image features were extracted using the OpenIMAJ library[8] as proposed by San Pedro and Siersdorfer [24] [31].

In addition to the low-level features, Image Embeddings were also extracted. Following the method proposed by [25] and also used in Trattner and Jannach [31] and Starke et al. [28], we used an embedding from the first fully-connected layer of a pre-trained (ImageNet) VGG-16 model.

Following the method used in Starke et al. [28], text similarity was calculated using two TF-IDF, as well as LDA topic modeling. In addition to the two TF-IDF algorithms used in [28], an algorithm utilizing lemmatized text was also used (TF-IDF-L), based on findings in Balakrishnan and Ethel [1]. Three metrics utilizing pre-trained large language models were also used. Following findings in Solberg [27], named entities were extracted and a metric utilizing Jaccard similarity was devised (NENTS). In addition to LDA, topics were modeled using BERTopic [9], the similarity metric for BERTopic compared vectors of topic predictions using cosine similarity. Finally, text embeddings were extracting using a pre-trained Sentence Transformer (SBERT) model [21][9] and compared using cosine similarity.

Similar to Starke et al. [28], the title similarity was evaluating using 4 edit-distance based metrics, as well as LDA topic modeling and TF-IDF. In addition we used the Sentence Transformer, BERTopic and Lemmatized TF-IDF metrics, which were also used on the main article text.

In line with Starke et al. [28], Section similarity was calculated using Jaccard similarity. In addition, similarity of the publication date was calculated by comparing the difference in publication date, divided by the total date range of the dataset. Finally, tags-similarity was calculated using Jaccard on the list of tags for each article.

3.4 Experiment

3.4.1 Procedure. Users were invited to join a study on news recommendation and similarity[10]. Upon starting the survey, they were first randomly assigned to a group of either Amedia context or Schibsted context. Once assigned, we semi-randomly formed 10 article pairs, which would be presented to each user: 5 from the local media outlet and 5 from the national outlet. Each pair belonged to a specific sample bin outlined in section 3.4.2.

For each pair, users needed to rate the similarity between the two news articles on a 5-point scale. As in [27, 28], the users were also asked about their familiarity with the presented articles and the confidence they had in their similarity ratings. In order to explore recommendation appropriateness, we also asked the users to what extent they would agree with the statement that they would like to be recommended article 1 after seeing article 2, and vice versa. In addition, we also inquired on basic demographics and news use frequency.

3.4.2 Sampling Strategy. The pairs were formed using methods similar to Starke et al. [28]. As outlined in section 3.1, the dataset was divided by outlet, and the 25 metrics (Table 3) were applied to each subset. This resulted in four similarity score matrices for each of the outlet's news article pairs, using equal weight calculations. To avoid problems with low similarity strength, as observed in [27, 28], we used a strategy that placed news articles in similarity strength 'bins'. We computed the standard deviation of the pairwise similarity scores and then divided pairs into the following sampling bins:

[8] http://www.openimaj.org/

[9] Specifically, using the *nb-sbert-base*, https://huggingface.co/NbAiLab/nb-sbert-base[16]

[10] This research adhered to the ethical guidelines of the Research council of [Country] and the guidelines of [University] for scientific research. It was judged to pass without further extensive review, for it contained no misleading information, stress tasks, nor would it elicit extreme emotions.

Table 3. Full list of similarity metrics and the features they are applied to. Metrics not used in [31] or [28] are denoted by *.

Name	**Metric**	**Explanation**
Image:BR	$sim_{BR}(s,t) = 1 - \lvert BR(s) - BR(t) \rvert$	Brightness Distance
Image:SH	$sim_{SH}(s,t) = 1 - \lvert SH(s) - SH(t) \rvert$	Sharpness Distance
Image:CO	$sim_{CO}(s,t) = 1 - \lvert CO(s) - CO(t) \rvert$	Contrast Distance
Image:COL	$sim_{COL}(s,t) = 1 - \lvert COL(s) - COL(t) \rvert$	Colorfulness Distance
Image:EN	$sim_{EN}(s,t) = 1 - \lvert EN(s) - EN(t) \rvert$	Entropy Distance
Image:EMB	$sim_{EMB}(s,t) = \frac{EMB(s) \cdot EMB(t)}{\lVert EMB(s) \rVert \, \lVert EMB(t) \rVert}$	Embedding Cosine
Text:BERTopic*	$sim_{BERTopic}(s,t) = \frac{BERTopic(s) \cdot BERTopic(t)}{\lVert BERTopic(s) \rVert \, \lVert BERTopic(t) \rVert}$	BERTopic Cosine
Text:LDA	$sim_{LDA}(s,t) = \frac{LDA(s) \cdot LDA(t)}{\lVert LDA(s) \rVert \, \lVert LDA(t) \rVert}$	LDA Cosine
Text:NENTS*	$sim_{NENTS}(s,t) = \frac{\lvert NENTS(s) \cap NENTS(t) \rvert}{\lvert NENTS(s) \cup NENTS(t) \rvert}$	Named-Entities Jaccard
Text:SBERT*	$sim_{SBERT}(s,t) = \frac{SBERT(s) \cdot SBERT(t)}{\lVert SBERT(s) \rVert \, \lVert SBERT(t) \rVert}$	SBERT Cosine
Text:TF-IDF	$sim_{TF-IDF}(s,t) = \frac{TF-IDF(s) \cdot TF-IDF(t)}{\lVert TF-IDF(s) \rVert \, \lVert TF-IDF(t) \rVert}$	Stem TF-IDF Cosine
Text:TF-IDF-50	$sim_{TF-IDF}(s,t) = \frac{TF-IDF(s) \cdot TF-IDF(t)}{\lVert TF-IDF(s) \rVert \, \lVert TF-IDF(t) \rVert}$	50 first TF-IDF Cosine
Text:TF-IDF-L*	$sim_{TF-IDF}(s,t) = \frac{TF-IDF(s) \cdot TF-IDF(t)}{\lVert TF-IDF(s) \rVert \, \lVert TF-IDF(t) \rVert}$	Lemma TF-IDF Cosine
Time:Days	$sim_{DAYS}(s,t) = \left\lvert \frac{s_d - t_d}{max(D) - min(D)} \right\rvert$	Days Distance
Section:JACC	$sim_{JACC}(s,t) = \frac{\lvert Section(s) \cap Section(t) \rvert}{\lvert Section(s) \cup Section(s) \rvert}$	Section Jaccard
Tags:JACC	$sim_{JACC}(s,t) = \frac{\lvert Tags(s) \cap Tags(t) \rvert}{\lvert Tags(s) \cup Tags(s) \rvert}$	Tags Jaccard
Title:BERTopic*	$sim_{BERTopic}(s,t) = \frac{BERTopic(s) \cdot BERTopic(t)}{\lVert BERTopic(s) \rVert \, \lVert BERTopic(t) \rVert}$	BERTopic Cosine
Title:LDA	$sim_{LDA}(s,t) = \frac{LDA(s) \cdot LDA(t)}{\lVert LDA(s) \rVert \, \lVert LDA(t) \rVert}$	LDA Cosine
Title:SBERT*	$sim_{SBERT}(s,t) = \frac{SBERT(s) \cdot SBERT(t)}{\lVert SBERT(s) \rVert \, \lVert SBERT(t) \rVert}$	SBERT Cosine
Title:TF-IDF	$sim_{TF-IDF}(s,t) = \frac{TF-IDF(s) \cdot TF-IDF(t)}{\lVert TF-IDF(s) \rVert \, \lVert TF-IDF(t) \rVert}$	Stem TF-IDF Cosine
Title:TF-IDF-L*	$sim_{TF-IDF}(s,t) = \frac{TF-IDF(s) \cdot TF-IDF(t)}{\lVert TF-IDF(s) \rVert \, \lVert TF-IDF(t) \rVert}$	Lemma TF-IDF Cosine
Title:BI	$sim_{BI}(s,t) = 1 - \lvert dist_{BI}(s,t) \rvert$	BiGram Distance
Title:JW	$sim_{JW}(s,t) = 1 - \lvert dist_{JW}(s,t) \rvert$	Jaro-Winkler Distance
Title:LCS	$sim_{LCS}(s,t) = 1 - \lvert dist_{LCS}(s,t) \rvert$	LCS Normalized
Title:LV	$sim_{LV}(s,t) = 1 - \lvert dist_{LV}(s,t) \rvert$	Levenshtein Distance

(1) Pairs below 2 standard deviations below the mean similarity strength.
(2) Pairs between 2 and 1 standard deviation below the mean similarity strength.
(3) Pairs between 1 standard deviation below the mean and 1 standard deviation above the mean similarity strength.
(4) Pairs between 1 and 2 standard deviations above the mean similarity strength.
(5) Pairs above 2 standard deviations above the mean similarity strength.

For each media outlet, we sampled one pair from each bin. The results of applying this strategy to the pairwise similarity scores can be seen in Table 4. Once the scores were divided into groups, 1 000 pairs were randomly sampled from each bin for each outlet and added to the survey database. This resulted in 5 000 pairs for each outlet and 20 000 pairs available in total.

Table 4. Amount of pairs and percentages per sample bin. Bin 1 is least similar and bin 5 is most similar.

	Nettavisen		BA		VG		BT	
Bin	# Pairs	%	# Pairs	%	# Pairs	%	# Pairs	%
1	97 506	0.3%	180 128	0.5%	1 099 918	0.6%	926 140	0.7%
2	3 569 870	11.9%	4 368 158	12.7%	24 675 638	12.9%	17 959 504	13.4%
3	21 826 506	73.0%	24 941 396	72.5%	135 158 032	70.9%	95 643 946	71.2%
4	3 058 926	10.2%	3 463 902	10.1%	22 400 166	11.7%	14 797 390	11.0%
5	1 340 748	4.5%	1 438 776	4.2%	7 313 302	3.8%	4 920 002	3.7%

Table 5. Segmentations of the participants and pairs for the analysis in the chapter. The pairs in the *pass* groups include the removal of the attention check ratings. Participants are divided into *Local* and *National* groups depending on their reported place of residence. *Bergen* and *Bergen Area* are considered *Local*.

	Participants			Pairs				
	Total	Local	National	Total	VG	BT	Nettavisen	BA
All	141	108	33	1410	365	365	340	340
Pass	119	91	28	1071	287	289	249	246

3.4.3 Participants. Participants were recruited by sharing the survey link across relevant social media channels. In total 329 participants started the survey with 143 completions. 2 of the participants were below 18 years old and were removed from the results, bringing the total number of participants to 141. 73 of the participants completed the Schibsted context, giving ratings to pairs from BT and VG, while 68 completed the Amedia context, giving ratings to pairs from BA and Nettavisen.

119 out of 141 participants, or 84.4%, passed the attention check. After accounting for the attention check, ratings for 1071 news pairs (featuring 1968 unique news articles) were available from users who passed the attention check. The final figures for the segmentation of participants and pairs are described in Table 5. The results are calculated using only the participants and pairs that passed the attention check. In addition, the pairs that had the attention check are removed as the attention check interfered with the ratings given[11].

A total of 112 participants, 79.4%, reported their frequency of news reading to be *approximately every day*. This is higher than in the previous work, and somewhat higher than expected. 81 participants were male while 59 were female. The largest age group was 25-34 with 55 participants, followed by 35-44 with 35 and 18-24 with 25.

4 RESULTS

4.1 Comparing Metrics to Human Judgments (RQ1)

We examined the extent to which *Feature-Specific Similarity Metrics* relate to *Human Similarity Judgments*. In order to compare the Similarity Metrics to the Similarity Judgments, Spearman correlations were computed between the metrics listed in Section 3.3 and the Human Similarity Judgments collected through the survey. The results per metric are described in Table 6, which are also divided on local vs national domains and outlet (to address RQ2 later).

We discuss Table 6 from top to bottom. Among the Image-based metrics, *Image:EMB* demonstrated the highest correlation to Human Similarity Judgments, registering a correlation of 0.30. This correlation was especially high for

[11]The attention check replaced the body text with a message to give ratings of 3 on all parameters if the text was read.

Table 6. Similarity metric correlation (Spearman) with human similarity judgments. Metrics are listed in the left column, with Spearman correlations for the various divisions of the datasets listed in the other columns. *All* combines the pair ratings of all outlets. *National* combines VG & Nettavisen, *Local* combines BT & BA. For the features with several metrics, the metric with the highest correlation can be seen in bold. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Metric	**News Outlet**						
	All	**National**	**Local**	**VG**	**BT**	**Nettavisen**	**BA**
Image:BR	0.24***	0.16***	**0.32*****	0.06	**0.36*****	0.26***	0.27***
Image:SH	0.26***	0.24***	0.28***	0.08	0.28***	0.40***	0.27***
Image:CO	0.13***	0.11*	0.15***	0.12*	0.15*	0.10	0.15*
Image:COL	0.07*	0.07	0.08	0.11	0.11	0.05	0.04
Image:EN	0.22***	0.15***	0.28***	0.09	0.29***	0.21***	0.27***
Image:EMB	**0.30*****	**0.39*****	0.23***	**0.32*****	0.20***	**0.46*****	**0.28*****
Text:BERTopic	0.40***	0.42***	0.37***	0.39***	0.36***	0.46***	0.39***
Text:LDA	0.29***	0.29***	0.29***	0.34***	0.33***	0.29***	0.26***
Text:NENTS	0.21***	0.22***	0.2***	0.12*	0.27***	0.36***	0.14*
Text:SBERT	**0.60*****	**0.58*****	**0.62*****	**0.51*****	**0.63*****	**0.65*****	**0.60*****
Text:TF-IDF	0.47***	0.45***	0.48***	0.38***	0.49***	0.52***	0.47***
Text:TF-IDF-50	0.17***	0.14**	0.2***	0.18**	0.17**	0.08	0.24***
Text:TF-IDF-L	0.47***	0.44***	0.49***	0.38***	0.49***	0.49***	0.49***
Time:Days	0.22***	0.20***	0.24***	0.17**	0.25***	0.23***	0.23***
Section:JACC	0.49***	0.47***	0.50***	0.36***	0.58***	0.62***	0.59***
Tags:JACC	0.33***	0.36***	0.30***	0.25***	0.25***	0.45***	0.42***
Title:BERTopic	0.30***	0.28***	0.32***	0.20***	0.24***	0.35***	0.43***
Title:LDA	0.07*	0.04	0.10	0.04*	0.20***	0.05	-0.07
Title:SBERT	**0.38*****	**0.38*****	**0.39*****	**0.35*****	**0.45*****	**0.41*****	**0.33*****
Title:TF-IDF	0.20***	0.19***	0.2***	0.09	0.16**	0.28***	0.24***
Title:TF-IDF-L	0.17***	0.15***	0.18***	0.09	0.11	0.20**	0.25***
Title:BI	0.18***	0.19***	0.16***	0.16**	0.13**	0.21***	0.21***
Title:JW	0.21***	0.2***	0.21***	0.14*	0.23***	0.26***	0.18**
Title:LCS	0.22***	0.27***	0.17***	0.19**	0.22***	0.35***	0.10
Title:LV	0.18***	0.19***	0.16***	0.16**	0.12*	0.22***	0.22***

Nettavisen (0.46). Curiously, in the VG dataset, the low level image feature metrics all demonstrated correlations too low to be statistically significant.

Overall, the *Text:SBERT* metric (0.60) presented the highest correlation across all divisions of the dataset. This would suggest that SBERT on body text was most representative of human similarity judgments. This outperformed the *Text:TF-IDF* metric (0.47), which was the highest correlating metric in studies of Starke et al. [28] (0.29) and Solberg [27] (0.53). The *Text:TF-IDF-L* metric showed similar correlations as the *Text:TF-IDF* metric. The *Text:BERTopic* metric (0.40) outperformed the other topic modeling metric, *Text:LDA* metric (0.29). The outlets with larger datasets showed higher correlations with the *Text:LDA* metric, specifically VG (0.34) and BT (0.33), compared to those with smaller datasets like Nettavisen (0.29) and BA (0.26). The same observation can not be made with *Text:BERTopic*. The *Text:NENTS* metric (0.21) had a wide range of correlations depending on the outlet, with VG showing the lowest correlation (0.12) and Nettavisen the highest (0.36).

Table 7. Results of similarity evaluations across national and local domains, as well as recommender appropriateness. Bin 1 is the least similar and 5 is the most similar article pairs (cf. Section 3.4.2). **Left section**: Similarity scores for all pairs, pairs from local outlets, and, pairs from national outlets. **Middle sections**: Student's t-test and Wilcoxon signed-rank test on the local pair ratings vs national pair ratings of the participants. **Right section**: Recommender appropriateness average response (Score), correlation between the score and article pair similarity (Sim.corr.), and, correlation between the recommender appropriateness of each of the two articles in the pair (Art.corr).

	Similarity scores			**Students t-test**		**Wilcoxon test**		**Appropriateness**		
Bin	All	Local	National	t	p	W	p	Score	Sim.corr.	Art.corr.
All	2.13	2.19	1.36	1.896	0.060	2724.0	0.048	2.46	0.54	0.84
5	3.45	3.59	3.30	1.280	0.204	608.5	0.169	3.26	0.40	0.92
4	2.63	2.89	2.38	2.801	0.006	834.5	0.006	2.81	0.50	0.85
3	1.79	1.71	1.88	-1.313	0.118	339.5	0.229	2.21	0.46	0.77
2	1.37	1.31	1.44	-1.682	0.480	100.0	0.080	2.00	0.22	0.79
1	1.38	1.41	1.36	0.575	0.566	175.0	0.507	2.04	0.42	0.77

The *Title:SBERT* metric demonstrated the highest correlation (0.38) among Title-based metrics, followed by *Title:BERTopic* (0.30). For *Title:SBERT*, BT showed a considerably higher score (0.45) than BA (0.33). Conversely, *Title:BERTopic* presented a higher score for BA (0.43) than for BT (0.24). The *Title:LDA* metric displayed very low scores (0.07), with the exception of BT, which showed a slightly higher correlation (0.2). This suggested NLP-based metrics, such as SBERT, would outperform TF-IDF metrics that were used previously. Furthermore, *Section:JACC* showed high correlations of 0.49. The correlations were particularly high for the Amedia outlets, with 0.59 for BA and 0.62 for Nettavisen, compared to lower correlations observed for the Schibsted outlets, specifically 0.36 for VG. The *Tags:JACC* metric showed high variation between the two datasets, with 0.25 for VG and BT, and 0.45 and 0.42 for Nettavisen and BA.

4.2 RQ2: National vs Local News Domains

4.2.1 Differences in human similarity judgments. We compared ratings given to pairs from local and national outlets using t-tests and Wilcoxon signed-rank tests. The latter, a non-parametric test, was necessary to account for users usually providing scores on the extremes of the 5-point similarity scale. Moreover, as the attention check replaced a random pair, the corresponding national or local pair were removed, bringing the total pairs evaluated to 952. This was due to the t-test was performed by evaluating the similarity rating of the pairs with the same similarity bin, across the two publications.

The results from the tests are outlined in Table 7. The most significant finding is that the ratings for bin 4 are higher for local outlets than for national outlets. The same findings can be seen in the Wilcoxon signed-rank test. In the Wilcoxon signed-rank test we also see that when considering all sample bins, the similarity ratings for the local outlets are slightly higher (p=0.48).

4.2.2 Change in metrics. In order to evaluate how the changes found in section 4.2 we performed Fisher r-to-z transformations on the correlations calculated in on a selection of the correlations calculated in Table 6. The z-values were then pairwise compared by performing a Z-test. This was performed on various compositions of national and local outlets.

By performing this analysis, 3 metrics stood out. These are *Image:EMB*, *Section:JACC* and *Title:BERTopic*. The results are described in Table 8. The *Image:EMB* did show similar differences across all divisions of the outlets. However, the

Table 8. Results of Z-test comparing national vs local news feature correlation after performing Fisher-r-to-z on the data in Table 6. All: VG and Nettavisen vs BT and BA. Schibsted: VG vs BT. Amedia: Nettavisen vs BA. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Metric	All	Schibsted	Amedia	VG vs BA
Image:EMB	2.819**	1.543	2.291*	0.416
Section:JACC	-0.726	-3.377***	0.574	-3.442***
Title:BERTopic	-0.760	-0.416	-1.100	-2.920**

strength was the weakest when comparing the most local and most national outlet. The *Section:JACC* metric also show high strength on some divisions. But it should also be considered that this metric shows weaker correlation when considering the ratings for VG alone. Finally, the *BERTopic* showed similar results across all divisions of the outlets. It also had the highest strength when evaluating VG and BA. However, it was only significantly higher when evaluating VG vs BA. To investigate this further, we also evaluated the *Title:BERTopic* z-score between *Nettavisen* and *VG* which returned a z-score of -1.788 with a p-value of 0.074.

4.3 RQ3: Recommender Appropriateness

We finally examined the user's perceived recommendation appropriateness, in relation to the inter-article similarity. This was based on whether users would like to be recommended one of the articles in a pair after seeing the other. The results are described in Table 7. It was observed that the overall Spearman correlation between similarity and recommender appropriateness is 0.54, which suggested a moderate relation between similarity and appropriateness. Most notably, the score for appropriateness increased per similarity strength bin, except between bins 1 and 2.

The final column of Table 7 describes the symmetry of the appropriateness rating. This meant whether the appropriateness rating for liking article 1 after 2 was similar to the rating for 2 after 1. We found this correlation to be relatively high: 0.84.

5 DISCUSSION & CONCLUSION

5.1 Representativeness of Feature-Specific Similarity Metrics (RQ1)

We have examined to what extent different feature-specific similarity metrics represent human judgments of similarity. The goal is to identify metrics can be used in content-based recommenders that users like to use, for they represent their judgment and preferences.

One of the primary findings is the effectiveness of the BERT-based metrics for news recommendation. Particularly SBERT, which has not been used often in this context [14], shows higher correlations than the other metrics on both of the features where it is used and also the highest correlation across all metrics when it is used on the body text of the article. This is surprising considering the basic implementation, including a limitation of the first 512 words of the article. This is lower than the median amount of words per article in the dataset. SBERT is primarily designed to create embeddings for sentences, and that may explain the higher relative correlations in the title feature than the text feature when compared to TF-IDF.

The BERTopic metrics also showed comparably high correlations, especially on the title feature where it is the second-highest correlating metric after SBERT when considering all ratings. Considering the VG and BA news outlets we see that the range of correlations is fairly high. When we also consider BT and Nettavisen, and the size of the various datasets, it may indicate that BERTopic's correlation decreases based on the number of articles in the dataset. This is most likely related to the training setup, and the high modularity of BERTopic might allow for setups that are more tailored toward finding document similarity, especially in larger datasets.

The high correlation in this study between human judgment and *Section:JACC* (0.49) compared to Starke et al. [28] (0.14) is notable. This could be due to the larger variety in the dataset, in terms of the different types of categories used. The difference in correlations between Schibsted and Amedia outlets is likely due to Amedia using predictive models to determine categories, while Schibsteds selection is editorial.

The *Tags:JACC* metric shows significantly higher correlations in Amedia outlets than in Schibsted outlets. This discrepancy could indicate differences in tagging strategy between the two, with potential implications for similar item recommendation purposes.

Curiously, the *Title:LDA* shows some weak correlation when looking at the BT pair ratings alone. Except for a study on similarity judgments in the the recipe domain [31], *Title:LDA* have failed to show any correlations with human judgment. This suggests that the amount of information in the titles of news articles is insufficient to generate a topic model using LDA. Hence, we would suggest to avoid this metric for content-based recommenders.

The correlations for the *Text:NENTS* metric are lower than expected and show a wide range across the different outlets. This suggests that it may be more effective in certain contexts. This aligns with findings from Solberg [27], where it was found to be more relevant for the *Sports* category than the *Recent Events* category.

5.2 Local and National domains (RQ2)

We have further examined the extent to which the performance of similarity metrics depends on the locality of a media outlet. We have observed some minor differences in human similarity judgments between national and local news domains, but most differences in correlational strength are not statistically significant. Local news article pairs are considered slightly more similar than national news overall, particularly in bins that were computationally more similar as well. This suggests that users mostly recognize well-matched news articles to be similar, but that worse matches are perceived as more distant than national news. While most differences are not significantly different, it does indicate that subtle changes in similar-item recommendation strategies can be made in news recommenders across different geographical granularities. However, the error that would be made by ignoring this is not large.

The differences in similarity judgments are not sufficient to impact the feature-specific similarity metrics to a large extent. While a couple of metrics do appear to be impacted, much of the differences probably could be explained by outlet-specific differences. Because of a lack of previous studies in this specific area, it is difficult to judge the magnitude of these small differences. The *Title:BERTopic* metric is an example of this. While it is intuitive that titles may differ between local and national domains, when checking the z-score between Nettavisen and VG, there are indications that these differences are primarily related to properties specific to the VG outlet, and not the national and local domains. While some studies outside the technological domains suggest regionality or locality may matter [26], recent studies within NRS indicate it is not as important [15]. We still recommend investigating these differences in more detail.

5.3 Recommender Appropriateness (RQ3)

We have also examined to what extent users perceive the two articles in a pair as good recommendations, when first seeing one or the other article. This research question has built upon earlier work from Yao and Harper [35] and Solberg [27], which indicate that similarity only may not be the most important factor in similar-item recommendation approaches. We have mainly turned the presented pairs into hypothetical recommendation scenarios, of which the appropriateness was judged.

We have found that the appropriateness of the recommendations is correlated with the perceived similarity, as well as the computed similarity. This is shown through the correlation between the judgment and the appropriateness, as well as through the increasing appropriateness along computational correlational strength. Although the overall correlation is only moderately strong (0.54), it does show a clear relation between similarity and appropriateness. Nonetheless, there still remains quite some unexplained variance which may be explained by other factors. These findings may support the understanding of what extent to news recommenders should evaluate similarity in the recommendations, contributing to the foundation of hybrid news recommenders.

6 LIMITATIONS AND FUTURE WORK

This study has faced a few limitations. Contrary to previous studies with US- and UK-based dataset, we have focused on Norway. While we do not see any particular reason to expect large cultural differences between these countries, the local news context is rather specific. The city of Bergen is used as the *local* domain for this study, which is a moderately large city for Northern European standards (200K-250K inhabitants). As such, the outlets chosen may not contain some properties that are associated with *local* news. BT in particular aims to provide users with the full spectrum of news, including foreign affairs. Amedia on the other hand mainly focuses on local news, with Nettavisen being their only general national newspaper, opposed to their 89 local newspapers. It can therefore be speculated that Nettavisen may contain properties otherwise reserved for local newspapers. While this study compares local and national domains, we do not compare different local domains. Investigating different local news across different populations may yield interesting results.

A main omission, observed in some other news recommender studies as well [12], is the lack of a naturalistic context. We have recruited participants from social media platforms and not, say, regular readers of a digital news website. Moreover, no recent news has been considered, which may impact the recommendation appropriateness beyond similarity. These factors should be incorporated when designing news recommenders of the future.

Building on the findings in this study, future research could test our findings in a news recommendation scenario. The SBERT metric used in the study has clear limitations in that SBERT models are designed for sentence level texts, and includes a limitation of 512 tokens. Metrics based on other LLM embeddings, especially those that are designed for full texts, should be investigated. Another next step would also be to further develop a recommendation metric by training models on the metrics in this dataset, to facilitate automated news recommendation, also to users who are not logged in.

ACKNOWLEDGMENTS

This research is funded by SFI MediaFutures partners and the Research Council of Norway (grant number 309339). Generative AI has been used during the writing of this paper. Specifically, it has been used to create drafts and to summarize and condense text, which has then been edited. Among other, both abstract and the title have produced by generating several drafts that has then been combined and edited to the final text. In all cases either OpenAI ChatGPT or Microsoft Bing Chat have been used, both based on GPT4.

REFERENCES

[1] Vimala Balakrishnan and Lloyd-Yemoh Ethel. 2014. Stemming and Lemmatization: A Comparison of Retrieval Performances. *Lecture Notes on Software Engineering* 2 (2014), 262–267. https://doi.org/10.7763/LNSE.2014.V2.134

[2] Daniel Billsus and Michael J. Pazzani. 2000. User Modeling for Adaptive News Access. *User Modeling and User-Adapted Interaction* 10, 2 (2000), 147–180. https://doi.org/10.1023/A:1026501525781

[3] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. 2007. Measuring semantic similarity between words using Web search engines. In *Proceedings of the 16th International Conference on World Wide Web* (Banff, Alberta, Canada) *(WWW '07)*. Association for Computing Machinery, New York, NY, USA, 757–766. https://doi.org/10.1145/1242572.1242675

[4] Benjamin P. Chamberlain, Emanuele Rossi, Dan Shiebler, Suvash Sedhain, and Michael M. Bronstein. 2020. Tuning Word2vec for Large Scale Recommendation Systems. In *Proceedings of the 14th ACM Conference on Recommender Systems* (Virtual Event, Brazil) *(RecSys '20)*. Association for Computing Machinery, New York, NY, USA, 732–737. https://doi.org/10.1145/3383313.3418486

[5] Wei Chu, Seung-Taek Park, Todd Beaupre, Nitin Motgi, Amit Phadke, Seinjuti Chakraborty, and Joe Zachariah. 2009. 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* (Paris, France) *(KDD '09)*. Association for Computing Machinery, New York, NY, USA, 1097–1104. https://doi.org/10.1145/1557019.1557138

[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* (Banff, Alberta, Canada) *(WWW '07)*. Association for Computing Machinery, New York, NY, USA, 271–280. https://doi.org/10.1145/1242572.1242610

[7] Mehdi Elahi, Dietmar Jannach, Lars Skjærven, Erik Knudsen, Helle Sjøvaag, Kristian Tolonen, Øyvind Holmstad, Igor Pipkin, Eivind Throndsen, Agnes Stenbom, Eivind Fiskerud, Adrian Oesch, Loek Vredenberg, and Christoph Trattner. 2022. Towards responsible media recommendation. *AI and Ethics* 2, 1 (2022), 103–114. https://doi.org/10.1007/s43681-021-00107-7

[8] Florent Garcin, Boi Faltings, Olivier Donatsch, Ayar Alazzawi, Christophe Bruttin, and Amr Huber. 2014. Offline and Online Evaluation of News Recommender Systems at Swissinfo.Ch. In *Proceedings of the 8th ACM Conference on Recommender Systems* (Foster City, Silicon Valley, California, USA) *(RecSys '14)*. Association for Computing Machinery, New York, NY, USA, 169–176. https://doi.org/10.1145/2645710.2645745

[9] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).

[10] Hebatallah A. Mohamed Hassan, Giuseppe Sansonetti, Fabio Gasparetti, Alessandro Micarelli, and Joeran Beel. 2019. BERT, ELMo, USE and InferSent Sentence Encoders: The Panacea for Research-Paper Recommendation?. In *ACM Conference on Recommender Systems*.

[11] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. 2010. *Recommender Systems an Introduction*. Cambridge University Press, Leiden. http://www.amazon.com/Recommender-Systems-Introduction-Dietmar-Jannach/dp/0521493366

[12] Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. 2018. News recommender systems – Survey and roads ahead. *Information Processing & Management* 54, 6 (2018), 1203–1227. https://doi.org/10.1016/j.ipm.2018.04.008

[13] Mohadeseh Kaviani and Hossein Rahmani. 2020. EmHash: Hashtag Recommendation using Neural Network based on BERT Embedding. In *2020 6th International Conference on Web Research (ICWR)*. 113–118. https://doi.org/10.1109/ICWR49608.2020.9122275

[14] Peter Kolbeinsen Klingenberg. 2023. *Using content-and behavioural data for recommendations in the Norwegian news market*. Master's thesis. The University of Bergen.

[15] Erik Knudsen, Alain D. Starke, and Christoph Trattner. 2023. Topical Preference Trumps Other Features in News Recommendation: A Conjoint Analysis on a Representative Sample from Norway. In *Proceedings of the International Workshop on News Recommendation and Analytics, co-located with the 2023 ACM Conference on Recommender Systems (RecSys 2023) (CEUR Workshop Proceedings, Vol. 3561)*, B. Kille (Ed.). CEUR-WS, Singapore, 14. https://hdl.handle.net/11245.1/7ef6ea51-8e5d-458c-bc0e-f134028fc912 Singapore, 18 September 2023.

[16] Per E Kummervold, Javier De la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. 2021. Operationalizing a National Digital Library: The Case for a Norwegian Transformer Model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Linköping University Electronic Press, Sweden, Reykjavik, Iceland (Online), 20–29. https://aclanthology.org/2021.nodalida-main.3

[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* (Shanghai, China) *(BDE 2020)*. Association for Computing Machinery, New York, NY, USA, 18–23. https://doi.org/10.1145/3404512.3404524

[18] 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* (Shanghai, China) *(BDE 2020)*. Association for Computing Machinery, New York, NY, USA, 18–23. https://doi.org/10.1145/3404512.3404524

[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* (Hyderabad, India) *(WWW '11)*. Association for Computing Machinery, New York, NY, USA, 57–66. https://doi.org/10.1145/1963405.1963417

[20] Özlem Özgöbek, Jon Atle Gulla, and Riza Cenk Erdur. 2014. A Survey on Challenges and Methods in News Recommendation. In *International Conference on Web Information Systems and Technologies*.

[21] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Conference on Empirical Methods in Natural Language Processing*.

[22] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2011. *Introduction to Recommender Systems Handbook*. Springer US, Boston, MA, 1–35. https://doi.org/10.1007/978-0-387-85820-3_1

[23] Julio Rieis, Fabrício de Souza, Pedro Vaz de Melo, Raquel Prates, Haewoon Kwak, and Jisun An. 2021. Breaking the News: First Impressions Matter on Online News. *Proceedings of the International AAAI Conference on Web and Social Media* 9, 1 (Aug. 2021), 357–366. https://doi.org/10.1609/icwsm.v9i1.14619

[24] Jose San Pedro and Stefan Siersdorfer. 2009. Ranking and classifying attractiveness of photos in folksonomies. 771–780. https://doi.org/10.1145/1526709.1526813

[25] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs.CV]

[26] Helle Sjøvaag, Hallvard Moe, and Eirik Stavelin. 2012. Public service news on the Web: A large-scale content analysis of the Norwegian Broadcasting Corporation's online news. *Journalism Studies* 13, 1 (2012), 90–106.

[27] Vegard Rygh Solberg. 2022. *News Recommendation based on Human Similarity Judgment.* Master thesis. The University of Bergen. Masteroppgave i informasjonsvitenskap, INFO390, MASV-INFO.

[28] A.D. Starke, Sebastian Øverhaug Larsen, and Christoph Trattner. 2021. Predicting Feature-based Similarity in the News Domain Using Human Judgments. In *Proceedings of the 9th International Workshop on News Recommendation and Analytics (INRA 2021)*.

[29] Nava Tintarev and Judith Masthoff. 2006. Similarity for News Recommender Systems. In *Workshop on Recommender Systems and Intelligent User Interfaces*, Gulden Uchyigit (Ed.). In conjunction with the International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, AH 2006, Dublin, Ireland, June 20-23, 2006.

[30] 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.

[31] 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. https://doi.org/10.1007/s11257-019-09245-4

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. arXiv:1706.03762 [cs.CL]

[33] Amy Winecoff, Florin Brasoveanu, Bryce Casavant, Pearce Washabaugh, and Matthew Graham. 2019. Users in the Loop: A Psychologically-Informed Approach to Similar Item Retrieval.

[34] Nakyeong Yang, Jeongje Jo, Myeongjun Jeon, Wooju Kim, and Juyoung Kang. 2022. Semantic and explainable research-related recommendation system based on semi-supervised methodology using BERT and LDA models. *Expert Systems with Applications* 190 (2022), 116209. https://doi.org/10.1016/j.eswa.2021.116209

[35] 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* (Vancouver, British Columbia, Canada) *(RecSys '18)*. Association for Computing Machinery, New York, NY, USA, 288–296. https://doi.org/10.1145/3240323.3240351

[36] Qi Zhang, Jingjie Li, Qinglin Jia, Chuyuan Wang, Jieming Zhu, Zhaowei Wang, and Xiuqiang He. 2021. UNBERT: User-News Matching BERT for News Recommendation. 3356–3362. https://doi.org/10.24963/ijcai.2021/462

[37] Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020. Towards Topic-Guided Conversational Recommender System. arXiv:2010.04125 [cs.CL]