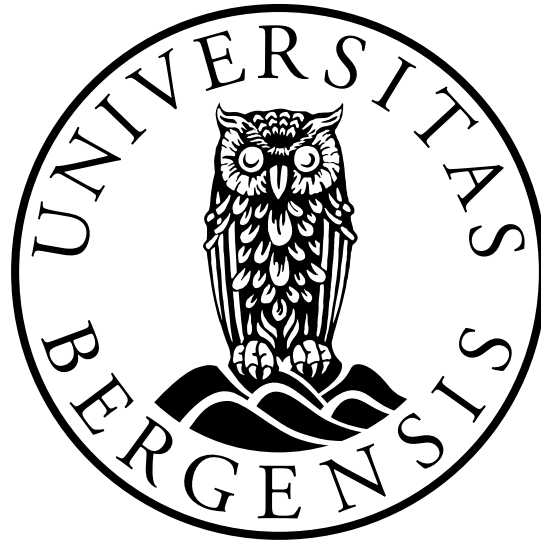


UNIVERSITY OF BERGEN



Department of Information Science and Media Studies

MASTER'S THESIS

**News Recommendation based on
Human Similarity Judgment**

Author: Vegard Rygh Solberg

Supervisor: Prof. Dr Christoph Trattner

Co-supervisor: Dr Alain D. Starke

December 6, 2022

Abstract

Similar item recommendation is one of the most popular types of recommender systems. As the name implies, the objective is to recommend items that are similar to a reference item. The news domain is one of the many that employ this form of recommendation, which utilize similarity functions in order to calculate the similarity. In this study, human judgments of article similarity were acquired using an online user study in which each of the 173 participants evaluated the similarity of 12 pairs of articles. Each of the 12 article pairs had their own unique characteristics. One pair would be made up of two completely dissimilar articles, while the other pairs had either a shared topic, a named entity in common, publication dates in close proximity, or some combination of these three characteristics. Half the pairs contained articles from the News (i.e., recent events) category, while the other half contained Sport articles. The similarity of the same article pairings was then calculated utilizing various similarity functions, and the correlation between human judgment and function scores was computed. This thesis found that the correlation ranged from weak to strong, depending on the function. The thesis also found that the correlation is largely dependent on whether the articles have certain characteristics in common. On average, the functions correlated more strongly to human judgment if the articles belonged to the category News (i.e., recent events) than Sport. The functions were also better at predicting human similarity when the articles in question were relatively similar to one another. The novel work presented in this thesis shows that the correlation between human judgment and similarity functions can be stronger than previous work has suggested, if news articles are paired in a meaningful way.

Acknowledgment

I would like to extend my sincere gratitude to my supervisors, Prof. Dr. Christoph Trattner and Dr. Alain D. Starke, for their guidance and support. The level of assistance they have offered throughout this process has been nothing short of remarkable.

A special thank you goes out to my roommate, who has routinely barged through my door at 8 in the morning these past few months, pushing me to get up and finish this project.

This research 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.

Bergen, Norway, December 2022

Vegard Rygh Solberg

Contents

Abstract	ii
Acknowledgment	v
1 Introduction	1
1.1 Motivation	1
1.2 Problem	2
1.3 Research Questions	2
1.4 Contribution	3
1.5 Thesis Outline	3
2 Background	5
2.1 News, Sport, and Recent Events	5
2.2 Recommender Systems	6
2.3 Recommender Scenario: News	7
2.4 Similarity Functions for News Retrieval	9
2.5 Human Perception of Similarity	10
2.6 Summary, Differences, and Contributions	12
3 Pre-study	15
3.1 Data Collection and Procedure	15
3.1.1 Research Design	15
3.1.2 Participants	16
3.2 Criteria for news recommendation (Q1)	17

3.2.1	Results	18
3.2.2	Discussion	18
3.3	Sport vs. Recent Events (Q2)	19
3.3.1	Results	20
3.3.2	Discussion	20
3.4	Similarity Factors (Q3)	21
3.4.1	Results	21
3.4.2	Discussion	21
3.5	Conclusion	22
4	Methodology	24
4.1	Constructing the Dataset	24
4.1.1	Requirements	24
Familiarity		24
Recency		25
Covid-19		25
4.1.2	Extracting Named Entities	25
4.1.3	Obtaining Articles	26
4.1.4	Data Structure	27
4.1.5	Categories	27
4.2	Similarity Functions	28
Title-based metrics		31
Subheading-based metrics		31
Topic-based metrics		32
Author-based metrics		32
Date-based metrics		32
BodyText-based metrics		32
AuthorBio-based metric		32
4.3	Research Design	32
4.3.1	Factors	32

Date	33
Topic	33
Named Entity	33
Category	33
4.3.2 Conditions	33
4.3.3 Dataset	34
4.3.4 Procedures and Measures	35
4.3.5 Participants	39
4.3.6 Statistical Analysis	41
5 Results	43
5.1 Similarity judgment and Function Correlation (RQ1)	44
5.2 Sports vs. Recent Events (RQ1.1)	47
5.3 Matching-characteristics (RQ1.2)	48
5.3.1 Multiple Linear Regression	50
5.3.2 Matching-characteristics and Correlation	53
5.4 Influence of demographic factors and user characteristics on Similarity judgment (RQ1.3)	56
5.4.1 Other factors	59
5.5 Information Cue Usage (RQ2)	59
6 Discussion and Future Work	62
6.1 Correlation between human judgment and various similarity functions (RQ1)	62
6.2 Is the correlation dependent on the category of the articles? (RQ1.1)	63
6.3 Is the correlation dependent on whether the articles have any shared matching-characteristics? (RQ1.2)	64
6.4 Are the similarity judgments dependent on demographic factors or user characteristics? (RQ1.3)	65
6.5 Which article features do readers employ when determining similarity between articles? (RQ2)	67
6.6 Limitations and Future Work	67

References**73**

List of Figures

2.1	Illustration of the categories used on the BBC homepage.	6
2.2	Illustration of how The Guardian recommends the most popular articles in addition to those that are related to the reference article.	9
3.1	A bar graph displaying the gender distribution among the participants.	16
3.2	A bar graph displaying the age distribution among the participants.	17
3.3	A bar graph displaying how often the participants read online news articles per week on average.	17
3.4	A bar graph displaying the criteria that individuals believe should be considered by news recommendation systems. Note: People were not confined to a single response; they could list as many factors as they wanted.	18
3.5	Illustration of the two BBC news articles used for Question 2 in the pre-study .	19
3.6	A bar graph displaying what people considered to be the single biggest factor that determines similarity between articles. Note: Six responses were omitted because it was impossible to discern with a high degree of certainty exactly what the respondents meant.	21
4.1	Illustration of the subcategories of Sport and Recent Events (News).	28
4.2	Illustration of the second phase of the user study.	36
4.3	Illustration of a pair of articles serving as an attention check. The first sentence of the body text has been replaced with text instructing users to give everything on the page a rating of 5.	37
4.4	Illustration of third phase of the user study.	38
4.5	A bar graph displaying the age distribution among the participants.	40
4.6	A bar graph displaying the gender distribution among the participants.	40

4.7	A bar graph displaying the gender distribution among the participants-	40
4.8	A bar graph displaying how often the participants read online news articles per week on average.	40
5.1	Spearman correlation matrix depicting how human judgment correlates with various feature-based functions. Reading tip: Correlation between humans and functions can be seen on the upper row. The rest is correlation between the various functions. Note: * $p < .05$, ** $p < .01$, *** $p < .001$	45
5.2	Tukey-HSD post hoc test result for human similarity judgment. News = Recent Events.	49
5.3	Tukey-HSD post hoc test result for BodyText:TF-IDF. News = Recent Events. . .	50
5.4	Tukey-HSD post hoc test result for information cue usage.	60

List of Tables

4.1	All articles were stored with the following information:	27
4.2	List of all the similarity function utilized in this work.	31
4.3	The 12 conditions from the 2x2x3 factorial design. Note: Dissimilar refers to the situation where an article pair is not matched on either Date, Topic, or Named Entity	34
4.4	Demographic questions and available responses. *News Reading Habits = On average, how many days a week do you read online news articles?	38
5.1	Correlation table depicting how correlation changes depending on demographics. The arrows indicate whether the correlation is higher or lower, compared the reference group (All). Note: * $p < .05$, ** $p < .01$, *** $p < .001$	46
5.2	Correlation table depicting the difference in correlation between Sports and Recent Events. The p value here refers to whether the difference in correlation across the domains is significant. Note: * $p < .05$, ** $p < .01$, *** $p < .001$	48
5.3	MLR table depicting the influence of matching-characteristics on similarity judgments. "LoConf" are similarity judgments with a confidence level below 4, while "HiConf" are similarity judgments with a confidence level of 5. Note: * $p < .05$, ** $p < .01$, *** $p < .001$	51
5.4	MLR table depicting the influence of matching-characteristics on function similarity scores. BodyText:TF-IDF. Note: * $p < .05$, ** $p < .01$, *** $p < .001$	52
5.5	MLR table for human judgment and all functions, using the matching characteristics as independent variables. Note: * $p < .05$, ** $p < .01$, *** $p < .001$	53
5.6	Spearman correlation between functions and all human similarity judgments. Note: * $p < .05$, ** $p < .01$, *** $p < .001$	54
5.7	Spearman correlation between functions and human similarity judgment with a confidence level of 5. Note: * $p < .05$, ** $p < .01$, *** $p < .001$	55

5.8 Average human judgment similarity score for all available combinations of matching-characteristics in the study. "All" denotes all participants in the study. "HiConf" are similarity judgments with a confidence level of 5, and "LoConf" are similarity judgments with a confidence level below 4. 45-55 and 18-24 refers to participants in those age groups. 57

5.9 MLR table depicting the influence of matching-characteristics on similarity judgment for participants aged 18-24 and 45-55. Note: * $p < .05$, ** $p < .01$, ** * $p < .001$ 58

5.10 MLR table depicting the influence of matching-characteristics on similarity judgments. "LoConf" are similarity judgments with a confidence level below 4. "HiConf" are similarity judgments with a confidence level of 5. Note: * $p < .05$, ** $p < .01$, ** * $p < .001$ 59

Chapter 1

Introduction

1.1 Motivation

The news industry changed forever when most newspapers moved online around the turn of the 21st century. From a news reader's perspective, one of the biggest changes brought on by the digital era is the sheer volume of easily accessible news. The news domain has dealt with this in the same way almost every other domain which can boast an impressive catalog of items has, through recommender systems [22]. These systems are there to help consumers discover relevant content, making the decision making process both easier and more enjoyable for the users, which in turn can lead to increased user activity. To be as effective as they have the potential to be, recommender systems must be tailored to the specific challenges presented by the domain in which they operate, and one of the main challenges in the news domain is the perpetual new user problem [25]. This is a situation in which users browse anonymously rather than identifying themselves by logging in to user profiles, preventing the system from learning about their unique preferences over time. One solution, which is now used by almost every online newspaper, is similar item recommendation [44]. As the name implies, the goal is to recommend items, in this case articles, which are similar to the one the user just finished reading.

Similarity functions have been used across recommender domains [44]. However, functions that may be appropriate for one domain (e.g., movies) might be less representative for another (e.g., recipes). For the news similarity functions, it should be determined whether these functions are actually measuring similarity that resonates with users of news websites. To this end, human judgment can be used as ground truth, and then determine if the similarity scores of the functions correlate with human judgment. Using human judgment as ground truth in recommender systems is not a novel concept, and it has been done in both the movie [46, 44] and recipe [44] domains; nevertheless, with one noteworthy exception

[41], it remains largely unexplored in the news domain.

1.2 Problem

This thesis addresses the question that remains pertinent in the news domain: How should similar articles be recommended? While similar article recommendation is common, the concept of similarity itself in the context of the news domain is not very well understood, in the sense that algorithmic similarity functions might be disconnected from a more human-based understanding of similarity. To rectify this, this thesis employs a framework in which similarity functions previously utilized in the news domain are compared to human similarity assessments. In an attempt to better understand the underlying mechanisms behind the correlation between humans and functions, the study will control for the influence of factors commonly affecting news articles. Based on this, the thesis problem statement is as follows:

Which feature-based similarity functions are most representative of human similarity judgments, and does this depend on whether the articles in question share various characteristics?

1.3 Research Questions

The focal point of this thesis is to determine which metrics are most closely related to the human notion of similarity, and which elements has an influence on this similarity. To that end, the following research questions are raised:

- **RQ1:** *To what extent are various feature-specific similarity functions correlated with human similarity judgement?*
 - **RQ1.1:** *To what degree is the correlation between human judgment and various feature-specific similarity dependent on the category of the articles?*
 - **RQ1.2:** *To what degree is the correlation between human judgment and various feature-specific similarity dependent on whether the articles have any shared characteristics?*
 - **RQ1.3:** *To what extent do various user characteristics and demographic factors affect the perception of similarity between articles?*
- **RQ2:** *Which article features do users employ when determining similarity between articles?*

1.4 Contribution

- Insight into which feature-based similarity functions are most representative of human similarity judgment, as well as how different characteristics influence human similarity judgment and its relationship to similarity function scores.
- Knowledge of how and why similarity judgements made with a high degree of confidence correlate significantly more strongly with similarity function scores than other similarity judgments.
- A dataset consisting of articles, obtained from The Guardian and manually tagged with relevant named entities and sorted into groups based on their various characteristics.
- A dataset of human similarity judgment of article pairs, which is the result of a comprehensive online study.

1.5 Thesis Outline

- **Background.** Chapter 2 offers a summary of the literature on four key points: Recommender systems in general, unique challenges in a news recommender scenario, commonly used similarity functions for news retrieval, and human perception of similarity.
- **Pre-study.** Chapter 3 chapter outlines a preliminary study conducted to determine which elements contribute to article similarity.
- **Methodology.** Chapter 4 describes the process of constructing the dataset of articles, the similarity functions utilized, as well as the research design.
- **Results.** Chapter 5 provides the findings of the statistical analysis conducted to answer the research questions.
- **Discussion and Future Work** Chapter 6 discuss the findings from Chapter 5, along with limitations of this thesis and suggestions for further studies.

Chapter 2

Background

The background chapter is organized into six sections and presents a summary of prior work relevant to this thesis.

- Section [2.1](#) clarifies some potentially confusing news-related terms.
- Section [2.2](#) discusses research on recommender systems in general.
- Section [2.3](#) examines the most prevalent difficulties encountered while generating news recommendations.
- Section [2.4](#) explores the most common algorithmic approaches for generating news recommendations.
- Section [2.5](#) examines prior research to see if similar-item recommendation generates recommendations that users find similar.
- Section [2.6](#) concludes the chapter and describes the differences between current research and previous studies.

2.1 News, Sport, and Recent Events

Before we can proceed, it is necessary to clarify a couple of things in order to avoid confusion. The "News Domain" is a commonly used term [33, 18] that refers to the whole news industry, whether it is online versions of newspapers and news aggregators like Google News. All articles published by a newspaper are commonly seen as part of the news domain, regardless of whether the topic is politics, sports, business, or anything else.

It is common practice in the news industry to categorize articles. Figure [2.1](#) depicts the categories used on the BBC homepage.

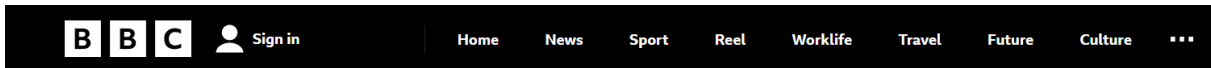


Figure 2.1: Illustration of the categories used on the BBC homepage.

The categories employed vary from newspaper to newspaper, although "News" and "Sport" are almost always present and by far the largest. The potentially confusing aspect here is the category called "News". This category is usually quite wide, but rarely does it encompass all articles in the newspaper. Typically, at the very least, sports-related articles are excluded from this category.

To prevent any misunderstanding over whether I am referring to the news domain as a whole or to the category named "News", I will refer to the category "News" as "Recent Events" from this point onwards.

2.2 Recommender Systems

News articles presented to users on the web can be personalized through recommender systems. A recommender system is a tool intended to assist the user in the decision-making process by providing recommendations, typically in the form of a top-N list [17, 19, 34]. It usually accomplishes this by analyzing past data to anticipate what the user would want in the present. The kind of data utilized depends on two factors: the types of data that are available and the types of data that are optimal for the recommender scenario at hand.

Collaborative Filtering (CF) is the most extensively used approach to designing recommender systems [34, 1]. The underlying premise of CF is that if individuals X and Y rate N items similarly or exhibit similar behavioral tendencies (I.e., watching, reading, listening), then they will rate or act similarly on other items [42]. Part of the allure of this approach is that in order to utilize CF, it is not necessary to have an in-depth grasp of the domain in which it will be used, as it is independent of the features of the items. A database containing user profiles and their interactions with the items in question is sufficient.

Content-Based Filtering is the second most common approach to personalized recommendations (CB). Instead of comparing user X to user Y, the focus here is on the items user X has engaged with in the past. Information regarding items (e.g., genre, author, director) is utilized to determine similarity between items, and then the user is recommended items that are similar to those they have previously interacted with [38].

CB and CF are frequently paired to produce hybrid approaches since they are capable of compensating for each others' limitations. For example, CF has a cold start problem for new items [43], which means that when you upload a new item, no one has rated it, and hence

the recommender system can not recommend it until enough people discover it on their own and rate it. When paired with a CB approach, you may immediately recommend new items, since the algorithm will be able to promote them to people who have previously liked similar items.

Whether CB, CF, or a hybrid approach is the optimal solution depends on the particular recommender scenario. However, in some cases the available data or website infrastructure only allows for one of the approaches. If users browse anonymously, that is, if they are not signed into the site, you cannot establish interest-tracking profiles for them, which rules out the approaches mentioned above. However, there are alternatives, such as non-personalized recommender systems. As the name implies, this sort of recommendation system generates identical recommendations for all users [30]. An example of of this is a "top N list" of some type, which may consist of the most popular items at the moment [10]. Another prominent method is to utilize a sort of content-based recommendation that does not require a user profile because it only considers the item the user is currently viewing when looking for similar items to recommend [18].

2.3 Recommender Scenario: News

The news domain has a number of unique or uncommon challenges compared to most other recommender scenarios. One complication is that the interests of readers are not constant; they vary dependent on elements such as the location of the reader and the time of day [18]. Time also plays a key role in the news domain in other ways, since the article's freshness sometimes influence how relevant readers perceive it to be [18]. There are numerous examples of personalized news recommenders that take article freshness into account while making recommendations [8, 9, 2, 29]. This is relatively common and hardly unexpected, given that the definition of news is "information about something that occurred recently" [45]. However, there are situations in which freshness is not very relevant, such as when a reader is attempting to acquire an overview of how a story has unfolded or when they are searching for articles that are related to the one they are presently reading [18].

Emotion also plays an essential role in the news domain, since it has been discovered that while individuals may forget the specifics of an article, their emotional reaction to the story typically lingers with them for a much longer period of time [39]. In a different study, designed to figure out what drives people to click on articles, Reis et al. [35] found that people were much more likely press articles with headlines that are either overtly negative or positive than neutral.

Another obstacle for recommender systems in the news domain is the necessity to handle

a vast corpus of articles that is continually evolving. Daily news articles are produced on a broad variety of subjects, making it challenging for a recommender system to keep up with current events and deliver relevant recommendations to readers. Furthermore, the sheer number of articles might make it challenging for a system to effectively process and evaluate the data in order to generate accurate recommendations [18].

Generating news recommendations involves more than simply predicting an article's relevance, even though that this is the most crucial component [18]. Recommending only Formula 1 articles to someone who is very interested in Formula 1 is a relatively safe bet, but it has been proposed that such repetitive recommendations may result in poor user engagement down the road [18]. According to various studies, diversity is a crucial aspect that might result in a more favorable opinion of the recommendations [31, 47]. Novelty is another quality factor to consider. According to Herlocker et al. [16], novel articles are ones that the user has not previously seen yet are relevant to them. Serendipity is yet another factor, one which is closely related to novelty, but also incorporates the level of unexpectedness of the recommendation [26]. Castells et al. [6] suggest that diversity, novelty, and serendipity are all desirable characteristics when constructing recommender systems, and that they should be balanced with accuracy, since the former often comes at the cost of the latter.

One of the challenges of news recommendation is what Mizgajski and Morzy [25] refer to as "the perpetual new user problem," which is characterized by the inability of news sites to track individual preferences over time. This is because the majority of individuals who read online newspapers do so anonymously by not signing into the site [27, 25]. However, as explained in Section 2.2, it is possible to generate recommendations without user profiles. Figure 2.2 depicts two of the most used solutions: recommending the most popular articles and articles that are related to the article currently being read.


Another approach in the news domain, which can be used in conjunction with what we are seeing in Figure 2.2 is session-based recommendation. Instead of incorporating the whole user history, session-based techniques concentrate on user-item interactions that occur within a set time frame, [27] which in a number of studies is set to 30 minutes [27, 23]. The objective of session-based recommendation is to predict the next item based on the sequence of previously consumed items in the session [7]. Predicting what a reader would desire after a single interaction with an article is clearly difficult, but adaptability of session-based system might be advantageous when dealing with the readers' ever-changing interests. An illustration of this is how one session-based news recommender takes into account the contextual qualities of the article, such as popularity and recency, and the reader context (I.e. time, location, and device) in order to construct their recommendations [40].

This does not mean that all news recommendations are either session-based or non-personalized.

Related stories


● ● ● ● ●

⏪



Fans, fun, fireworks: Qatar World Cup 2022 enjoys rare moment of normality

● 23h ago




Gianni Infantino does his Football Jesus act during strange monologue on Qatar

● 1d ago



Fifa president Gianni Infantino defends Qatar World Cup in bizarre speech - video

● 1d ago



England and Wales plan to defy Fifa with OneLove armbands at World Cup

● 1d ago

Most viewed

In Football	
<p>1 Qatar 0-2 Ecuador: World Cup 2022 kicks off after opening ceremony - as it happened</p>	<p>6 USA have questions of their own as controversy flares at World Cup</p>
<p>2 BBC ignores World Cup opening ceremony in favour of Qatar criticism</p>	<p>7 England's Harry Kane may abandon 'OneLove' armband over booking fear</p>
<p>3 Goalkeeper sent off for confronting fan who allegedly urinated in his drink</p>	<p>8 Fifa World Cup revenue up by more than \$1bn after taking tournament to Qatar</p>
<p>4 World Cup opening ceremony: six things we learned in Qatar</p>	<p>9 Strike a pose: the best World Cup 2022 portraits - in pictures</p>
<p>5 World Cup 2022 opening ceremony - in pictures</p>	<p>10 Fans paid to attend World Cup by Qatar have daily allowance cancelled</p>

Figure 2.2: Illustration of how The Guardian recommends the most popular articles in addition to those that are related to the reference article.

However, personalized recommenders based on profiles that maps the preferences of the users recommenders typically do not operate on online versions of newspapers, but rather on news aggregators such as Google News and Bing News [21].

2.4 Similarity Functions for News Retrieval

The way news recommenders operate is by focusing largely on text-based attributes of articles, such as the article's body of text and title, while the author is typically overlooked. [18, 3]. In other domains, such as movies and recipes, it is common to utilize images [44], but this is uncommon in the news domain [18]. Starke et al. [41] compared the correlation between various feature-based similarity functions and human judgment and found that while the body of text had the highest correlation, the author-based and image-based functions generally had higher correlation than the title-based functions, indicating that these features should perhaps also be considered.

Most news recommender systems are usually based on one out of two approaches: 1) The use of topic models to generate latent topics from texts, where Latent Dirichlet Allocation (LDA) [24, 20, 11] is one of the most often employed techniques for this purpose. 2) The use of vector space models. Many of the traditional algorithms are based on Term Frequency-Inverse Document Frequency (TF-IDF), a vector space model commonly used in information retrieval [14]. As a result, it is frequently used as a benchmark against which to evaluate

other functions [4, 5, 37]. TF-IDF is used to extract vectors from articles that readers have previously interacted with, as well as from unseen articles in the system. More precisely, it operates as follows: $TF - IDF(t, d, D) = tf(t, d) * idf(t, D)$, where $tf(t, d)$ indicates the rate with which a term appears in a document, and $idf(t, D)$ represents the number of documents in which a term appears [32]. Cosine similarity may then be utilized to determine the degree of similarity between the vectors of previously viewed and unseen articles: $Sim = \frac{A * B}{\|A\| \|B\|}$

Because TF-IDF has been around since 1975 [36], the version you see above has been modified over the years to address its shortcomings. For example, because articles are structured in an inverted pyramid pattern, which implies that the most important information appears first, it has been argued that shortening the length of articles may lead to improved TF-IDF performance. Bogers et al. [4] discovered that as articles grew longer, the performance of TF-IDF declined, albeit marginally.

One of TF-IDF's flaws is that it is incapable of capturing the meaning of words. For example, the phrases "United" and "Nations" directly succeeding each other obviously refer to the international organization; nevertheless, TF-IDF does not recognize this. Instead, if a second document says "Manchester United player set to play Nations League", TF-IDF will infer similarities between the two texts because they both include "United" and "Nations," even though the words clearly refer to different things. The opposite of this is also represents a challenge, where two different words with the same meaning are counted as separate terms [5].

This is the background for several different variations of TF-IDF which has the explicit goal of incorporating semantic meaning into the computation of similarity, such as Synset Frequency-Inverse Document Frequency (SF-IDF) [5] and Concept Frequency-Inverse Document Frequency (CF-IDF) [14]. Both SF-IDF and CF-IDF were shown to be superior to TF-IDF in terms of news recommendation [5, 14].

2.5 Human Perception of Similarity

The click-through rate is a popular metric for analyzing the success of news recommender systems' recommendations [12]. In the context of similar news recommendation, however, it is not particularly helpful if our goal is to determine whether or not people agree that the recommended items are truly similar. People may click on a recommended article for reasons other than similarity, or they may opt not to click on a certain article despite the fact that they believe it is similar, which is why a different approach is needed.

In the movie domain, Yao and Harper [46] gathered human judgments of similarity between

movies to utilize as ground truth in order to determine if different algorithmic methods to similarity actually correlate with the human idea of similarity. Trattner and Jannach [44] also utilized human judgment as ground truth in a study including both the movie domain and the recipe domain. Their study went farther by additionally focusing on determining what makes two items similar. They identified all available features, such as "Title" and "Genre," and applied a variety of similarity functions to these. This allowed them to identify which particular features and functions correlate most strongly with human judgment.

Inspired by the work of Trattner and Jannach [44], Starke et al. [41] adapted the same set of functions to the news domain. Seven different features were identified: subcategory, title, image, author, date, body of text, and author biography. The dataset consisted of Washington Post articles from 2012 to 2017 that were confined to the "National Politics" news category. Human judgments of similarity were collected and then compared to similarity scores computed by similarity functions for the seven features. While Trattner and Jannach's research revealed correlation scores of more than 0.50 for many features in the movie domain, the correlation between the same functions and human judgment in the news domain was considerably weaker. With a correlation of 0.29, TF-IDF (Body of Text) was the winner, while an image-based function ranked a distant second with 0.17. It is not a given that the same set of functions that did well in one domain would also do well in another domain, but there are reasons to examine these results further. On a scale from 1 to 5, where 1 represents complete dissimilarity and 5 represents complete similarity, the vast majority of article pairings were judged as 1 or 2 by the humans, leading to a mean judgment score of 1.72.

Consider the following hypothetical situation: We possess a dataset containing article pairings. Every article pairing that humans have judged as 4 (very similar) is likewise rated as 4 by the function. Every pair that is scored 5 (nearly identical) by humans is likewise rated 5 by the function. So far, I have described the ideal function, which can predict human notions of similarity with 100 percent accuracy and has a perfect correlation of 1.0 with humans. Now, suppose we add a large number of article pairings that users have rated as 1 or 2, such that they account for 80% of all article pairs in the dataset. Let's assume that for every article pair that people score as 1, the function rates it as 2, and for every pair that humans rate as 2, the function rates it as 1. The underlying reason for examining the correlation between functions and human judgment is so that we might potentially utilize the functions to recommend items that we are fairly confident people would find similar. Given that we are attempting to recommend similar articles, the function's poor performance in predicting whether article pairings should be scored as a 1 or a 2 is irrelevant. However, the correlation score in the scenario above would now be 0.12, making it appear as though the function is incapable of accurately estimating human similarity.

The point is, the correlation score alone does not provide the full picture of what is going on. Hence, the correlation results of Starke et al. [41] should probably not be used to draw any conclusions, beyond that the functions do not seem like they are suited to predict human judgment of similarity for dissimilar articles. If we instead could evaluate the correlation across several data-sets, each with unique characteristics and differing degrees of article similarity, this would provide us more comprehensive knowledge of whether or not there is a link between human judgment and this collection of similarity functions. In order to do this though, we would first need to have an idea about what factors causes people to perceive similarity between two articles in the first place.

The profiles created to capture a user's reading interests in news recommender systems is a reasonable starting point if we want to look for factors contributing to similarity, and keyword profiles used to be the most prevalent user profile type [13]. These are comprised of a list of keywords that indicate topics of interest, with each term being assigned a numerical value that reflects its significance to the profile [13]. Later it was argued by Li et al. [20] that a user profile expressed as a weighted topic distribution does not adequately reflect the user's specific reading preference. He proposed that users' interest in named entities be added in the profiles to make them more accurate. Li et al. [20] found that those recommender systems that incorporate preferred named entities perform better than those that do not. A different factor, one which has nothing to do with user profiles, but nevertheless seem to have an impact on the similarity between two articles is the proximity in publication date between the articles in question. Starke et al. [41] employed a exponentially declining function of publication date in their study, and while the correlation with human judgments of similarity was rather modest, it was strong enough to be statistically significant. In conclusion, it is reasonable to believe that topic, named entities, and closeness in date of publication all have a role in whether or not two articles are perceived to be similar.

2.6 Summary, Differences, and Contributions

The literature review indicates that the news domain is difficult to navigate due to numerous challenges, the biggest one being the absence of user profiles that map the long-term preferences of the readers. Similar article recommendations continue to be one of the most popular ways to make recommendations in such situations, and there does not seem to be anything in the literature to suggest that this will change in the near future.

Despite the importance of similarity in news recommenders, the literature offers limited insight into whether algorithmic representations of similarity accurately reflect what humans consider similarity to mean. The only previous study to calculate correlation between hu-

man judgment and similarity functions in the news domain is Starke et al. [41]. However, it seems plausible that the research design might have had a role in this study's finding of a weak correlation between human judgment and similarity functions.

This thesis will elaborate on the findings of Starke et al. [41] by employing the same set of similarity functions and features, but also taking into consideration a number of factors that may have influenced the correlation between human judgment and similarity functions. These factors include those that we have identified as likely to influence how similar people find article pairings to be, such as having the same topic, a shared named entity, and publishing dates that are close together. The study will also be undertaken using articles from the categories "Recent Events" and "Sport" to determine if article category may potentially be a factor influencing the correlation as well.

In conclusion, this thesis will address the following unanswered questions in the existing literature:

- Which particular combinations of features and functions are most representative of human judgment?
- Is the correlation dependent on whether the articles come from the category "Recent Events" or "Sport"?
- Is the correlation dependent on whether or not the articles being compared share a topic, a named entity, or proximity in publication date? These factors will henceforth be referred to as matching-characteristics.
- How do the matching-characteristics impact human evaluation of similarity when considered separately and when combined?
- How do the matching-characteristics impact the score of similarity functions when considered separately and when combined?

Chapter 3

Pre-study

After examining the existing literature, there were still some unanswered questions; thus, a preliminary study was conducted before moving on the main study of the thesis. The two principal goals for the pre-study were as follows:

First, the reason for wanting to examine whether similarity functions might have different correlation with human judgment depending on the category of the article was based on the assumption that people's perception of similarity might vary based on whether an article belongs to the "Recent Events" or "Sport" category. However, no information was found in the literature which could support this assumption. Therefore, the first objective was to produce data which could either support or refute the assumption. Second, we wished to either confirm that the similarity factors found in the background section were the most essential ones, or, alternatively, identify new factors that had been overlooked in the background section.

3.1 Data Collection and Procedure

Participants were recruited through Amazon Mechanical Turk, which were then directed to the survey, which was developed and hosted at SurveyXact. The participants were initially questioned about their gender, age, and the frequency with which they read online news articles per week. Following this, participants were asked three open-ended questions on news recommender systems.

3.1.1 Research Design

There were two separate versions of the preliminary study that were identical in every way except for one. The second question in the survey asks respondents to describe a fictitious

or nonfictional article that they would consider similar to the reference article they have been given. Version one utilizes an article from "Recent Events," whereas version two uses an article from "Sport." The study was run in two batches, and a between-subjects research design was utilized, with the first half of the participants exposed to version one and the second half to version two.

3.1.2 Participants

To help ensure the quality of the data, only Amazon MTurk Master workers were recruited. This indicates that the participants have high approval rates across a wide variety of work in the past. 45 workers were recruited and compensated \$1 each to complete the survey, which took an average of 4 minutes and 49 seconds.

As per Figure 3.1, sixty percent of the participants are women and forty percent are men. Figure 3.2 reveals that the age distribution is skewed toward the younger end, as only six participants are aged 45 or older. Figure 3.3 shows that there while 13 people read online newspapers daily, as many as 8 participants read online articles just one day a week or less. This number is surprisingly high, especially given that there were so few older participants.

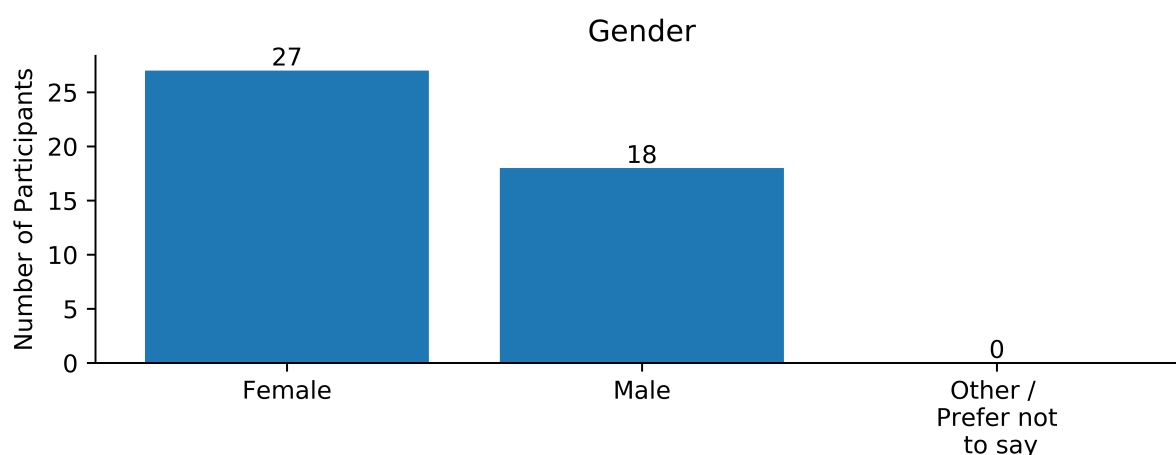


Figure 3.1: A bar graph displaying the gender distribution among the participants.

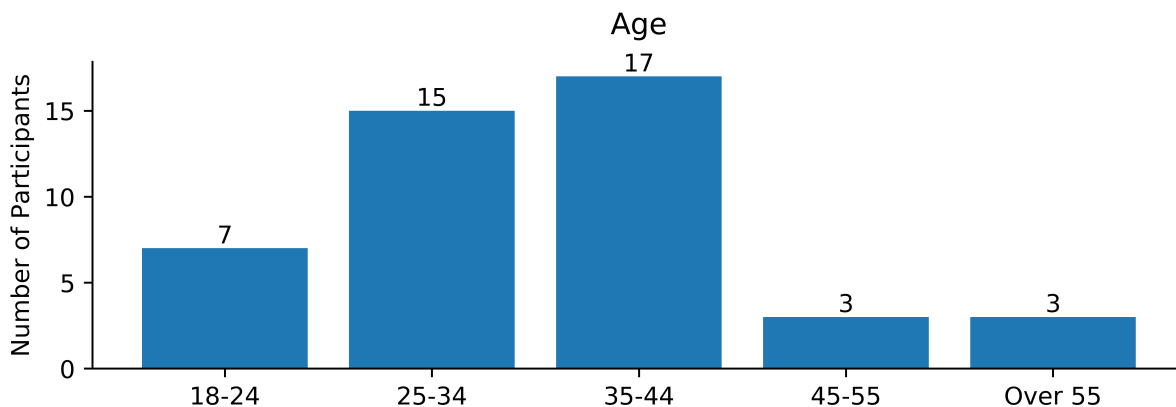


Figure 3.2: A bar graph displaying the age distribution among the participants.

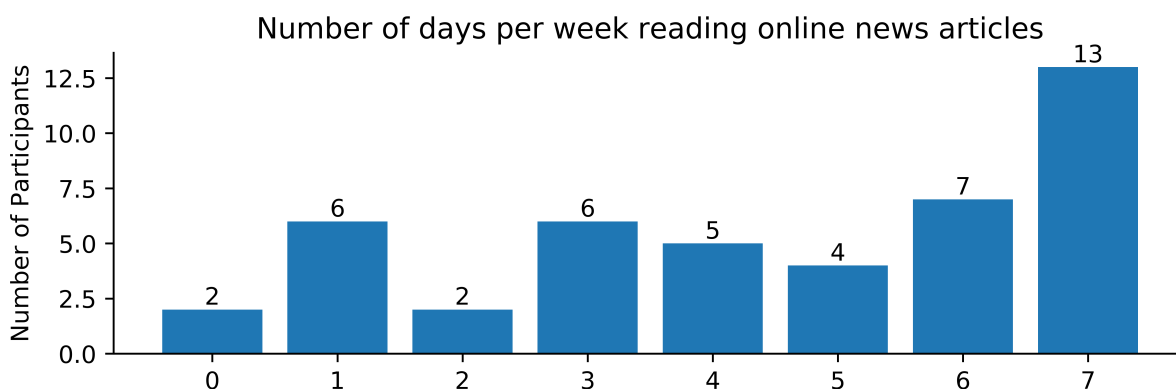


Figure 3.3: A bar graph displaying how often the participants read online news articles per week on average.

3.2 Criteria for news recommendation (Q1)

The purpose of the first question was to learn more about the kind of recommendations that people desire:

"News recommenders are encountered on news websites, where they suggest articles to you that you might be interested in reading next, after you have finished reading a news article. We want to know more about your thoughts on what information should be used for such news recommendations. Imagine that you have just finished reading an article, and you reach the list of potentially interesting articles for you to read next. What do you think should be the criteria for an article to appear on this list?"

3.2.1 Results

To transform qualitative responses into quantitative data, each response was labeled. The findings are depicted in Figure 3.4.

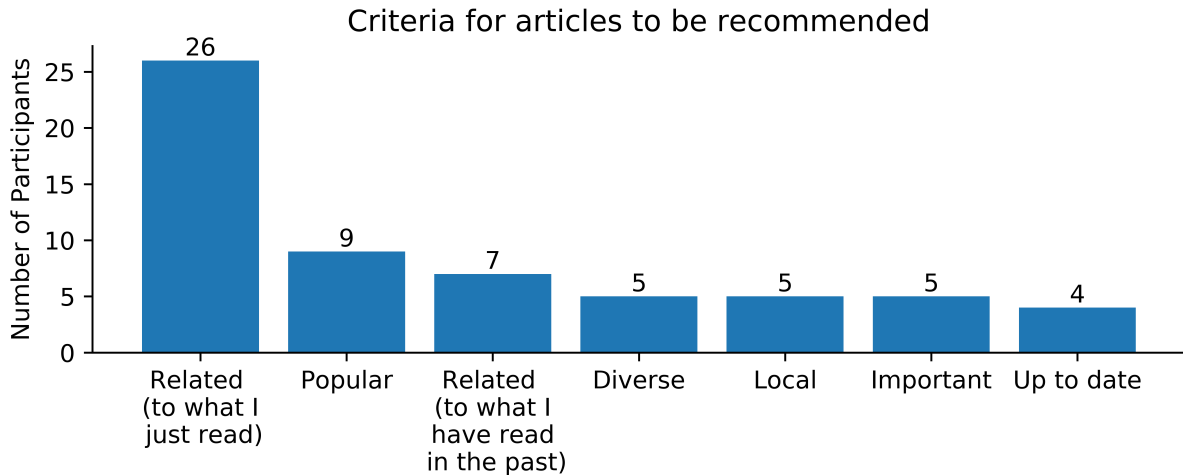


Figure 3.4: A bar graph displaying the criteria that individuals believe should be considered by news recommendation systems. Note: People were not confined to a single response; they could list as many factors as they wanted.

26 participants out of 45 stated that a criteria for a news article to appear on the list of recommended articles should be that it is related to the article they just read. This was by far the most frequently proposed criterion, followed by nine individuals who believed that news that is presently trending would be suitable criteria. Seven individuals said that relevance to previously read articles would be a fitting criterion, while diversity among the suggested articles (5), news that is geographically relevant (5), news that is of significant importance the public (5), and recency (4) were also listed as possible criteria.

3.2.2 Discussion

Evidently, the participants in the study believed similar article recommendation was the way to go. However, considering the prominence of this sort of recommendation in online newspapers, this is also the type of result one would expect if people simply responded the first thing that came to mind. Apart from that, it is noteworthy that several respondents brought up other factors highlighted in the background section, such as Diversity and Recency (Up to date).

The question was posed with the intention that it might uncover some aspect of news recommendation that I had overlooked and that could be incorporated into the main study. However, this did not occur, thus these findings had no impact on the main study.

3.3 Sport vs. Recent Events (Q2)

The writing styles of articles in the "Sport" category and the "Recent Events" category are unique. As its primary purpose is to inform readers, "Recent Events" articles are often written in a straightforward fashion. The purpose of "Sport" articles is to both inform and entertain. This is reflected in how articles on sports are written in a much more colorful and lively language. Based on this, I hypothesized that people may perceive similarity differently in the "Recent Event" and "Sport" categories. However, I was unable to locate any literature in the background section that could either support or refute this. The pre-study was a chance to determine whether this idea had any merit before committing to anything in the main-study.

The goal of the second question was to determine whether individuals had different definitions of similarity depending on whether an article belonged to the "Recent Events" or "Sport" category:

Covid: Boris Johnson sets new booster target over 'Omicron tidal wave'

© 13 December 2021



Watch Boris Johnson set out the latest plans to tackle Omicron

Sadio Mane: Senegal forward to return to Liverpool after X-rays on rib injury

© 13 November 2021 | Sport Africa



Sadio Mane has scored eight goals in 15 appearances in all competitions for Liverpool this season

Figure 3.5: Illustration of the two BBC news articles used for Question 2 in the pre-study

"Given the news article above, give us a short description of a either made up or real news article you would consider to be very similar."

Note: The first article was shown to half of the participants, while the second article was shown to the other half.

3.3.1 Results

The half who read the article regarding COVID and Boris Johnson all, without exception, described COVID-related articles. Boris Johnson was not mentioned by any of the participants. The group who read the report regarding Sadio Mané's injury responded differently. The majority of responses were either about Mané or someone you might argue is directly related to him, such as his teammates or manager. A few participants mentioned other footballers, whereas others talked about different sports, such as rugby.

3.3.2 Discussion

Everyone who was tasked with describing an article similar to the one on COVID stayed on-topic. While this is primarily a COVID-related article, it also discusses what Boris Johnson, the British prime minister at the time of the research, was doing to address the issue. And still, nobody paid any attention to Boris Johnson. In summary, the topic was of the utmost importance, but the people involved were less so.

This is in contrast to the manner in which the participants discussed their fictitious sports articles, where the focus was typically on Mané or someone closely related to him. There was also a tendency to stray considerably further from the original article, with some even describing non-football related, merely sport-related articles. In short, the person involved in the story was hugely important, while the topic was important too, but less so than in the case of the COVID article.

If we consider these articles to be representative of the "Sport" and "Recent Events" categories, it could be argued that people perceive similarity differently in the two categories. The implication this had for the main study was that if people perceive similarity differently based on the category of the article, then the similarity functions may have a different correlation to human judgment depending on the category of the articles that are being compared. This argument would not have been explored in the main study if it had not been supported in any way by the data presented here.

While there are differences in how people perceive what is similar to the two articles, this should not be simply attributed to the different categories. Clearly, the "Recent Events" article is story-driven, with COVID, not Boris Johnson, serving as the primary focus. In the "Sport" article, the story that someone is no longer injured is interesting *because* of the individual involved. However, it could also be that these traits are typical for the two categories. Either way, the support for the idea that correlation between human judgment and functions might differ depending on the category of the articles is clearly not very strong based on these findings, but strong enough that it is worth examining further in the main study.

3.4 Similarity Factors (Q3)

The objective of the third question was to determine which characteristics the readers consider most important when evaluating article similarity:

"When comparing two news articles, what is to you the single biggest factor that determines whether they are similar?"

3.4.1 Results

The coded versions of the responses are depicted in Figure 3.6. 29 of 39 participants with valid responses cited the topic as the single most influential aspect in the similarity of articles. According to five respondents, the title was the most critical feature, while two pointed to similar keywords. Two others cited shared named entities as the most significant indicator of similarity, while one individual identified journalistic quality as the most significant factor.

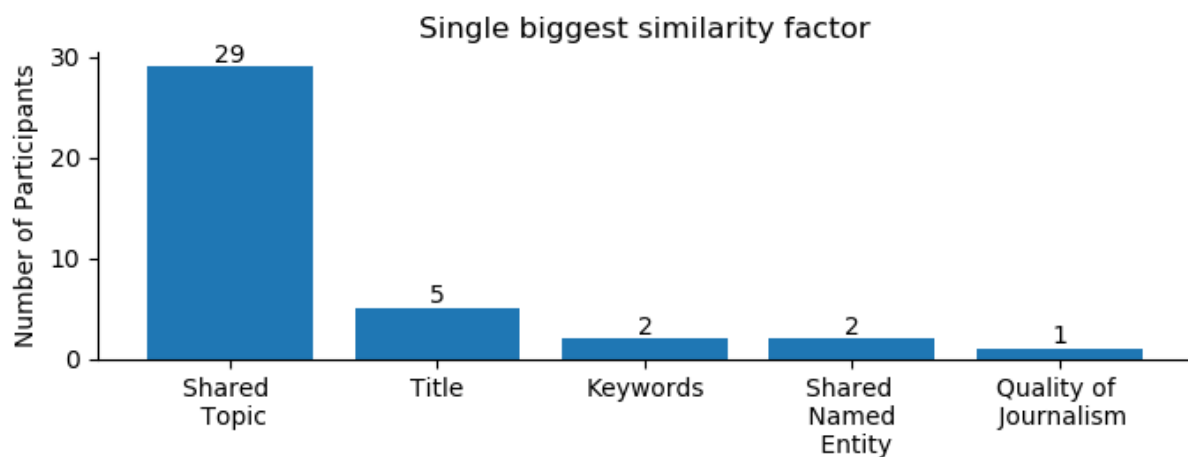


Figure 3.6: A bar graph displaying what people considered to be the single biggest factor that determines similarity between articles. Note: Six responses were omitted because it was impossible to discern with a high degree of certainty exactly what the respondents meant.

3.4.2 Discussion

As noted in the Background section, the main study will evaluate the influence certain factors have on human similarity judgment and the correlation between human judgment and similarity functions. The idea behind this question was it would either highlight some similarity factor that had been overlooked in the background section, or confirm that there was no obvious similarity factor that had been neglected. The latter turned out to be the case,

as there were no major surprises in the results. The topic was clearly the single most influential element in terms of similarity, while shared named entity was also cited.

Several people pointed towards having a similar title as the most significant criteria, which speaks to the perceived importance of this feature. Some of the functions used to calculate similarity in the main study are designed to utilize the title, thus this aspect is also covered.

Two individuals also listed similar keywords as the most essential criteria. This could have been a consideration, but there is a substantial overlap between having a shared topic and shared named entity on one side and having similar keywords on the other. If two articles share the same topic and named entity, it is almost certain that they share many of the same keywords. Keywords will therefore not be considered in the main study.

The quality of journalism is an intriguing answer that could have an impact on the perceived similarity between articles. However, there is no objective way to measure this, and it would likely play a far larger role in a different study in which the article pairings are not all selected from the same newspaper.

In summation, out of the three similarity factors chosen based on the literature, two of them - shared topic and shared named entity - were mentioned as the single most important factor by participants. The results also seem to indicate that there had been no major oversights in regards to any factors which should have been included.

3.5 Conclusion

The purpose of the preliminary study was to answer two questions that had to be resolved before the main study could be conducted: Are there any vital factors contributing to similarity which had been overlooked, and do people have a different perception of similarity entails based on whether an article belongs to the "Recent Events" or "Sport" category.

Participants identified two out of the three chosen factors, "Shared Topic" and "Shared Named Entity" as the single most important criteria. "Articles with close publication dates" was not addressed, but just because no one perceives it as the single most important element does not imply it is irrelevant. Also, based on the results, it was concluded that no major oversights have been done in terms of similarity factors which should have been included.

The study also supported the notion that there could be a difference in between what people think similarity entails, depending on whether the reference article belongs to Sport or Recent Events. Because of this, it was decided to not just use articles from Recent Events in the main study, but also articles from Sport, so that the two categories could be compared to each other.

Chapter 4

Methodology

This chapter is divided into three parts and details the methodology and data of this study. Section one discusses how the dataset utilized in this study was constructed. The second section offers a description of the similarity functions employed, while the third section describes the research design and the statistical methods for the study.

4.1 Constructing the Dataset

This study employs a dataset of 385 articles from 2019 to 2021, obtained from the British newspaper The Guardian ¹. The process of collecting these articles began with the identification of the requirements that each article in the collection had to meet.

4.1.1 Requirements

Familiarity

To assure the quality of responses when individuals are asked to rate the similarity between news articles, it was essential for them to have some level of familiarity with the articles. To this end, it is necessary to avoid niche articles that cover events and topics that people are unlikely to have heard of. For instance, articles about Formula 1, which is a massively popular sport, is fine, while articles about the sport of orienteering, are not.

¹<https://www.theguardian.com/>

Recency

In real life, when deciding which article to select next from a list of recommended articles, people typically choose from a selection of recent articles. Therefore, the overall scenario in which people are expected to read and evaluate the similarity between old articles is one that could feel a little unnatural. As a result, it was determined that the dataset should not contain articles that are too old, as it is likely that the older the articles, the more unnatural the task would feel.

2019 was chosen as the cutoff point going back in time, but we also need a cutoff point in the other direction. If we had selected articles published within the past several weeks or months, it is impossible to predict how the recency would have affected similarity assessments. To prevent this issue entirely, it was determined that too-recent items should be avoided, hence no articles from 2022 were selected. Thus, the chosen time frame for articles to be eligible for this dataset was between 2019 and 2021.

Covid-19

Given that we have collectively experienced a pandemic that has been all-consuming in terms of our lives and media articles, it was determined to omit Covid-19-related articles. Because it has become so entrenched in our lives, it is probable that any pair of news articles mentioning Covid-19, even if they are about entirely different things, will be perceived as quite similar due to the Covid-19 link. The results in the pre-study could indicate this as well, as not a single person strayed off the topic of Covid when asked to describe an article which would be similar to the reference article. Of course, this might not be the case too, but the downside of removing Covid-19 articles is virtually non-existent, while they potentially could have been problematic if they were included.

4.1.2 Extracting Named Entities

Looking back at the requirements, it is clear that they could all be achieved in some form through an automated process. However, the named entities from the articles also needed to be extracted, a procedure that is far more difficult to automate effectively. While it can be done, it is much harder due issues like how articles are sometimes inconsistent in their use of names in headlines. One article might write "Trump", another one "Donald Trump", and then there is a third one writing about Eric Trump, but just writing "Trump". This is not problematic for a person who is reading, as there will be other contextual clues as to who is the subject of the article, such as an image. Due to this, it was determined that the extraction

of named entities from articles would be performed manually.

The following were the rules for extracting named entities:

- Named entities are defined as a real-world object that can be identified by a proper name, such as a person, place, organization, or product.
- Named entities will be extracted from the headline and subheading of articles.
- Any sequential capitalized words will be part of the same entity (Monaco Grand Prix is one entity)
- When countries are referred to as adjectives (Russian man), the name of the country will be extracted (Russia).
- Always write down the full name of people, even if only the surname is used.

4.1.3 Obtaining Articles

Tagging articles with the named entities they mention requires manual inspection of the articles. It is easier to explore and inspect articles in an environment designed to display articles, such as an online news website, than in a dataset containing articles. Thus, articles were obtained from an online newspaper, rather than using articles from a pre-existing dataset. The chosen paper was the UK's most popular newspaper website, The Guardian [15].

The procedure began by selecting one of the two categories, "Sport" or "Recent Events," and then a subcategory of that, such as "UK Politics." A random date between 2019 and 2021 would be chosen, and a reference article that met all of the above listed requirements would be picked. Then, five to ten articles that were within two weeks of the reference article in terms of publication date would be picked as well. A new date would be picked and the procedure would be repeated up to five times more. Following that, a new subcategory would be chosen, and the entire process would be repeated. The amount of articles from each subcategory naturally fluctuated, since some featured a wide variety of diverse topics, resulting in more articles, whilst those with less diversity ended up with fewer articles, ensuring that no single topic was overrepresented. Once a sufficient number of articles had been obtained, the process was concluded. The research design section goes into further detail about what sufficient entails.

4.1.4 Data Structure

Instead of utilizing a pre-existing dataset, obtaining the data straight from a newspaper provided complete control over what type of data was collected and in what format it was stored. The data was stored as CSV as the simplicity made it the preferred option. The full list of features every article was stored with can be seen in Table 4.1.

Table 4.1: All articles were stored with the following information:

Article Features
ID
Category (Sport or Recent Events)
Topic
Title
Subheading
Main Image
Image Caption
Body of Text
Date
Time
Day of the Week
Author
Author Bio
Article URL
Named Entities

4.1.5 Categories

Sport and Recent Events (News) are the two categories from which items are obtained. These categories were chosen since they are by far the two most prominent categories in most newspapers, and as we observed in the preliminary study, people may have different ideas about what constitutes similarity depending on whether the reference article belongs to Recent Events or Sport. Before proceeding, we must establish what types of articles these two groups contain. Figure 4.1 depicts the subcategories which appear once you click on either Sport or Recent Events (News) on the website of The Guardian.



Figure 4.1: Illustration of the subcategories of Sport and Recent Events (News).

Note that the subcategories depicted in Figure 4.1 do not represent all subcategories for either Sport or Recent Events (News). It is just a list of the most popular subcategories, which is also why the majority of obtained articles originate from these subcategories, since we strive for a high degree of familiarity between participants and articles. Occasionally, though, articles from subcategories outside these were picked if they did not cover a narrow niche topic.

Sport is as self explanatory as a category can be. Any article pertaining to any sport is published in this category. However, Recent Events (News) is as vague as Sport is specific. First of all, we are just using the Recent Events (UK) category and not the full Recent Events category. In order to describe the category, we might begin by examining what it does not include. Sport articles are not present, nor any other articles where the objective is to entertain as much as it is to inform. Figure 4.1's subcategories offer a fairly accurate portrayal of the contents of this category. The names of the subcategories UK Politics, Media, Society, and Law are pretty indicative of the types of articles to expect. The vast majority, if not all, of the dataset's articles would fall into one of these subcategories. Scotland, Wales, and Northern Ireland were disregarded throughout the process. Again, we are seeking as much familiarity as possible between articles and participants, which is why only articles pertinent to the entire United Kingdom or England were obtained.

4.2 Similarity Functions

As stated in the background section, this thesis is based upon the work of Starke et al. [41]. Consequently, the same similarity functions they used are also utilized in this work. In this work, however, an additional feature, article subheadings, is taken into account. Thus, the complete list of features utilized in this work includes Title, Subheading, Date, Image, Body of Text, Author, and Author Bio.

LDA and TF-IDF, which are both commonly employed in the news domain, are also utilized here in conjunction with cosine similarity as similarity functions. The complete set of feature-function combinations is displayed in Table 4.2 below.

Name	Metric	Explanation
Subcat:Jacc	$\text{sim}(n_i, n_j) = 1 - \frac{\text{Subcat}(n_i) \cap \text{Subcat}(n_j)}{\text{Subcat}(n_i) \cup \text{Subcat}(n_j)}$	Subcategory Jaccard-based similarity
Title:LV	$\text{sim}(n_i, n_j) = 1 - \text{dist}_{LV}(n_i, n_j) $	Title Levenshtein distance-based similarity
Title:JW	$\text{sim}(n_i, n_j) = 1 - \text{dist}_{JW}(n_i, n_j) $	Title Jaro-Winkler distance-based similarity
Title:LCS	$\text{sim}(n_i, n_j) = 1 - \text{dist}_{LCS}(n_i, n_j) $	Title Longest common subsequence distance-based similarity
Title:BI	$\text{sim}(n_i, n_j) = 1 - \text{dist}_{BI}(n_i, n_j) $	Title Bi-gram distance-based similarity
Title:LDA	$\text{sim}(n_i, n_j) = \frac{\text{LDA}(\text{Title}(n_i)) * \text{LDA}(\text{Title}(n_j))}{\ \text{LDA}(\text{Title}(n_i))\ \ \text{LDA}(\text{Title}(n_j))\ }$	Title LDA cosine-based similarity
Subheading:BI	$\text{sim}(n_i, n_j) = 1 - \text{dist}_{BI}(n_i, n_j) $	Subheading Bi-gram distance-based similarity

Subheading:LCS	$\text{sim}(n_i, n_j) = 1 - \text{dist}_{LCS}(n_i, n_j) $	Subheading Longest common subsequence distance-based similarity
Subheading:TF-IDF	$\text{sim}(n_i, n_j) = \frac{TF-IDF(\text{Text}(n_i)) * TF-IDF(\text{Text}(n_j))}{\ TF-IDF(\text{Text}(n_i))\ \ TF-IDF(\text{Text}(n_j))\ }$	Subheading text cosine-based similarity
Image:EMB	$\text{sim}(n_i, n_j) = \frac{EMB(n_i) * EMB(n_j)}{\ EMB(n_i)\ \ EMB(n_j)\ }$	Image Embedding cosine-based similarity
Author:Jacc	$\text{sim}(n_i, n_j) = 1 - \frac{\text{Author}(n_i) \cap \text{Author}(n_j)}{\text{Author}(n_i) \cup \text{Author}(n_j)}$	Author Jaccard-based similarity
Date:ND	$\text{sim}(n_i, n_j) = 1 - \text{dist}_{Days}(n_i, n_j) $	Date published distance-based similarity (unit = days)
BodyText:TF-IDF	$\text{sim}(n_i, n_j) = \frac{TF-IDF(\text{Text}(n_i)) * TF-IDF(\text{Text}(n_j))}{\ TF-IDF(\text{Text}(n_i))\ \ TF-IDF(\text{Text}(n_j))\ }$	Body text cosine-based similarity
BodyText:LDA	$\text{sim}(n_i, n_j) = \frac{LDA(\text{Text}(n_i)) * LDA(\text{Text}(n_j))}{\ LDA(\text{Text}(n_i))\ \ LDA(\text{Text}(n_j))\ }$	Body text LDA cosine-based similarity
BodyText:Senti	$\text{sim}(n_i, n_j) = 1 - \text{SENTI}(n_i) - \text{SENTI}(n_j) $	Body text sentiment distance-based

		similarity
AuthorBio:TF-IDF	$\text{sim}(n_i, n_j) = \frac{TF-IDF(\text{Bio}(n_i)) * TF-IDF(\text{Bio}(n_j))}{\ TF-IDF(\text{Bio}(n_i))\ \ TF-IDF(\text{Bio}(n_j))\ }$	Author bio cosine-based similarity
AuthorBio:LDA	$\text{sim}(n_i, n_j) = \frac{LDA(\text{Bio}(n_i)) * LDA(\text{Bio}(n_j))}{\ LDA(\text{Bio}(n_i))\ \ LDA(\text{Bio}(n_j))\ }$	Author Bio LDA cosine-based similarity

Table 4.2: List of all the similarity function utilized in this work.

Title-based metrics

There are a total of five title-based metrics, four of which are string-based and one of which is topic-based.

The topic-based metric relies on LDA topic modeling for the article titles. In the interest of consistency, the same parameters as Starke et al. [41] are applied, hence the number of topics is set at 100. Cosine similarity is computed to compare two articles using weight vectors $LDA(n_i)$, and $LDA(n_j)$:

$$\text{sim}(n_i, n_j) = \cos(LDA(\text{Title}(n_i)), LDA(\text{Title}(n_j))) \quad (4.1)$$

The four string metrics used are Levensthein, Longest Common Subsequence, BI-gram, and Jaro-Winkler. The similarity is calculated by measuring the distance (dist) between two news articles, n_i and n_j :

$$\text{sim}(n_i, n_j) = 1 - |\text{dist}(n_i, n_j)| \quad (4.2)$$

Subheading-based metrics

Three subheading-based metrics were utilized, two of which were string-based: Bi-gram and Longest Common Subsequence. Similarity was calculated using the same method as discussed in the section on title-based metrics. The final metric is based on TF-IDF, and the similarity between two articles is calculated using cosine similarity.

Topic-based metrics

Only one topic-based metric was utilized: Jaccard Coefficient. See Table 4.2.

Author-based metrics

Again, only one metric was used: Jaccard Coefficient. See Table 4.2.

Date-based metrics

One date-based measure was employed, consisting of a linear function that estimates similarity based on the number of days between the publication dates of two articles:

$$\text{sim}(n_i, n_j) = 1 - |\text{dist}_{\text{days}}(n_i, n_j)| \quad (4.3)$$

BodyText-based metrics

Two BodyText-based metrics were used, TF-IDF and LDA. Similarity was calculated using the same methods as discussed in the section on title-based and subheading-based metrics.

AuthorBio-based metric

The same two metrics as for Body of Text were utilized for the Author Biography as well: TF-IDF and LDA. Similarity was calculated using the same methods as discussed in the section in title-based and subheading-based metrics.

4.3 Research Design

This section begins with an overview of the relevant factors pertinent to the research design, followed by a subsection describing each phase of the study's execution. It concludes with a summary of the statistical methods utilized.

4.3.1 Factors

How news articles were paired and whether they matched was subject to four factors: Date, Topic, Named Entity, and Category.

Date

The factor Date consists of two levels, similar and dissimilar, based on whether or not the articles' publication dates are near each other. The cutoff was set at 14 days, indicating that articles published within 14 days of one another are considered date-wise similar. Articles published between 14 and 28 days apart were deemed neither similar nor dissimilar and were eliminated as possible article pairing candidates for the study. Any pair of articles separated by more than 28 days was deemed dissimilar.

Topic

The two levels of the "Topic" factor are similar and dissimilar. Every article in the dataset is labeled with a topic, such as "Brexit" or "Formula 1." If the pair of articles possess the same topic, they are regarded similar; otherwise, they are deemed dissimilar.

Named Entity

The "Named Entity" factor comprises two levels: similar and dissimilar. If two articles share a named entity, they are considered similar, and if they do not, they are dissimilar. This factor is binary and makes no distinction between articles with a single shared named entity and articles with numerous shared named entities.

Category

The factor Category has two levels, "Sport" and "Recent Events," and each article pair in the study belongs to one of these two categories. This means that articles in the "Sport" and "Recent Events" categories are only compared to other articles in the same category.

4.3.2 Conditions

It was determined that the factor Named Entity was unsuitable as a standalone factor. Named entities are hardly ever referenced outside of a specific topic, especially in sports. Therefore, Named Entity only appears alongside Topic in this study. Thus, Topic and Named Entity are not separate factors, like Date and Category, but rather a single factor with three levels: Dissimilar, Topic, and Topic + Named Entity. This means that the online study was subject to a 2x2x3 within-subjects design. All 12 conditions are depicted in Table 4.3.

Table 4.3: The 12 conditions from the 2x2x3 factorial design. Note: Dissimilar refers to the situation where an article pair is not matched on either Date, Topic, or Named Entity

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

4.3.3 Dataset

A total of 385 news articles were obtained. The process stopped once it was possible to divide the set of articles into 60 groups of 12 pairs, with one unique pair from each of Table 4.3's 12 conditions in every group. Any given article could appear in a maximum of two different article pairs, but never in two article pairs from the same group. This way, if the 173 participants were spread evenly among the groups, all article pairings would be rated at least two times, but never more than three times.

Participants recruited at Prolific were directed to a web application built with HTML, CSS, and JavaScript from the ground up. When a participant opened the web application, a random number between 1 and 60 would be generated to determine which group of article pairs would be displayed. When an individual completed the study, the number generated at the beginning of the session would be recorded in a database. When a new participant accessed the study, the web application would compare the number generated to this database to ensure that the same group of pairs was never used more than three times². This provided an even distribution of participants among the study's various groups of article pairs.

²This means that no new people can get a particular group of article pairs as soon as three people have submitted their results with this group. However, if multiple people roll the same number in quick succession, more than just three participants can end up with the same group of articles. This happened for a few groups which ended up with more than three people rating the articles in it. Not often enough to be problematic, but the design flaw should still be pointed out.

4.3.4 Procedures and Measures

The web app's landing page is a consent form that provides participants with an overview of the study and the tasks they will be asked to perform. In order to proceed, participants must confirm that they have read and comprehended the instructions.

The user then proceeds to the second phase, depicted in Figure 4.2 which involves rating article pairs based on their degree of similarity. The users must evaluate 13 article pairings, one for each of the twelve conditions listed in Table 4.3, plus one pair that serves as an attention check (See Figure 4.3. For all article pairings, participants are asked to use a one-to-five Likert scale to indicate how similar the articles are, how confident they are in their similarity evaluation, and how familiar they are with the articles. The presentation order of the article pairings is randomized. At some point, users will encounter the article pair illustrated in Figure 4.3, which serves as an attention check. The opening sentence of the body text has been substituted with text asking viewers to rate everything on the page a 5, which is unlikely to occur naturally given the dissimilarity of the articles.

[Question 2 / 13]

Inspect both articles below. Click on 'show more' to read both articles in full. Afterwards, please respond to the four statements at the bottom of the page.

Article 1

Category: Athletics

Setback for UK Athletics as BBC balks at new £3m TV rights deal

- *Current deal expires this summer and new one not yet agreed BBC believed to only be offering a fraction of previous price*



BBC pundits Michael Johnson and Jessica Ennis-Hill broadcast from the Diamond League in London in 2018, one of the elite events. Photograph: David Klein/Reuters

Sean Ingle
12/02/2020

UK Athletics is facing a fresh crisis over the renewal of its £3m-a-year TV deal with the BBC which runs out this summer, the Guardian has learned.

Insiders fear the BBC is only willing to pay a fraction of what it currently pays for the rights for elite athletics in...

[Show more](#)

Article 2

Category: Cycling

'Do or die': Australian cycling in limbo amid landmark reform

- *Opposition to governance unification, along with the high voting thresholds, risks derailing nation-wide changes*



Under the new proposed structure, road cycling will be integrated with BMX and mountain biking under one administrative umbrella. Photograph: Charlie Crowhurst/Getty Images

Kieran Pender
21/10/2019

Cycling in Australia faces unprecedented upheaval, with voting underway to unify the 19 separate entities responsible for the sport across the country. While the proposed restructuring is intended to facilitate better outcomes for elite and recreational cyclists, create "one voice" for advocacy and improve the sport's financial position, resistance to...

[Show more](#)

The articles above are very similar

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

(Completely Disagree) (Completely Agree)

I am confident in my provided similarity rating

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

(Completely Disagree) (Completely Agree)

I am familiar with Article 1 shown above

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

(Completely Disagree) (Completely Agree)

I am familiar with Article 2 shown above

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

(Completely Disagree) (Completely Agree)

<< Previous

Next >>


Figure 4.2: Illustration of the second phase of the user study.

Article 1

Article 1: Team GB

Team GB swimmers come home after most successful Olympics

- *Brits return from Tokyo with haul of four golds, having finished third in swimming medal table*



Adam Peaty, left, and Tom Dean arrive back at Heathrow after their Olympic Games successes. Photograph: Aaron Chown/PA

Nicola Slawson
2021-08-02

This is an attention check: Please give a rating of 5 to all 4 scales on this page.

Team GB finished third in the swimming medal table behind the US and Australia. They won eight medals this past week. The squad returned with a haul of four golds, beating the...


[Show more](#)

Article 2

Article 2: Collingwood

Collingwood players write open letter apologising to anyone who suffered racism at the AFL club

- *The letter, signed by 150 footballers and netballers, states that 'through our silence we feel responsible for these injustices'*



The Collingwood players say they will 'confront the history of our club in order to learn, heal and determine how best to walk forward together'. Photograph: Darren England/AAP

Mike Hytner
2021-02-04

This is an attention check: Please give a rating of 5 to all 4 scales on this page.

Collingwood players have issued an apology for their part in allowing a culture of racism to develop at the club. In an open letter attributed to "the 150 footballers and netballers of...

[Show more](#)

Figure 4.3: Illustration of a pair of articles serving as an attention check. The first sentence of the body text has been replaced with text instructing users to give everything on the page a rating of 5.

Once all thirteen articles have been evaluated, users go to the third phase, depicted in Figure 4.4. The participants are shown a picture of an article which points out the individual features that an article in this study is made up of: Category, Title, Subheading, Image, Author, Date of Publication, and Body of Text. The user is then asked to indicate for each of these features, how important they were while making similarity judgments on a Likert scale from one to five.

Information Cues

Depicted below is an example of how a news article consists of different cues. Please indicate for each of these cues how important they were while making similarity judgements.

Category ---->

Title ---->

Subheading ---->

Image ---->

Author ---->

Date of Publication ---->


Body Text ---->

Article 1

Category: Champions League

Title: Manchester City v Chelsea: bravura final battle is held up by healthy bottom line

Subheading: *These furiously evolving teams deserve their place in the final, but wealth of their owners and Premier League certainly helps*



Author: **Barney Ronay**

Date of Publication: 2021-05-06

Body Text: *Les Anglais sont arrivés. Yes, it's that time of the year again. The darling buds of May are, with all due sense of caution, beginning to bloom. And the Premier League is all set for its annual European away day.*

First things first: the presence of two English clubs in...

Figure 4.4: Illustration of third phase of the user study.

Following this, the user advances to the fourth and final phase of the user research. The user is then asked to respond to a series of demographic questions, which are all depicted in Table 4.4.

Table 4.4: Demographic questions and available responses. *News Reading Habits = On average, how many days a week do you read online news articles?

Age	Gender	Education	News Reading Habits*
<18	Male	Less than high school	0
18-24	Female	High school or equivalent	1
25-34	Non-binary	Vocational School	2
35-44	Other / Prefer not to say	Bachelor Degree (e.g., BA, BSc)	3
45-55		Master Degree (eg., MA, MSc)	4
>55		Doctorate (e.g., PhD)	5
		Prefer not to say	6
			7

The study's web application's source code is available in a Github repository containing all

pertinent code for this thesis ³.

4.3.5 Participants

The crowd sourcing platform Prolific⁴ was used to recruit participants for the user study. The quality of responses was expected to be higher than those obtained from Amazon MTurk⁵, as research indicates that Prolific provides significantly greater data quality on parameters like as attentiveness, comprehension, and reliability [28].

To further assure the quality of the data, various constraints on who could participate in the study were imposed. Only workers with an approval rating of 99% from prior studies in which they had participated were employed, and all participants were required to be from England because the articles in the study were gathered specifically with that in mind. This was done to ensure that the participants had some level of familiarity with the articles. The median completion time was 14 minutes and 20 seconds, and the participants received 2.25 pounds for their work.

173 individuals were recruited in order to conduct the user study, resulting in 2,076 evaluated news article pairings. The median completion time of 14 minutes and 20 seconds was significantly longer than anticipated and suggests that the majority of users carefully considered their options before selecting an answer. The attention check was passed by 65 percent of the participants, which is a relatively high percentage given that it was far more concealed than attention checks typically are. However, after excluding individuals who failed the attention check, we are left with 1356 article pair ratings.

Overall, the dataset appears to be quite diverse, with no glaring imbalances that are big enough to present a problem in terms of the validity of the data. However, as depicted in Figure 4.6, there is bit of a gender imbalance in the dataset in the sense that women account for over 60% of the participants, men 38%, while 2% responded "Other / Prefer not to say". In terms of age, the dataset in Table 4.5 is fairly diverse, with a good spread. 25-34 is clearly the most populous age group with 58 people, while the smallest group is >55, with 19 people. With the exception of those without a high school diploma and those with a Ph.D., most groups are well-represented in terms of education.

The participants' self-reported assessments of their news consumption patterns in Figure 4.8 are quite different. The most frequent response is that a person reads online publications seven days per week, which accounts for around 35 percent of responses. However, nearly 30 percent of users report reading articles two or fewer days each week on average.

³<https://github.com/VRS-MT>

⁴Prolific platform: <https://www.prolific.co/>

⁵Amazon MTurk platform: <https://www.mturk.com/>

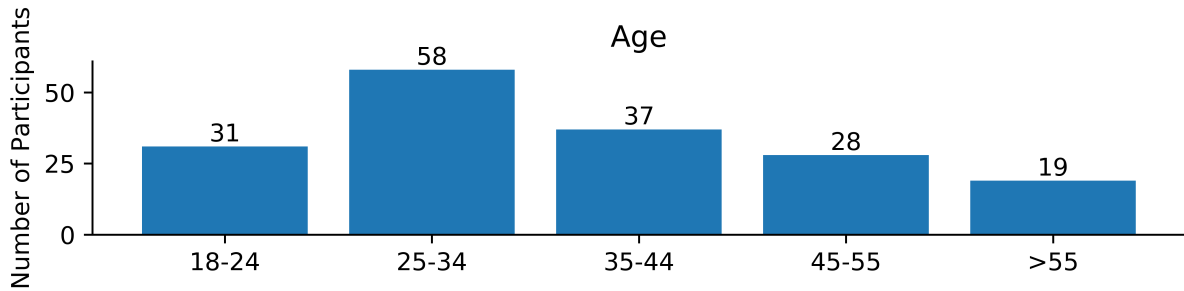


Figure 4.5: A bar graph displaying the age distribution among the participants.

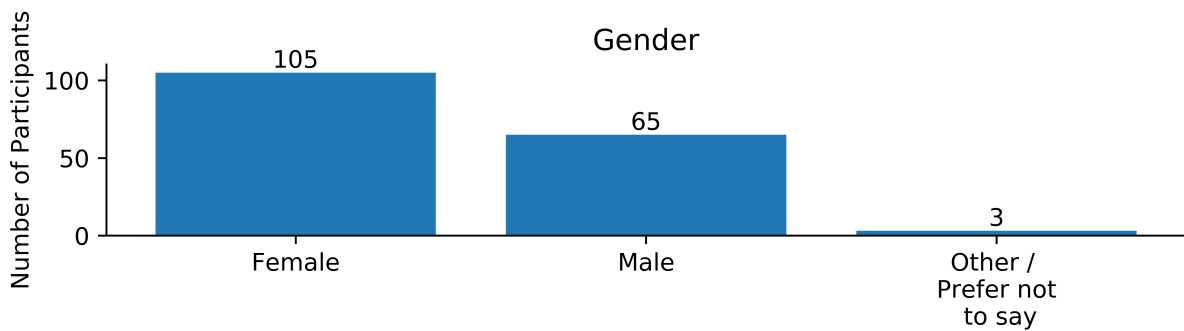


Figure 4.6: A bar graph displaying the gender distribution among the participants.

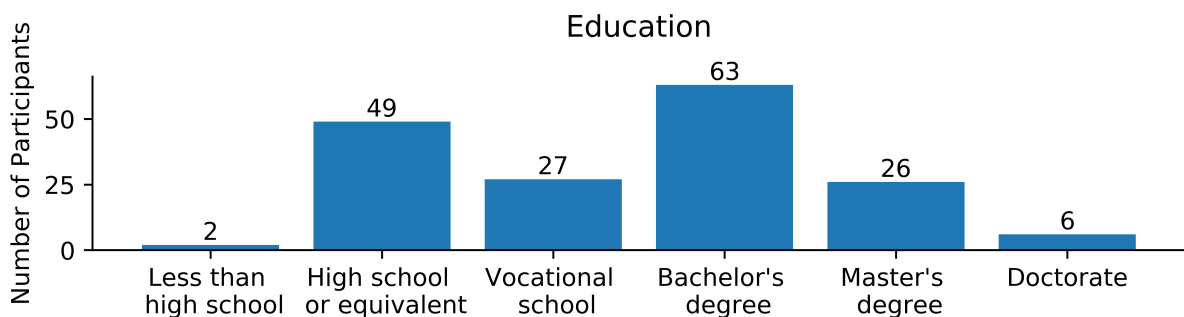


Figure 4.7: A bar graph displaying the gender distribution among the participants-

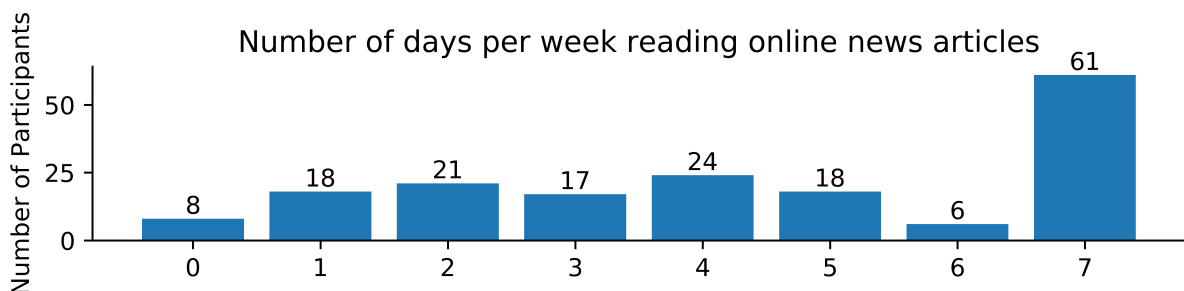


Figure 4.8: A bar graph displaying how often the participants read online news articles per week on average.

4.3.6 Statistical Analysis

Throughout the analysis, multiple statistical techniques were employed. The Spearman's Rank correlation coefficient was computed between human judgments of similarity and function scores across various data-sets. Fisher r-to-z transformation was utilized numerous times to determine whether or not these correlation coefficients were statistically significant different from one another. One-way ANOVA with Tukey Post Hoc Analysis was utilized to see if the variations in mean average human judgment across all 12 conditions were statistically significant. It was also utilized to assess the mean scores on which features were most important in determining article similarity. Finally, Multiple Linear Regression Analysis was employed to ascertain the extent to which the matching-characteristics (Date, Named Entity, Topic) could explain the variation in both human judgment and function scores.

Chapter 5

Results

This chapter describes the findings of the statistical analysis of the research data. It is organized in accordance with the research questions as follows:

- **RQ1:** Section 5.1 describes the analysis performed to evaluate to what extent various similarity functions correlates with human similarity judgment.
- **RQ1.1:** Section 5.2 describes the analysis performed to evaluate if the correlation is dependent on whether the articles originate from the Recent Events domain or the Sport domain.
- **RQ1.2:** Section 5.3 describes the analysis performed to evaluate if the correlation is dependent on whether the articles have any shared characteristics.
- **RQ1.3:** Section 5.4 describes the analysis performed to evaluate whether similarity judgment is dependent on user characteristics or demographic factors.
- **RQ2:** Section 5.5 describes the analysis performed to evaluate which article features readers employ when determining similarity between articles.

5.1 Similarity judgment and Function Correlation (RQ1)

To determine the degree to which similarity functions are representative of human judgment, the Spearman correlation between each function and human judgment was calculated. The Correlation Matrix in Figure 5.1 reveals that BodyText:TF-IDF ($\rho = 0.54$, $p < 0.001$) had the strongest correlation, which is not surprising given that TF-IDF has performed well in earlier research [41]. Clearly, the Body of Text feature is fit for purpose, yet BodyText:Senti ($\rho = 0.07$, $p < 0.01$) was one of the worst performing functions, and BodyText:LDA ($\rho = 0.17$, $p < 0.001$) was also a long way from the top. Topic:Jaccard had the second highest correlation ($\rho = 0.45$, $p < 0.001$), indicating a potentially significant role for shared topics in article recommendation. Title:BI ($\rho = 0.30$, $p < 0.001$) had the third strongest correlation overall, placing it well ahead of the other four title-based functions, which all had correlation coefficients below 0.20. Author was the only feature where the functions performed similarly, as indicated by the strong correlation between the functions themselves (0.71, 0.71 & 0.57). All correlations to human judgments were however modest with Author:Jaccard ($\rho = 0.26$, $p < 0.001$) being the strongest, followed by AuthorBio:TF-IDF ($\rho = 0.22$, $p < 0.001$) and AuthorBIO:LDA ($\rho = 0.21$, $p < 0.001$).

For the correlation matrix in Figure 5.1, the similarity ratings of all participants were utilized; however, Table 5.1 depicts how the correlation changes if we filter for individuals who passed the attention check or only use similarity ratings with a confidence level of 5. The correlational difference between all participants and those who passed the attention test proved to be insignificant. The correlation coefficient never changed by more than 0.03, and that occurred just once. There was also no consistency in terms of whether the correlation increased or decreased. Comparing all ratings to only those with a confidence level of 5 revealed significant differences in favor of those with high confidence. The three functions that had the strongest link with human judgment to begin with were also the ones whose correlation increased the most, each by at least 0.10 points. There is a clear pattern for functions that performed well in the first place to have a stronger correlation when all judgments that were not made with absolute confidence are filtered out, whereas functions that performed poorly to begin with do not see much of an increase, and some even decrease. All highlighted correlation increases in Figure 4.5 are statistically significant, which was established by applying Fisher r-to-z transformation to produce a z value, which was then used to evaluate the significance of the difference between the correlation coefficients.

Table 5.1: Correlation table depicting how correlation changes depending on demographics. The arrows indicate whether the correlation is higher or lower, compared the reference group (All). Note: * $p < .05$, ** $p < .01$, *** $p < .001$

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

5.2 Sports vs. Recent Events (RQ1.1)

To evaluate whether any of the functions may perform better in one domain than the other, Spearman correlation was once again calculated, but this time the dataset was divided into two sections: sport and recent events. Fisher r-to-z transformation was then utilized to see if the correlation coefficients differed in a statistically significant manner. Table 5.2 reveals that five functions had statistically significant better correlation in the recent events domain. The function with the greatest correlation difference among these five was Title:LV ($\rho = 0.00$ vs $\rho = 0.19$, $p < 0.001$). Both Subheading:LCS ($\rho = 0.10$ vs. $\rho = 0.18$, $p < 0.05$) and BodyText:LDA ($\rho = 0.12$ vs. $\rho = 0.21$, $p < 0.05$) fared much better in the recent events domain as well, although the correlation is still rather low for all three functions. However, BodyText:TF-IDF ($\rho = 0.48$ vs. $\rho = 0.56$, $p < 0.01$), and Topic:Jacc ($\rho = 0.41$ vs. $\rho = 0.49$, $p < 0.05$) - the two top functions from RQ1.1 - also performed better in the recent events domain. Sports, on the other hand, had two functions with stronger correlation: Image:EMB ($\rho = 0.15$ vs. $\rho = 0.07$, $p < 0.05$) and Subheading:TF-IDF ($\rho = 0.24$ vs. $\rho = 0.17$, $p < 0.05$), yet their correlation remains weak even in their strongest domain.

Table 5.2 also demonstrates how function correlation varies by domain when only similarity ratings with a confidence level of 5 was utilized. Because this group employs fewer ratings than the group using all ratings, the difference in correlation coefficients must be greater to be declared statistically significant. For instance, Topic:Jacc ($\rho = 0.51$ vs. $\rho = 0.59$) still favored recent events by a margin of 0.08, but the difference was no longer statistically significant. In general, the tendency was for the observed discrepancies between the domains to become more pronounced, as Subheading:TF-IDF ($\rho = 0.36$ vs. $\rho = 0.24$, $p < 0.05$) was 0.12 points ahead in sports and was now the third highest correlation function there. The most prominent exception to this pattern was BodyText:TF-IDF ($\rho = 0.62$ vs. $\rho = 0.67$), the function with the strongest correlation in both domains. There was still a minor gap between them, but it had shrunk and was no longer significant.

Table 5.2: Correlation table depicting the difference in correlation between Sports and Recent Events. The p value here refers to whether the difference in correlation across the domains is significant. Note: * $p < .05$, ** $p < .01$, *** $p < .001$

Function	Sport: All	ReEv: All		Sport: HiConf	ReEv: HiConf
Topic:Jacc	0.41	0.49 *		0.51	0.59
Title:LV	0.002	0.19 ***		0.00	0.28 ***
Title:JW	0.15	0.15		0.09	0.16
Title:LCS	0.15	0.21		0.13	0.25
Title:BI	0.30	0.31		0.36	0.42
Title:LDA	-0.04	-0.02		0.00	-0.08
Subheading:BI	0.14	0.11		-0.11	0.11
Subheading:LCS	0.10	0.18 *		0.08	0.23 *
Subheading:TF-IDF	0.24	0.17 *		0.36	0.24 *
Image:EMB	0.15	0.07 *		0.08	0.10
Date:ND	0.04	0.04		0.13	0.09
BodyText:TF-IDF	0.48	0.56 **		0.62	0.67
BodyText:LDA	0.12	0.21 *		0.21	0.28
BodyText:Senti	0.05	0.10		0.07	0.13
Author:Jacc	0.27	0.27		0.36	0.33
AuthorBio:TF-IDF	0.23	0.22		0.29	0.28
AuthorBio:LDA	0.22	0.20		0.31	0.24

Given the number of functions that perform better in either sports or recent events, and the fact that a number of these functions are among those with the highest correlation, it would appear that at least some of these functions are more suited to one domain than the other, though generally not by a large margin.

5.3 Matching-characteristics (RQ1.2)

To examine the influence of matching-characteristics, a one-way ANOVA and a Tukey's HSD post hoc test was conducted on human judgment across the various matching-characteristics. Figure 5.2 illustrates that adding "Date" to either "Dissimilar" or "Topic" did not yield significantly different results. This was true in both the sports and recent events domain. In other words, the proximity of two articles' publication dates did not appear to have much influence

how similar readers perceived them to be, at least by itself. The difference between "Dissimilar" and "Topic" was significant in both domains, but it was larger in the recent events domain, suggesting that being topically related may be a greater indicator of similarity in this domain. This was also in line with the result from RQ1.2 that Jacc:Topic has a greater correlation in the recent events domain. Adding "Named Entity" to "Topic" was not statistically significant in either domain, although it had absolutely no effect in the recent events domain while it was nearly statistically significant in the sport domain. While "Named Entity" and "Date" didn't contribute much on their own, there was a significant difference between "Topic" and "Topic + Date + Named Entity" for both domains. Furthermore, the difference was larger than what would be obtained by merely stacking the changes from "Topic" to "Date" and "Named Entity," indicating that there may have been a small interaction effect between "Date" and "Named Entity."

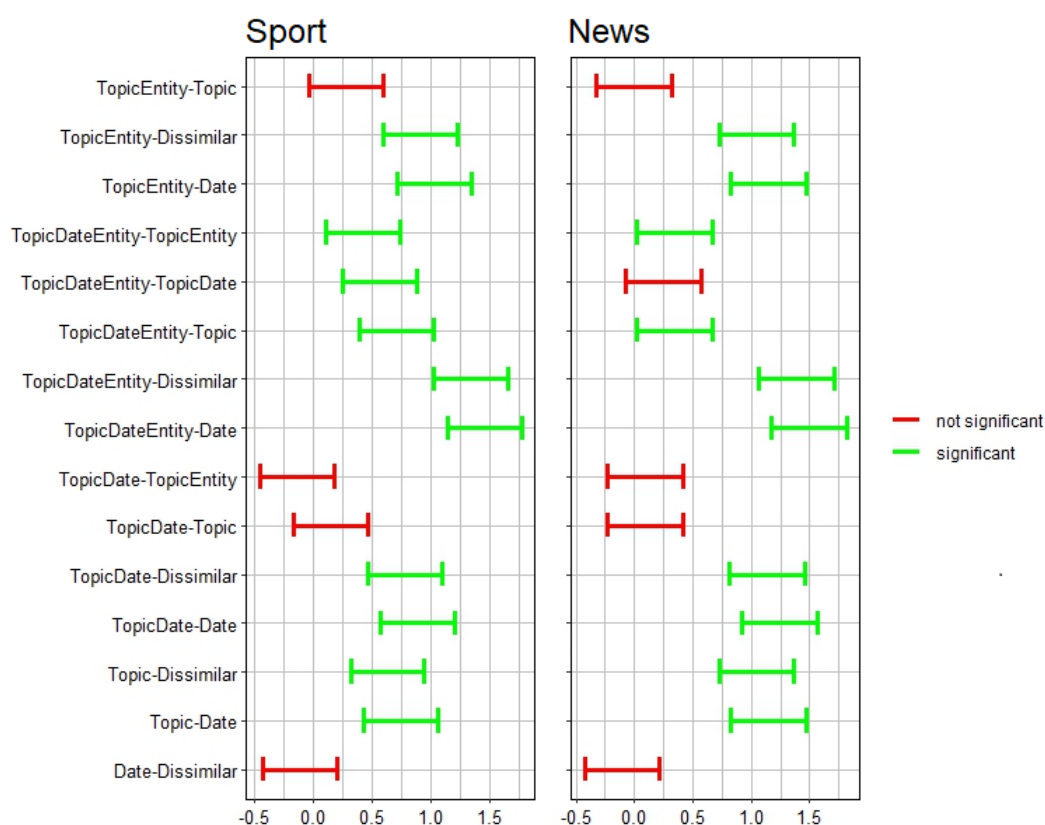


Figure 5.2: Tukey-HSD post hoc test result for human similarity judgment. News = Recent Events.

BodyText:TF-IDF similarity ratings were examined in the same manner. It appears, based on Figure 5.3, that the function was more sensitive than humans to matching-characteristics. The difference between "Dissimilar" and "Date" was insignificant, but everything else in the sport domain was significant. However, "Named Entity" was not even close to being a substantial factor in the recent events domain. The parallels with human judgment did not end

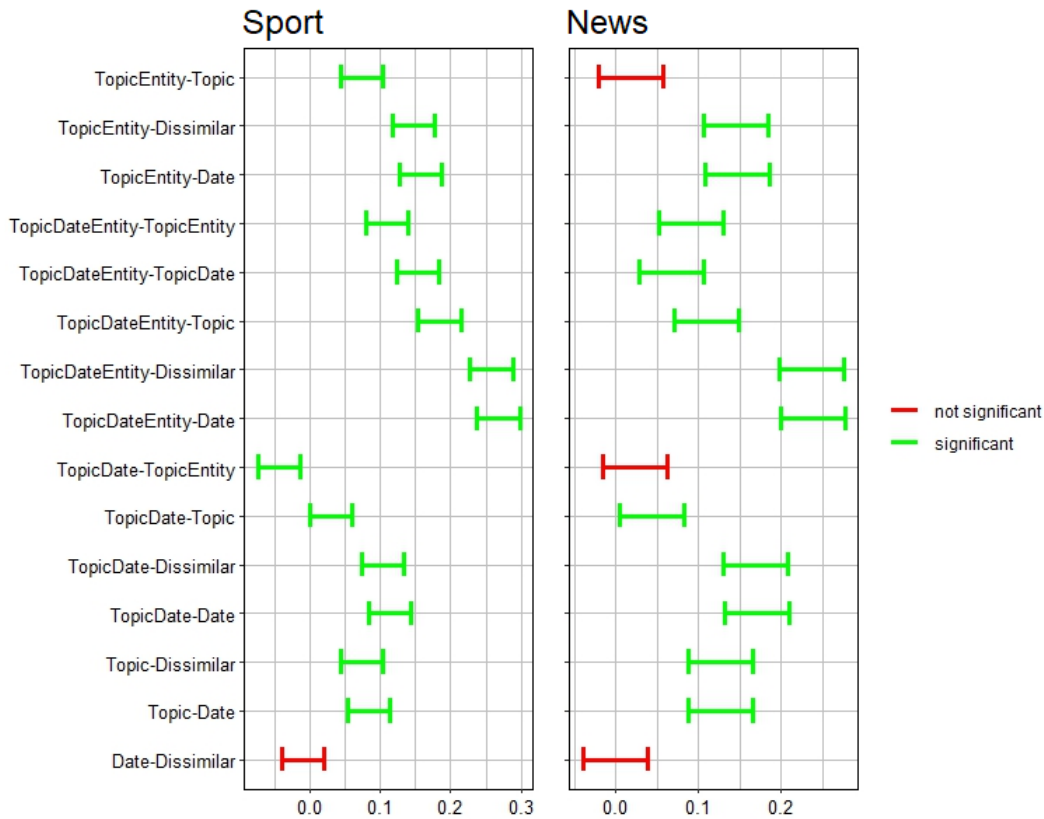


Figure 5.3: Tukey-HSD post hoc test result for BodyText:TF-IDF. News = Recent Events.

there, since "Topic" also appeared to be a more significant factor in the recent events domain than in the sports domain for TF-IDF as well.

5.3.1 Multiple Linear Regression

Using multiple linear regression analysis with the similarity score as the dependent variable and the matching-characteristics as the independent variables, the relationship between similarity scores and matching-characteristics was further explored. The R^2 value for human judgment (Human-All) in the recent events domain was 0.189, and 0.226 in the sport domain (Table 5.3). The standardized β coefficients in the recent events domain were as follows: Topic ($\beta = 0.451$, $p < 0.001$), Named Entity ($\beta = 0.049$), Date ($\beta = 0.046$). This supports the post hoc test findings: "Named Entity" and "Date" were not of great importance to people when they made similarity judgments in the recent events domain while topic on the other hand was a major factor. In sport, the findings are also in line with the post hoc results, in that "Topic" ($\beta = 0.313$, $p < 0.001$) was important, but less so than in recent events. Also, while "Topic" was nearly twice as important as "Named Entity" ($\beta = 0.176$, $p < 0.01$) when it comes to predicting human judgment in sports, this still means that "Named Entity" was a vital factor in this domain. The role of "Date" ($\beta = 0.068$, $p < 0.05$) was minor, but big enough

to just about be statistically significant.

Table 5.3 also includes findings for human judgment with a confidence level of 5 (HiConf). The R^2 score for HiConf in the sport domain was 0.304 and in the recent events domain it was 0.325, which is much higher than the R^2 value for Human-All. The significance of "Topic" ($\beta = 0.477$, $p < 0.001$) increased a little in the recent events domain, while "Named Entity" ($\beta = 0.144$, $p < 0.01$) tripled. "Date" ($\beta = 0.065$) remained insignificant. In sports, the relevance of "Topic" ($\beta = 0.349$, $p < 0.001$) increased slightly, "Named Entity" ($\beta = 0.265$, $p < 0.001$) grew significantly, and "Date" doubled ($\beta = 0.125$, $p < 0.01$). Overall, the matching characteristics appeared to have had a far greater effect on similarity judgments for HiConf than for Human-All.

Table 5.3: MLR table depicting the influence of matching-characteristics on similarity judgments. "LoConf" are similarity judgments with a confidence level below 4, while "HiConf" are similarity judgments with a confidence level of 5. Note: * $p < .05$, ** $p < .01$, *** $p < .001$

	Sport: Human (All)			ReEv: Human (All)		
	β	S.E.	STD. β	β	S.E.	STD. β
Topic	0.189 ***	0.020	0.313	0.286 ***	0.020	0.451
Named Entity	0.106 ***	0.020	0.176	0.031	0.020	0.049
Date	0.039 *	0.016	0.068	0.027	0.016	0.046
	$R^2 = 0.189$			$R^2 = 0.226$		
	Sport: Human (HiConf)			ReEv: Human (HiConf)		
	β	S.E.	STD. β	β	S.E.	STD. β
Topic	0.248 ***	0.038	0.349	0.327 ***	0.038	0.477
Named Entity	0.187 ***	0.038	0.265	0.107 **	0.041	0.144
Date	0.084 **	0.030	0.125	0.044	0.031	0.065
	$R^2 = 0.304$			$R^2 = 0.325$		
	Sport + ReEv: Human (All)			Sport + ReEv: Human (HiConf)		
	β	S.E.	STD. β	β	S.E.	STD. β
Topic	0.238 ***	0.014	0.384	0.286 ***	0.027	0.410
Named Entity	0.687 ***	0.014	0.110	0.151 ***	0.028	0.208
Date	0.033 **	0.011	0.056	0.065 **	0.022	0.100
	$R^2 = 0.205$			$R^2 = 0.316$		

As shown in Table 5.4, BodyText:TF-IDF had an R^2 value of 0.308 in the recent events domain and 0.429 in the sports domain. The relative importance of the matching characteristics in the recent events domain was also nearly identical to that of HiConf for "Topic" ($\beta = 0.456$, $p < 0.001$) and "Named Entity" ($\beta = 0.132$, $p < 0.001$), while "Date" ($\beta = 0.145$, $p < 0.001$) was

a quite a bit more important to BodyText:TF-IDF. Based on this, it is no surprise that the highest correlation observed in the study was between HiConf and BodyText:TF-IDF in the recent events domain. In the sport domain, the importance of "Topic" ($\beta = 0.330$, $p < 0.001$) and "Date" ($\beta = 0.163$, $p < 0.001$) was pretty similar to that of HiConf, but "Named Entity" ($\beta = 0.401$, $p < 0.001$) was far more important to the function than humans. This is trend which can be seen in other high correlation functions as well, such as Title:BI, where "Named Entity" ($\beta = 0.445$, $p < 0.001$) was more than two and half times as important as "Topic" ($\beta = 0.172$, $p < 0.001$) in predicting the function score. This may be part of the reason why the correlation is somewhat higher in the recent events domain for both of these functions, and it also explains why HiConf had a greater correlation than Human All, given that the former places a greater emphasis on "Named Entity."

Table 5.4: MLR table depicting the influence of matching-characteristics on function similarity scores. BodyText:TF-IDF. Note: * $p < .05$, ** $p < .01$, *** $p < .001$

	Sport: BodyText:TF-IDF			ReEv: BodyText:TF-IDF		
	β	S.E.	STD. β	β	S.E.	STD. β
Topic	0.094 ***	0.008	0.330	0.149 ***	0.010	0.456
Named Entity	0.114 ***	0.008	0.401	0.043 ***	0.010	0.132
Date	0.044 ***	0.006	0.163	0.045 ***	0.008	0.145
	$R^2 = 0.429$			$R^2 = 0.308$		
	Sport: Title:Bi			ReEv: Title:Bi		
	β	S.E.	STD. β	β	S.E.	STD. β
Topic	0.022 ***	0.003	0.172	0.039 ***	0.004	0.309
Named Entity	0.058 ***	0.004	0.445	0.020 ***	0.004	0.156
Date	0.009 **	0.003	0.072	0.007 *	0.003	0.057
	$R^2 = 0.310$			$R^2 = 0.171$		

Table 5.5 presents an overview of the R^2 value and the standardized β coefficient for matching-characteristics for humans and all similarity functions. While BodyText:TF-IDF and Title:BI appear to be impacted by matching-characteristics in a manner akin to human judgment, this was not the case for the majority of the other functions. The functions that were influenced by the characteristics in a manner most similar to those of humans also had the highest correlation. Body-Text:TF-IDF was more sensitive than humans to matching-characteristics, Title:BI was roughly as sensitive as humans, and all other functions were much less sensitive.

Table 5.5: MLR table for human judgment and all functions, using the matching characteristics as independent variables. Note: * $p < .05$, ** $p < .01$, *** $p < .001$

	Topic	Named Entity	Date	R ²
Human:All	0.383 ***	0.111 ***	0.056**	0.205
Human:HiConf	0.410 ***	0.208 ***	0.095 **	0.316
Title:LV	0.084 ***	0.212 ***	0.031	0.07
Title:JW	0.201 ***	0.088 ***	0.059 **	0.07
Title:LCS	0.184 ***	0.209 ***	0.055 **	0.12
Title:LDA	0.079 **	0.073**	-0.029	0.018
Title:BI	0.238 ***	0.303 ***	0.064 ***	0.226
Subheading:BI	0.102 ***	0.084 ***	-0.017	0.026
Subheading:LCS	0.143 ***	0.204 ***	0.039	0.092
Subheading:TF-IDF	0.163 ***	0.174 ***	0.029	0.085
Image:EMB	-0.081 **	0.201 ***	0.022	0.031
BodyText:TF-IDF	0.396 ***	0.256 ***	0.153 ***	0.348
BodyText:LDA	0.042	0.218 ***	0.093 ***	0.066
BodyText:Senti	0.020	0.053 *	-0.042	0.004
AuthorBio:TF-IDF	0.370 ***	-0.043	0.027	0.122
AuthorBio:LDA	0.348 ***	-0.026	0.041 *	0.114

5.3.2 Matching-characteristics and Correlation

Matching-characteristics appears to have a comparable effect on human judgment and BodyText:TF-IDF. Logically, this should indicate that TF-IDF is capable of distinguishing between dissimilar article pairs and pairs that have "Topic" and other matching-characteristics in common. What is less obvious, however, is whether BodyText:TF-IDF is capable of distinguishing similar article pairings. We observed in Table 5.8 that there is a vast difference in similarity between dissimilar article pairings and those that have a topic in common. In Table 5.1, we observed that the correlation of Topic:Jaccard was close to that of BodyText:TF-IDF. How can we know that BodyText:TF-IDF is not doing the same thing as Topic:Jaccard, that it is only effective at distinguishing between pairs of obviously similar and obviously dissimilar articles?

As we saw in the Tukey Post Hoc results in Figure 5.2, "Topic", "Topic" + Date" and "Topic" + "Named Entity" are not significantly different from each other, in either sports or recent events. These conditions were therefore merged to represent similar article pairs. Since "Date" and "Dissimilar" were also a nearly identical, they have been merged to create a bigger set of dissimilar articles. Spearman correlation was then calculated between human judgment and the similarity scores of the functions for both the set of only dissimilar article pairs, and the set of only similar article pairs. The results are displayed in Table 5.6 (All

ratings) 5.7 (High confidence ratings).

Table 5.6: Spearman correlation between functions and all human similarity judgments.

Note: * $p < .05$, ** $p < .01$, *** $p < .001$

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

Table 5.7: Spearman correlation between functions and human similarity judgment with a confidence level of 5. Note: * $p < .05$, ** $p < .01$, *** $p < .001$

Function	Sport: Dissimilar	Sport: Similar	News: Dissimilar	News: Similar
Title:LV	0.12	-0.03	-0.02	0.19
Title:JW	0.04	0.03	0.01	-0.01
Title:LCS	0.08	0.02	-0.07	0.22
Title:BI	0.15	0.22	-0.05	0.33 **
Title:LDA	-0.03	0.09	-0.19	0.06
Subheading:BI	0.04	0.08	-0.12	0.14
Subheading:LCS	0.12	-0.02	0.03	0.19
Subheading:TF-IDF	-0.02	0.34 **	0.00	0.11
Image:EMB	-0.07	0.10	0.13	0.16
Date:ND	-0.05	0.26 *	-0.04	0.13
BodyText:TF-IDF	0.12	0.51 ***	0.19	0.53 ***
BodyText:LDA	0.12	0.10	0.13	0.32 *
BodyText:Senti	0.01	0.10	0.02	0.07
Author:Jacc	-0.06	0.26 *	0.21	0.23
AuthorBio:TF-IDF	0.01	0.19	0.15	0.21
AuthorBio:LDA	0.26	0.16	0.16	0.21

Most functions exhibited greater correlation in the dataset of similar article pairings than in the dataset of dissimilar article pairs, although correlation remained low in both datasets when Human-All ratings were utilized. BodyText:TF-IDF had a correlation of 0.37 ($p < 0.001$) in recent events and 0.37 ($p < 0.001$) in sports, however none of the other functions had a correlation greater than 0.22. When comparing functions to HiConf human judgment, the general tendency of functions performing better with the dataset of similar article pairs persisted, but to a considerably greater extent. Compared to Human-All, the function correlation in the dissimilar dataset was substantially lower while the correlation in the similar dataset was much greater. The TF-IDF correlation in the recent events domain with the similar dataset was 0.51 ($p < 0.001$) and 0.53 ($p < 0.001$) in the sport domain. Subheading:TF-IDF ($\rho = 0.36$, $p < 0.001$), which was the second highest correlation in the sports domain for HiConf saw its correlation almost double from what it was compared to Human-All. The correlation for the same function in the same domain was -0.02 when the dissimilar dataset was being used instead.

These results demonstrate two points. HiConf has a significantly stronger correlation to sim-

ilarity functions than Human-All in a dataset consisting of pairs of articles that are largely similar. Also, matching-characteristics matter in terms of correlation, since the best functions consistently achieved much greater correlation in the dataset containing article pairs that were matched on topic than in the dataset containing article pairs that were dissimilar.

5.4 Influence of demographic factors and user characteristics on Similarity judgment (RQ1.3)

This section presents the findings of the analysis conducted to examine if demographic factors and user characteristics influence article pair similarity ratings. Similar article recommendation hinges on different people being in agreement as to to which articles are similar and not. That is why the similarity judgment of almost all available user characteristics and demographic groups was compared against each other. This includes the following factors: Gender, Age, Education, Use of Online Newspapers, Attention Check, Confidence

Table 5.8 provides an overview of the average similarity judgment score from the participants, for various demographic groups across the 12 conditions. The results of the group "All" in Table 5.8 indicate that Dissimilar and Date are the two conditions with the lowest human judgment scores. Date has a lower score than Dissimilar in both Sport and Recent Events, but the difference is so small that it is not statistically significant, as previously seen in Table 5.2. Also noteworthy is the fact that both Topic and Topic + Named Entity have the same similarity rating in Recent Events, supporting the former claim that named entities have minimal influence on human similarity in this category. For both categories, the condition with highest average human judgment, is the one with all three matching characteristics stacked on top of each other: Topic + Named Entity + Date.

The similarity ratings were compared across all 12 conditions made up of the different combinations of matching-characteristics, and one-way ANOVA was used to assess whether or not the differences in similarity rating between demographic groups were statistically significant ($p < 0.05$).

Table 5.8: Average human judgment similarity score for all available combinations of matching-characteristics in the study. "All" denotes all participants in the study. "HiConf" are similarity judgments with a confidence level of 5, and "LoConf" are similarity judgments with a confidence level below 4. 45-55 and 18-24 refers to participants in those age groups.

Condition	All	HiConf	LoConf	45-55	18-24
Sport: Dissimilar	1.94	1.50	2.28	1.89	2.26
Sport: Date	1.83	1.36	2.41	1.61	2.29
Sport: Topic	2.57	2.28	2.59	2.25	2.55
Sport: Topic + NE	2.86	2.72	3.06	2.61	3.10
Sport: Topic + Date	2.72	2.55	2.73	2.50	2.81
Sport: Topic + Date + NE	3.28	3.58	2.96	3.11	3.68
ReEv: Dissimilar	1.74	1.26	2.31	1.54	1.94
ReEv: Date	1.64	1.26	2.15	1.29	2.00
ReEv: Topic	2.79	2.57	2.98	2.75	3.19
ReEv: Topic + NE	2.79	2.69	2.67	2.68	2.87
ReEv: Topic + Date	2.88	2.57	2.73	2.46	3.26
ReEv: Topic + Date + NE	3.13	3.30	3.00	3.00	3.39

Age

There were six age categories: <18, 18-24, 25-34, 35-44, 45-55, and >55. <18 and >55 were omitted due to insufficient numbers. For the two youngest age categories, 18-24 and 25-34, there were no conditions with significantly differing scores. The same held true for the two oldest age categories, those aged 35-44 and 45-55. 18-24 and 25-34 each had two conditions that were different from 35-44, whereas 24-34 and 44-55 had just one condition with different scores. The two groups that differed the most from one another was the youngest (18-24) and the oldest (45-55), with five out of twelve conditions having statistically significant different scores. Table 5.8 displays the similarity ratings for these two groups. There is a clear trend in the differences between these categories, as participants in the older age category consistently rated article pairings as less similar than their younger counterparts across all 12 conditions. The two groups were further analyzed using multiple linear regression (MLR), with the similarity rating as the dependent variable and the matching-characteristics as the independent variables. Table 5.9 reveals that with the exception of "Date", the matching-characteristics were of equal importance to both groups, while the squared R values were nearly identical at 0.20 and 0.22. All of this seems to indicate that the youngest and oldest groups were largely in agreement regarding the types of articles that are similar to one

another, and that the observed difference between the groups was the result of two groups using slightly different parts of the similarity scale to express roughly the same sentiment. As these were the only two age groups with significant differences for more than two of the conditions, the overall influence of age on article similarity judgments appears to have been modest.

Table 5.9: MLR table depicting the influence of matching-characteristics on similarity judgment for participants aged 18-24 and 45-55. Note: * $p < .05$, ** $p < .01$, *** $p < .001$

	Age: 18-24			Age: 45-55		
	β	S.E.	STD. β	β	S.E.	STD. β
Topic	0.207 ***	0.031	0.357	0.228 ***	0.033	0.378
Named Entity	0.077 *	0.031	0.132	0.089 **	0.033	0.148
Date	0.063 *	0.0256	0.115	0.010	0.028	0.018
	$R^2 = 0.205$			$R^2 = 0.221$		

Confidence

When comparing similarity ratings with a confidence level of 3 or less to those with a confidence level of 5, four of the conditions had ratings which differed significantly. The similarity ratings, as depicted in Table 5.8, follow a clear trend in which the high confidence ratings appear to be far more sensitive to the matching-characteristics. Using multiple linear regression in the same manner as for the age demographic supported this notion, since the matching-characteristics explain 31.6% of the variation in similarity judgments for the high confidence ratings, but only 8.2% of the variance for the low confidence ratings (See Table 5.10). This results in far larger differences in the average similarity rating between the various conditions for the high confidence ratings than is the case for low confidence ratings. Even when the high confidence ratings are compared to all similarity ratings as opposed to the low confidence ratings, the same tendency can be observed, albeit the disparity is not as pronounced. This suggests that the level of confidence in the similarity rating had a significant effect on similarity judgments.

Table 5.10: MLR table depicting the influence of matching-characteristics on similarity judgments. "LoConf" are similarity judgments with a confidence level below 4. "HiConf" are similarity judgments with a confidence level of 5. Note: * $p < .05$, ** $p < .01$, *** $p < .001$

	LoConf			HiConf		
	β	S.E.	STD. β	β	S.E.	STD. β
Topic	0.117 ***	0.023	0.241	0.286 ***	0.027	0.410
Named Entity	0.041	0.021	0.092	0.151 ***	0.028	0.208
Date	0.006	0.018	0.015	0.065 **	0.022	0.096
	$R^2 = 0.082$			$R^2 = 0.316$		

5.4.1 Other factors

None of the remaining four factors, Gender, Education, Use of Online Newspapers, and Attention Check, appeared to affect the similarity scores.

5.5 Information Cue Usage (RQ2)

Participants were asked to rank the importance of article features on a scale from 1 to 5 when determining similarity between articles. The Body of Text had the highest importance rating at 4.17, followed by Title with a rating of 4.04. The results of the one-way ANOVA and Tukey's HSD post hoc test depicted in Figure 5.4 show that the difference between these two is not statistically significant. A bit farther behind were Subheading (3.50) and Topic (3.37), whose differences were also not statistically significant. At the bottom we find Image (2.92), Date (2.42), and Author (1.76).

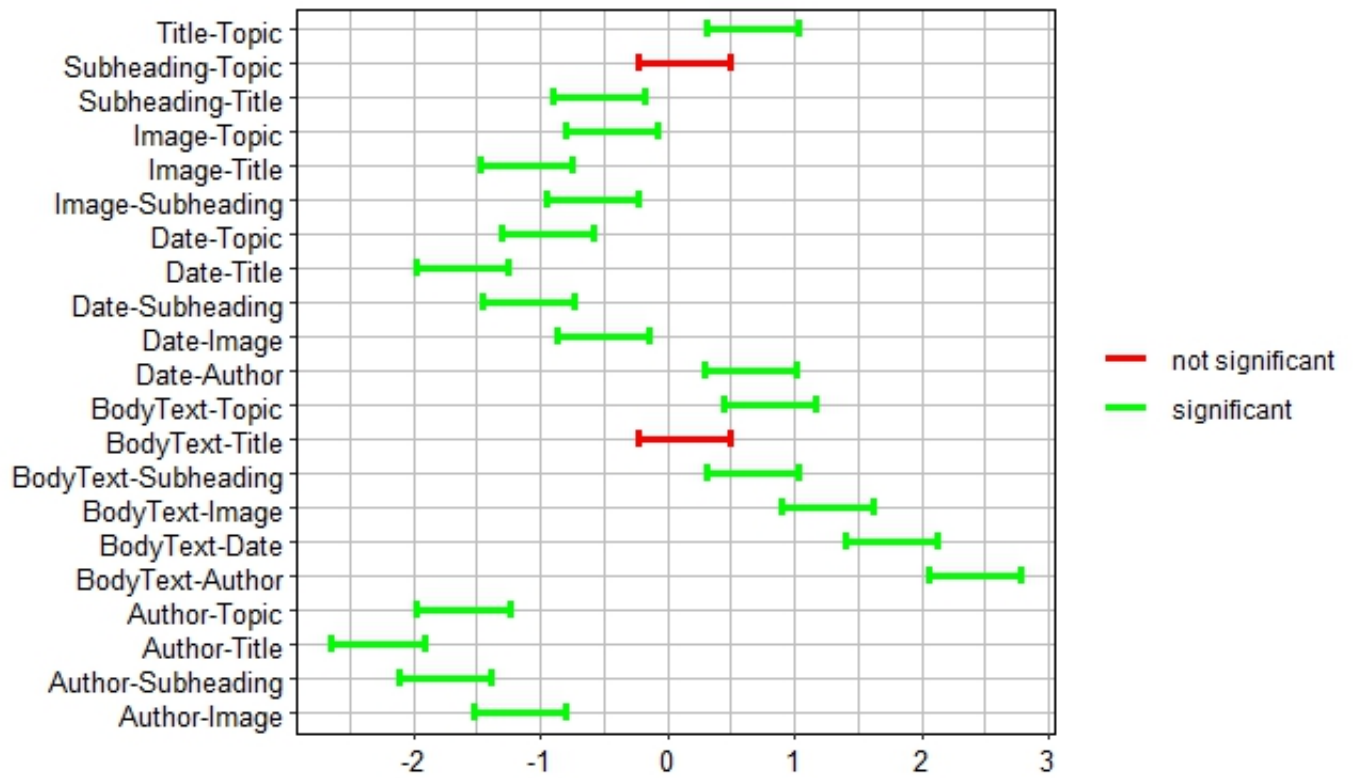


Figure 5.4: Tukey-HSD post hoc test result for information cue usage.

Chapter 6

Discussion and Future Work

In this thesis, it has been investigated how representative various feature-based similarity functions are of human similarity judgments. In doing so, we have expanded upon earlier work by Starke et al. [41] in which human similarity judgment are compared to similarity functions for news article that are paired randomly to each other. In this thesis, however, it has been determined how matching characteristics between news articles based on similarity in Date, Named Entities, Topic, and Category impact human judgment of similarity, similarity scores of various functions, as well as the correlation between human judgment and function scores. As Starke et al. [41] relied on a dataset with various news articles that were dissimilar, the rationale for this thesis work has been that presenting news articles that are matching is a more realistic news recommendation scenario. On top of that, it might have been easier to judge for humans, which might increase the low correlation scores found in Starke et al. [41].

To this end, a new dataset of articles has been generated in this thesis, that allowed the research to control for the effect of the aforementioned characteristics. The similarity of the article pairings in the dataset was then evaluated in a comprehensive user study and a long list of similarity functions. To answer the research questions, thorough statistical analysis was undertaken, and the following is a discussion of the analytical results and conclusions relevant to each research question.

6.1 Correlation between human judgment and various similarity functions (RQ1)

The correlation between human judgment and the various similarity functions is found to be largely reliant different ways of matching data, as seen by the numerous correlation co-

efficients reported throughout the results section. Furthermore, it is clear that the majority of the functions utilized in this thesis have a rather weak correlation with human judgment. Nevertheless, some functions stand out from the rest: BodyText:TF-IDF, Topic:JACC, and Title:BI.

BodyText:TF-IDF has been found to have the strongest correlation. This is in line with the findings of Starke et al. [41], but the correlational strength in this work is higher overall. However, Topic:Jacc was the second strongest function in our study, while Starke et al. [41] found it to be relatively weak. It is argued that this can be attributed to how the articles are paired in the datasets. If there is a vastly disproportionate number of article pairings in the dataset that do not have a shared topic, as was likely the case for Starke et al. [41], this function will yield a low correlation since it will infer that the vast majority of article pairs are identical. Two-thirds of the article pairs in the dataset used for this study are paired on a shared topic, which has arguably led to these higher scores. This is confirmed by the positive impact that such article pairing factors have on similarity scores across functions and human judgments. Moreover, if topic is a good indicator of similarity for humans, which this research suggests that it is, then this function should have a significantly greater correlation in this work, which it does.

Title:BI also has low correlation to human judgment in the work of Starke et al. [41]. When using a dataset containing only pairings of dissimilar articles, this was also the case in our study. However, the correlation was significantly greater when the employed dataset consisted of either strictly similar article pairs or a mix of similar and dissimilar pairs. It is unclear why it is so much better at predicting similarity between similar articles than between dissimilar articles, but this is a clear example of how a function's poor performance at predicting human similarity for dissimilar articles does not necessarily preclude its use in a recommender scenario.

6.2 Is the correlation dependent on the category of the articles? (RQ1.1)

The simple answer to the question of whether the correlation depends on category is yes, at least for some of them. Five of the seventeen feature/function combinations have shown a statistically significant greater correlation for Recent Events, while two had a statistically significant greater correlation for Sport. However, the functions whose correlation is too low to be of any use in a recommender scenario should be disregarded in terms of practical use. Hence, one should focus on the top three functions from RQ1 (6.1). The trend of

continues here though, as both Topic:Jacc and BodyText:TF-IDF have correlation scores for Recent Events that are statistically significantly higher compared to those for Sport, while Title:BI has roughly the same correlation in both categories. The correlation between similarity functions and human judgment appears to be influenced by categories, as suggested by these findings.

6.3 Is the correlation dependent on whether the articles have any shared matching-characteristics? (RQ1.2)

The findings of the one-way ANOVA, Post Hoc Tukey, and Multiple Linear Regression analyses demonstrated that the influence of matching-characteristics on human judgment of similarity varies significantly depending on whether the articles belong to Recent Events or Sport. In Sport, if two articles had the same named entity, their average similarity judgment scores were quite a bit higher than that of article pairs which did not. However, in Recent Events, named entities did very little to the similarity rating. Two possible explanations for this could be the following: A) Named entities are just not regarded as particularly significant in Recent Events articles. This would be consistent with the findings of the preliminary study, which indicated that when tasked with describing an article similar to a reference article, people appeared to care much more about the topic than the named entity. B) The way named entities were tagged probably put Recent Events at a disadvantage. In the interest of consistency, the rules for extracting named entities were identical for both categories, which lead to there being a number of articles tagged "England" or "UK" in the Recent Events category. The fact that two articles share the named entity "UK" was bound to have minimal bearing on their similarity, given that all of the Recent Events articles were about news that occurred in the UK, albeit the majority of them did not explicitly state this. However, there were not enough such articles that this alone could explain why named entities appeared to have no effect whatsoever in the Recent Events category.

While Named Entity is shown to have a greater influence on similarity in the Sport category, this was not the case for Topic. Despite the fact that Topic was certainly an essential and contributing element to similarity in the sport category as well, it was far more significant in the Recent Events category. This appears to be consistent with the findings of the preliminary study, in which participants tasked with describing a similar article to a reference article were considerably more ready to stray from the topic when the reference article was about sports. Date has a minimal impact on similarity in both sports and Recent Events, but slightly more in Sport. However, there appears to be an interaction effect between Date and Named Entity

that causes the similarity to spike a bit when both are present, particularly for Sport articles. This leads to the question on how matching-characteristics affect functions? As it turns out, the answer is fully dependent on the function, as there is no answer that describes functions as a whole. Upon closer inspection of the function with the highest correlation to human judgment, BodyText:TF-IDF, it is evident that the effects of matching characteristics resemble the effect it has on humans more than any other function, which is likely part of the explanation for why it has the highest correlation. The primary difference between humans and BodyText:TF-IDF TF-IDF appears to be that TF-IDF is quite a bit more sensitive to matching characteristics. Multiple Linear Regression revealed that matching characteristics account for approximately 20% of the variance in human similarity judgment. For TF-IDF, this figure was roughly 40%. In addition, the relative importance of the various matching-characteristics differed, as BodyText:TF-IDF was substantially more sensitive to articles being matched on named entities.

To address the RQ, it seems that the correlation indeed depends on whether articles share matching-characteristics. When correlation was calculated for the dataset consisting of strictly dissimilar articles, the correlation was generally lower than when correlation was calculated for the dataset consisting of articles matched on the matching-characteristics, particularly for the functions with the highest correlation.

6.4 Are the similarity judgments dependent on demographic factors or user characteristics? (RQ1.3)

Seeing as similar item recommendation recommends the same items to everyone, it seemed pertinent to examine whether any particular demographic groups have conflicting views on what similarity entails. This was not the case, but there were still two non-demographic groups that rated articles differently: those who were confident in their judgments and those who were not.

High-confidence ratings tend to be low for dissimilar articles, considerably higher for article pairs matched on topic, and again quite a bit higher when article pairs are matched on all matching-characteristics. In contrast, ratings with low confidence tend to be quite similar to one another. Sure, articles matched on topic receive higher ratings than those who are dissimilar, but not by much. As it turns out though, the high confidence ratings also tends to have much higher correlation with the functions than the other ratings in the study. This can be seen throughout the various tables depicting correlation in the Results chapter.

This is interesting because not all similarity judgements are equally valuable. If we suppose

that those who are confident in their ratings are so for a reason (familiarity with the content of the articles, for instance), then you would want a function to correlate with those ratings. If you were reading an article and were given two lists of similar articles, one compiled by a person who is highly interested in the article's topic and the other by a person who knows nothing about the topic, you would not have a difficult time picking one of the lists. You do not want a function to correlate with human judgment on average. You want a function that correlates to the human judgment of those with knowledge on the the particular topic in question. The question therefore becomes whether or not confidence is a valid metric to measure expertise. This question cannot be answered with the data found in this study, therefore for the time being, let us presume that it is.

Assuming then that the similarity ratings which comes with a confidence rating of 5 is the ones you actually want, you would then want to know more about the distribution of the 5s. Is it a few people which are very confident in everything they rate, or are they more spread out? The latter would obviously make the 5s carry more weight. 72.2% of the people in the study have rated at least one article pair with a confidence of 5, which means the 5 as a rating is a least being used by nearly $\frac{3}{4}$ of the participants. Of course some are more liberal with their use of the 5s, as nearly 50% of the 5s come from people who have rated at least half of the article pairs with a confidence rating of 5. A little on the high end, but not alarmingly so.

One possibility though, is that certain article pairs are very easy to rate with confidence, while others are very hard to give a similarity rating. If there is a huge chunk of article pairs where no one are confident in their similarity rating, the whole argument that the similarity ratings which are made with confidence are the real ratings we should be looking at falls completely apart. I therefore looked at the article pairs which has at least three ratings, and where one of these ratings was made by someone who said their confidence in this rating was 3 or lower. In 69% of these cases, at least one of the two remaining similarity scores were made by someone who said their confidence when rating the same article pair was 5. In 98% of the cases, at least one of the two remaining similarity scores were made by someone who said their confidence when rating the same article pair was either 4 or 5. In other words: regardless of the article pair in question, there is "always" someone who is confident in their rating for said article.

A case may also be made for why the confident ratings may in fact indicate at least some expertise. It seems likely that individuals with knowledge of a certain topic would be more likely to recognize that two articles share a named entity, since they would already be familiar with the name and it would thus stand out to them. According to the multiple linear regression analysis, Named Entity is a more significant predictor of high confidence human judgments than human judgments in general. The fact that Named Entity is indeed more im-

portant to high confidence ratings, could also go a long way in answering why these ratings have higher correlation with BodyText:TF-IDF, given how sensitive the function seemingly is to named entities.

6.5 Which article features do readers employ when determining similarity between articles? (RQ2)

The results when people were asked for the most important features were in line with the answers Starke et al. [41] saw in their work. Body Text was most important, closely followed by Title, with Subheading and Topic a little behind that. Image, Date, and Author was not particularly important according to the readers. Self-assessment just represents what people believe they do, therefore it is uncertain if this really reflects which features individuals actually employ. However, they do match perfectly with the feature/function combinations with the strongest correlation to human judgment. The four highest were Topic:Jacc, Subheading:TF-IDF, BodyText:TF-IDF, and Title:BI, which correspond to the four characteristics that users ranked as the most important.

It is interesting that Title is seen as such an essential feature, yet the majority of title-based functions perform so poorly. We can only speculate as to why this is the case, but it appears likely that it is because title is a difficult feature to utilize, rather than because people overestimate its worth. The fact is that people in the study who did not pass the attention test, and so did not read the body of text, but instead likely only read the title and/or subheading, had essentially the same correlation to functions as those who read the body of text. Being able to extract all the information a title has to offer relies heavily on context and pre-existing knowledge. This is easy for humans, but nearly impossible for simple function like the ones used in this work, which are just looking at the titles in isolation. The fact that Title:BI actually was among the best functions was definitely a positive though, as using the same features as humans do, might be key when it comes to predicting article similarity in a similar fashion to humans.

6.6 Limitations and Future Work

This thesis has built upon the work of Starke et al. [41]. A limiting factor herein is that a different dataset has been used, which might introduce a bias in comparing the results from both studies. Although this has made a comparison harder, it is unlikely that the low similarity scores in the work of Starke et al. [41] only boils down to the database (i.e., the Washington

Post Corpus), which is a large and versatile database. Moreover, by comparing two different news domains in this thesis (i.e., sports and recent events), we have also safeguarded against the use of only news articles about politics, as has been done in Starke et al. [41].

The similarity judgment made with high confidence exhibited a stronger correlation with similarity functions. However, the assumption that confidence is also a sign of a higher level of news expertise has not been examined, even though this seems sensible. Future studies would benefit from validating whether this assumption actually holds.

The dataset used in this study is smaller than those used in comparable previous studies. In particular, there were only 173 article pair ratings for every condition due to the fact that there were 12 different conditions. I had intended to also examine the correlation between similarity functions and human judgments using only the condition with by far the greatest average similarity rating (Date + Topic + Named Entity). However, with so few ratings, it is very difficult to make reliable statistical inferences. Hence, instead, we have focused on comparing similarity functions to human judgments across all data or larger subsets, while predicting the impact of the news article matching (e.g., based on Date + Topic + Named Entity) in different ways.

In future work, it may be worthwhile to examine the impact of additional factors. This can include tone, style, or quality of journalism, which were also mentioned in the preliminary research. Additionally, utilizing different and superior functions may also increase correlation. As described in the background section, there are newer and improved versions of TF-IDF [5, 14], which, given the correlation TF-IDF achieved in this thesis, seem like the natural functions to test next. Other functions it could be worth testing, would be some form of a named entity-based function, for instance in the same vein as topic was utilized as a function in conjunction with jaccard in this thesis.

Furthermore, the insights from this study should be validated further. In line with Trattner and Jannach [44], the insights should be tested in a recommendation scenario. One of the objectives should be to investigate whether news articles retrieved using similarity functions that have been found to be most representative of human judgment, are also perceived as satisfactory to the use and are actually similar, compared to other similarity functions. Ideally, this would not only be in a mock-up scenario, but only on a news website with actual users, in order to test retention effects. Since it is envisioned that similar-item retrieval will continue to play an important role in personalized news recommendation, it should be done in such a way that it resonates most with users.

Bibliography

- [1] Gediminas Adomavicius and Alexander Tuzhilin. “Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions”. In: *IEEE transactions on knowledge and data engineering* 17.6 (2005), pp. 734–749.
- [2] Mária Bieliková, Michal Kompan, and Dušan Zeleník. “Effective hierarchical vector-based news representation for personalized recommendation”. In: *Computer Science and Information Systems* 9.1 (2012), pp. 303–322.
- [3] Daniel Billsus and Michael J Pazzani. “User modeling for adaptive news access”. In: *User modeling and user-adapted interaction* 10.2 (2000), pp. 147–180.
- [4] Toine Bogers and Antal Van den Bosch. “Comparing and evaluating information retrieval algorithms for news recommendation”. In: *Proceedings of the 2007 ACM conference on Recommender systems*. 2007, pp. 141–144.
- [5] Michel Capelle et al. “Semantics-based news recommendation”. In: *Proceedings of the 2nd international conference on web intelligence, mining and semantics*. 2012, pp. 1–9.
- [6] Pablo Castells, Neil Hurley, and Saul Vargas. “Novelty and diversity in recommender systems”. In: *Recommender systems handbook*. Springer, 2022, pp. 603–646.
- [7] Minjin Choi et al. “Session-aware linear item-item models for session-based recommendation”. In: *Proceedings of the Web Conference 2021*. 2021, pp. 2186–2197.
- [8] Abhinandan S Das et al. “Google news personalization: scalable online collaborative filtering”. In: *Proceedings of the 16th international conference on World Wide Web*. 2007, pp. 271–280.
- [9] Maunendra Sankar Desarkar and Neha Shinde. “Diversification in news recommendation for privacy concerned users”. In: *2014 international conference on data science and advanced analytics (DSAA)*. IEEE. 2014, pp. 135–141.
- [10] K. Falk. *Practical Recommender Systems*. Manning, 2019. ISBN: 9781617292705. URL: https://books.google.no/books?id=%5C_dbdnAAACAAJ.

- [11] Florent Garcin and Boi Faltings. “Pen recsys: A personalized news recommender systems framework”. In: *Proceedings of the 2013 International News Recommender Systems Workshop and Challenge*. 2013, pp. 3–9.
- [12] Florent Garcin et al. “Offline and online evaluation of news recommender systems at swissinfo. ch”. In: *Proceedings of the 8th ACM Conference on Recommender systems*. 2014, pp. 169–176.
- [13] Susan Gauch et al. “User profiles for personalized information access”. In: *The adaptive web* (2007), pp. 54–89.
- [14] Frank Goossen et al. “News personalization using the CF-IDF semantic recommender”. In: *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*. 2011, pp. 1–12.
- [15] The Guardian. *The Guardian Most popular newspaper*. URL: <https://www.theguardian.com/media/2021/jul/28/the-guardian-most-widely-used-newspaper-website-and-app-for-news-according-to-ofcom> (visited on 11/28/2022).
- [16] Jonathan L Herlocker et al. “Evaluating collaborative filtering recommender systems”. In: *ACM Transactions on Information Systems (TOIS)* 22.1 (2004), pp. 5–53.
- [17] Dietmar Jannach et al. *Recommender systems: an introduction*. Cambridge University Press, 2010.
- [18] Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. “News recommender systems—Survey and roads ahead”. In: *Information Processing & Management* 54.6 (2018), pp. 1203–1227.
- [19] George Karypis. “Evaluation of item-based top-n recommendation algorithms”. In: *Proceedings of the tenth international conference on Information and knowledge management*. 2001, pp. 247–254.
- [20] Lei Li et al. “Scene: a scalable two-stage personalized news recommendation system”. In: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 2011, pp. 125–134.
- [21] Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. “Personalized news recommendation based on click behavior”. In: *Proceedings of the 15th international conference on Intelligent user interfaces*. 2010, pp. 31–40.
- [22] Linyuan Lü et al. “Recommender systems”. In: *Physics reports* 519.1 (2012), pp. 1–49.
- [23] Malte Ludewig and Dietmar Jannach. “Evaluation of session-based recommendation algorithms”. In: *User Modeling and User-Adapted Interaction* 28.4 (2018), pp. 331–390.

- [24] Tapio Luostarinen and Oskar Kohonen. “Using topic models in content-based news recommender systems”. In: *Proceedings of the 19th Nordic conference of computational linguistics (NODALIDA 2013)*. 2013, pp. 239–251.
- [25] Jan Mizgajski and Mikołaj Morzy. “Affective recommender systems in online news industry: how emotions influence reading choices”. In: *User Modeling and User-Adapted Interaction* 29.2 (2019), pp. 345–379.
- [26] Tomoko Murakami, Koichiro Mori, and Ryohei Orihara. “Metrics for evaluating the serendipity of recommendation lists”. In: *Annual conference of the Japanese society for artificial intelligence*. Springer. 2007, pp. 40–46.
- [27] Maria Panteli et al. “Recommendation Systems for News Articles at the BBC.” In: *INRA@RecSys*. 2019, pp. 44–52.
- [28] Eyal Peer et al. “Data quality of platforms and panels for online behavioral research”. In: *Behavior Research Methods* 54.4 (2022), pp. 1643–1662.
- [29] Raymond K Pon et al. “Tracking multiple topics for finding interesting articles”. In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2007, pp. 560–569.
- [30] Anil Poriya et al. “Non-personalized recommender systems and user-based collaborative recommender systems”. In: *Int. J. Appl. Inf. Syst* 6.9 (2014), pp. 22–27.
- [31] Pearl Pu, Li Chen, and Rong Hu. “A user-centric evaluation framework for recommender systems”. In: *Proceedings of the fifth ACM conference on Recommender systems*. 2011, pp. 157–164.
- [32] Anand Rajaraman and Jeffrey David Ullman. *Data mining (pp. 1–17)*. 2011.
- [33] Shaina Raza and Chen Ding. “News recommender system: a review of recent progress, challenges, and opportunities”. In: *Artificial Intelligence Review* (2021), pp. 1–52.
- [34] Francesco Ricci, Lior Rokach, and Bracha Shapira. “Introduction to recommender systems handbook”. In: *Recommender systems handbook*. Springer, 2011, pp. 1–35.
- [35] Julio Rieis et al. “Breaking the news: First impressions matter on online news”. In: *Proceedings of the international AAAI conference on web and social media*. Vol. 9. 1. 2015, pp. 357–366.
- [36] Gerard Salton, Anita Wong, and Chung-Shu Yang. “A vector space model for automatic indexing”. In: *Communications of the ACM* 18.11 (1975), pp. 613–620.
- [37] Amit Singhal et al. “Document length normalization”. In: *Information Processing & Management* 32.5 (1996), pp. 619–633.

- [38] Jieun Son and Seoung Bum Kim. “Content-based filtering for recommendation systems using multiattribute networks”. In: *Expert Systems with Applications* 89 (2017), pp. 404–412.
- [39] Stuart Soroka, Lori Young, and Meital Balmas. “Bad news or mad news? Sentiment scoring of negativity, fear, and anger in news content”. In: *The ANNALS of the American Academy of Political and Social Science* 659.1 (2015), pp. 108–121.
- [40] Gabriel de Souza Pereira Moreira, Felipe Ferreira, and Adilson Marques da Cunha. “News session-based recommendations using deep neural networks”. In: *Proceedings of the 3rd workshop on deep learning for recommender systems*. 2018, pp. 15–23.
- [41] Alain D Starke, Sebastian Øverhaug, and Christoph Trattner. “Predicting Feature-based Similarity in the News Domain Using Human Judgments”. In: *15th ACM Conference on Recommender Systems, RecSys 2021*. 2021.
- [42] Xiaoyuan Su and Taghi M Khoshgoftaar. “A survey of collaborative filtering techniques”. In: *Advances in artificial intelligence* 2009 (2009).
- [43] Dongting Sun, Zhigang Luo, and Fuhai Zhang. “A novel approach for collaborative filtering to alleviate the new item cold-start problem”. In: *2011 11th International Symposium on Communications & Information Technologies (ISCIT)*. IEEE. 2011, pp. 402–406.
- [44] Christoph Trattner and Dietmar Jannach. “Learning to recommend similar items from human judgments”. In: *User Modeling and User-Adapted Interaction* 30.1 (2020), pp. 1–49.
- [45] Merriam Webster. *MerriamWebster News Definition*. URL: <https://www.merriam-webster.com/dictionary/news> (visited on 11/15/2022).
- [46] Yuan Yao and F Maxwell Harper. “Judging similarity: a user-centric study of related item recommendations”. In: *Proceedings of the 12th ACM Conference on Recommender Systems*. 2018, pp. 288–296.
- [47] Cai-Nicolas Ziegler et al. “Improving recommendation lists through topic diversification”. In: *Proceedings of the 14th international conference on World Wide Web*. 2005, pp. 22–32.

../references

