UNIVERSITY OF BERGEN

Department of Information Science and Media Studies

MASTERS THESIS

Evaluating Feature-Specific Similarity Metrics using Human Judgments for Norwegian News

Author: Daniel Rosnes

Supervisor: Prof. Dr. Christoph Trattner

Co-supervisor: Assoc. Prof. Dr. Alain D. Starke

June 2, 2023

Abstract

This master's thesis delves into the measurement of similarity between news articles within the Norwegian news domain. Four central questions form the basis of the thesis: the identification of information cues utilized by readers, the effectiveness of specific similarity metrics, the comparison with other domains, and the exploration of differences in human similarity ratings between national and local news.

Key findings include that a Sentence-BERT metric, applied to the body text, best represented human similarity judgments. Compared to other news domains, the Norwegian news domain showed stronger correlations for a majority of the metrics.

A minimal contrast was observed between human ratings for local and national news, with local news considered slightly more similar. This disparity between local and national levels, however, did not markedly impact how metrics represented human similarity judgments. The findings from this thesis may provide valuable insights for enhancing news recommendation systems within the news sector.

Acknowledgment

I want to express my sincere gratitude to my main supervisor, Full. Prof. Dr. Christoph Trattner, for his support and invaluable guidance throughout the work on this thesis, and for allowing me to work with my thesis during my time at MediaFutures. Additionally, my co-supervisor Assoc. Prof. Dr. Alain Starke deserves substancial recognition for his significant contribution during the later stages of the thesis.

I'd like to extend my appreciation to Dr. Samia Toulieb at MediaFutures, whose expertise in NLP proved invaluable for my thesis. Equally, my thanks go to Dr. Erik Knudsen, also at MediaFutures, whose deep understanding of Norwegian News Media and practical advice about Norwegian participants were significant in my work.

I want to thank Assoc. Prof. Dr. Mehdi Elahi with work package 2 at MediaFutures for enlightening me about recommender systems, and for his gracious invitation to the work package 2 meetings, which offered constructive feedback from key partners at MediaFutures.

I also express my gratitude to Emiliano Guevara, whose help and dedicated support in acquiring and curating the Amedia dataset were invaluable. His explanations of the dataset's diverse features were valuable to my work.

Next, I would like to thank Eivind Throndsen with Schibsted and Thomas Husken with Bergens Tidende for their assistance with the Schibsted dataset and for arranging meetings with key individuals within Schibsted.

Finally, I want to extend my heartfelt thanks to all of my fellow students and colleagues at MediaFutures and the Department of Information Science and Media Studies at the University of Bergen. Your camaraderie and collaborative spirit enriched my academic journey immensely.

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

Thank you all!

D. R.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The abundance of information in today's digital landscape, particularly in news dissemination, underscores the need for tools that can effectively sift through vast content repositories and guide users toward relevant and engaging materials. To this end, recommender systems have emerged as crucial instruments, helping to streamline information discovery, optimize content delivery, and enhance the overall user experience [24].

The news domain faces several domain-specific challenges that make the introductions of common recommender system strategies difficult [18, 26]. Similar-item recommenders are able to circumvent many of these challenges [26]. While such recommenders are popular with news websites, there is limited knowledge surrounding whether the recommendations they provide represent what users consider similarity [49]. While there are studies exploring this [52, 49], the studies are generally done with limited data. Such as using single publications, a limited number of categories within publications, and/or a limited amount of news articles.

No studies I could find have explored this problem in the Norwegian language, and there is limited knowledge of how new advances within the field of Natural Language Processing (NLP), such as BERT models, compare to traditional techniques of similarity measurements when evaluated against human similarity judgments.

In this thesis, I attempt to explore these issues by exploring how feature-specific similarity metrics represent human similarity judgments in four different Norwegian publications that span the local and national domains.

1.2 Problem Formulation

The problem addressed in this thesis is the analysis of human similarity judgment representations by similar-item recommenders across local and national levels of Norwegian news publications. The primary objective is to examine the variations in these representations by exploring select publications, which encompass both local and national newspapers. Additionally, the study aims to assess the efficacy of a set of feature-specific similarity metrics, derived from recent advancements in language technologies, in comparison to traditional measures of similarity for news articles.

1.3 Objectives

- **RQ1:** Which specific features do Norwegian users use to evaluate the similarity between news articles?
- **RQ2:** To what extent do feature-specific similarity metrics represent human similarity judgments in the Norwegian news domain?
- **RQ3:** How do the feature-specific metric representations in the Norwegian news domain compare with feature-specific metric representations in other domains?
- **RQ4.1:** Does the strength of human similarity judgments towards Norwegian news media differ across local and national outlets?
- **RQ4.2:** To what extent do feature-specific similarity functions represent human judgment across local and national Norwegian news media outlets?

1.4 Contribution

The goal of this master's thesis is to explore and evaluate feature-specific similarity metrics and how they represent human similarity judgments in the Norwegian news domain. A primary task is also to explore how human similarity judgments may vary between national and local news domains, and how this may affect feature-specific similarity metrics. By doing this I will make the following contributions:

- To the best of my knowledge, this is the first study of its kind investigating feature-specific similarity functions using human judgments for Norwegian language news. All previous work found have done so only for English language news.

- I add to the knowledge of similarity functions used in recommender systems, by comparing these findings to similar work in other domains, specifically Trattner and Jannach [53], Starke et al. [49], and Solberg [48].
- Compared to the previous studies, which have focused on national-level news across single publications, this study also looks at local-level news across multiple publications.
- Finally, I provide insights into how current state-of-the-art NLP methods represent human similarity judgments. The previous work by Trattner and Jannach [53], Starke et al. [49], and Solberg [48] did not include such methods in their analyses.

1.5 Outline

This thesis is split into five Chapters. The first chapter is the introduction chapter you are currently reading. This chapter is followed by Chapter 2 which lists the most relevant work for this thesis. Explaining the literature surrounding the similar-item news recommenders as well as detailing the work comparing feature-specific similarity metrics to human similarity judgments that this thesis is a direct follow-up to. Chapter 3 details the data used and gives a thorough explanation of all the metrics used in the study. It also details the survey in which the human similarity judgments were collected. Chapter 4 lists the results of the study, and attempts to answer each research question. Finally, Chapter 5 discusses the results and limitations of the study, and suggests possible directions for further research.

Chapter 2

Background

In this chapter, I give an overview of relevant work to the problem I am seeking to research in the thesis. It is split into five sections. Section 2.1 introduces the domain of Similar Item Retrieval. Section 2.2 further narrows down the domain and explain the area of News Recommenders. Section 2.3 explains how recommender systems in general, and news recommenders particularly, are evaluated. Section 2.4 takes a deep-dive into the specific work that this thesis builds directly upon. Finally, Section 2.5 summarizes the problem and lists some key differences between the previous work and this thesis.

2.1 Similar Item Retrieval

One of the core fields of Information Science is that of Information Retrieval (IR). It forms the basis of many of the online services we use every day [28]. The objective in this field, as implied by its name, is to fulfill the user's request by providing them with a desired item. In *Similar Item Retreival*, it is to provide a *unseen* or *novel* item that is similar to a specific base item [49]. A key question then becomes how to compute the similarity between the base item and possible items to be retrieved. [41, 57].

Similar Item Retrieval is often performed with the use of recommender systems. These are systems that attempt to provide a user with recommendations, usually by providing a list of ranked recommendations given a specific input [24, 45]. The approach to solving this is generally categorized into three different types: Collaborative Filtering (CF), Content-Based (CB), and Knowledge-Based (KB). Approaches combining any of these techniques are referred to as Hybrid recommenders (H) [24].

CF is the approach of using historic interactions between users and items in order to calculate the probability of a specific user preferring a specific item. CB is the approach of

evaluating the content characteristics, or *features*, of items and/or users, and the similarity between these features, to estimate a probability of a specific user preferring a specific item. KB approaches use specific domain knowledge expertise to pair users and items [24, 35, 45].

CB approaches are particularly used when there is no user information available. This is often because it is not affected by the *cold-start* problem that many recommender systems suffer from [24, 14]. In such situations, a CB recommender will instead evaluate *specific features* of the items available in the recommender, and provide a recommendation based on the feature-based similarity of the items. This approach is formalized through the use of *similarity functions* [24].

Similarity functions generally follow a convention of taking in two items and returning a coefficient with the range 0 to 1, or -1 to 1, where a higher value indicates higher similarity between the items that are compared [54]. Given a large set of items, we can then rank the items by their resulting similarity coefficients in order to retrieve similar items [24].

2.2 News Recommenders

Recommender systems often have to overcome domain-specific challenges [24, 45]. This leads to their categorization based on the domain for which they are designed. One notable domain is the *news* domain. A survey conducted by Karimi et al. [26] identifies various domain-specific challenges for this domain. Among these challenges is the high volatility of a news article's relevance. Furthermore, a user's interest in news articles may vary due to several factors, such as the time of day, the user's location, and the features of the device through which they are consuming news [26].

Karimi et al. [26] also detail what they refer to as the *permanent cold-start problem*. This problem arises from the lack of historic information from users. It is also caused by the prevalence of one-time and first-time users. Further compounding the problem is the high frequency of new news items. This situation creates a challenge for common recommendation algorithms, which typically utilize CF. These algorithms are not suitable due to their susceptibility to the cold-start problem. In their survey, Karimi et al. [26] shows that a majority of news recommenders use CB algorithms or hybrid algorithms with a CB component. They present that out of 112 articles proposing news recommendation algorithms, 104 propose either CB algorithms or hybrids with a CB component.

In addition to the issues already mentioned, Elahi et al. [18] lists several potential undesired effects, like *filter bubbles*, *echo chambers*, and *spread of misinformation* that may occur with *personalized* news recommendations. By using item-based similar item recommendations

many of these issues can be avoided, as content-based recommenders generally are not affected by cold start problems [24], and can be used without any form of personalization. This may be the reason for their popularity on news websites. Visit any online news website and you are likely to be met by a content-based similar item recommender.

Such similar item news recommenders generally employ *feature-specific similarity metrics*. In particular, they usually involve evaluating the article's text or title, while other features are ignored [26]. A traditional method to compute the similarity between text items is by deriving vectors from the text [49]. *Term Frequency-Inverse Document Frequency* (TF-IDF) remains one of the most commonly used IR methods to create similarity vectors from text [5][49]. It works by taking the frequency of a set of words, the *terms*, in a document and multiplying it by the inverse of the frequency of the words across all documents [36]. The end result is that texts that have a higher frequency of the same *rare* words will have similar vectors. These vectors can then be compared using cosine similarity [10, 24][49].

While TF-IDF is still popular, it has been outperformed by other metrics, such as BM25 [37][49]. In recent years approaches using transformer models and Word2Vec also show better performance than TF-IDF on text similarity tasks [11, 33]. Since the introduction of transformer models with the Bidirectional Encoder Representations from Transformers (BERT) model in [55], the use of such models has received immense popularity. In recommender systems there are several approaches utilizing the embeddings provided by various transformer models [22, 27, 61], and combining transformer models with topic modeling techniques [34, 58, 62]. As with TF-IDF these approaches generally extract vectors from text which can then be compared using cosine similarity.

2.3 Recommendation Evaluation

Recommender systems are typically evaluated in one of the following three approaches: through offline experimentation and simulation based on historical data, through laboratory studies, or through A/B (field) tests on real-world websites [26]. In their survey Karimi et al. [26] found that a large majority of studies relied on traditional IR measures like precision and recall, rank-based measures like *Mean Reciprocal Rank* or *Normalized Discounted Cumulative Gain*, or prediction measures like the *Root Mean Square Error*. These methods all rely on a dataset annotated based on the task the recommender is meant to solve. Such datasets are not readily available in the news domain [26].

While only 19 of the 112 papers surveyed by Karimi et al. [26] utilize it, *click-through-rate* (CTR) is a popular way to evaluate the performance of news recommenders [20]. However, CTR is not helpful in determining if the items are similar, as the user may click on the item

for other reasons than similarity [15].

In order to validate the performance of similar-item recommenders, *human judgments* are typically used [8]. A critical question is to what degree similarity functions mirror a user's judgment of the similarity between pairs of items. Problems could arise if a user undervalues or overemphasizes specific item features compared to which is calculated, and how the similarity is being calculated [49, 57].

Yao and Harper [59] collect human similarity judgments using movie pairs collected from the MovieLens[1] dataset. As part of their study, users are asked to what extent the movies are similar, and whether they would recommend the second movie to someone who likes the first. Their goal is to explore whether CF or CB algorithms provide similar item recommendations that are closer to human similarity judgments. Yao and Harper [59] suggest that CB algorithms perform better in matching human similarity judgments. Another key observation in Yao and Harper [59] is that similarity is not everything in a similar item recommender: Over 60% of the users in their survey choose a compromise over being recommended the most similar item.

2.4 Directly Related Work

Other studies where human judgments have been collected in order to evaluate similar item recommenders include Trattner and Jannach [53], Starke et al. [49], and Solberg [48]. This thesis builds directly on the work done in these studies. The main methodology of calculating feature-specific similarity metrics and comparing them with human similarity judgments used in this thesis is introduced by Trattner and Jannach [53]. Starke et al. [49] then applies the same methodology to the news domain. Solberg [48] first attempts to discover *news recommender criteria*, before he uses a similar methodology to that of Trattner and Jannach [53] and Starke et al. [49] to examine differences between categories in the news domain. The next subsections details the different approaches.

Movie and Recipe Domain

In the initial work by Trattner and Jannach [53] two main studies are performed across the movie and recipe domains. The studies follow a novel approach where the goal is not to evaluate existing algorithms, but to develop new similarity functions from human similarity judgments. The human similarity judgments are used as baselines for how similar the items are, and what makes the two items similar. Trattner and Jannach [53] also asks the users

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

which *similarity cues* the users used while evaluating the similarity. These similarity cues represent the features the feature-specific metrics are based on.

In the study in the recipe domain, Trattner and Jannach [53] use a dataset based on *allrecipes.com*, with the following recipe features: *Title*, *image*, *ingredients*, and *directions*. They employ a total of 17 feature-specific similarity metrics. In the study in the movie domain, the researchers utilize the MovieLens dataset, which includes the following features for each movie: *Title*, *image*, *plot*, *genre*, *director*, *date*, and *tags*. They evaluate 20 feature-specific similarity metrics in this study.

The pairs that the users evaluate are prepared by first calculating the metrics across the dataset and then combining and averaging them. Afterward, a biased stratified sampling strategy is applied, dividing the pairs into three groups based on the mean computed similarity score. These groups consist of the 20% lowest-scoring pairs, the 60% middle-scoring pairs, and the 20% highest-scoring pairs. Finally, the pairs are sampled equally across these groups.

For both studies, Trattner and Jannach [53] collect human similarity judgments using crowdworkers on Amazon Mechanical Turk. The task of the workers is to evaluate the similarity of pairs of items on a 5-point Likert scale. Each item is presented to the user along with all of its features. Subsequently, the workers are asked to indicate the extent to which they used each of the features to evaluate the similarity of each pair.

The objective of Trattner and Jannach [53] is to develop a specific similarity function by employing machine learning approaches. In each domain, they develop and test an offline similarity function based on the feature-specific similarity metrics. Later, the best-performing function is evaluated using a new survey. This survey serves to validate the approach and demonstrate the feasibility of creating similarity functions by training models based on human similarity judgments.

Another important finding in Trattner and Jannach [53] is that the reported *information cues*, which refer to the features that participants reported using in their evaluation of item similarity, do not serve as reliable predictors of which features and feature-specific similarity metrics yield accurate predictions of similarity.

News Domain

In Starke et al. [49], a similar approach to Trattner and Jannach [53] is employed, but this time in the news domain. The articles used in the study are from the TREC Washington Post dataset[2]. A total of 2400 articles are included, with 400 articles randomly sampled from

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

each year between 2012 and 2017. Additionally, the articles are restricted to the 'Politics' category. The dataset provides several features for each article, including *Subcategory*, *title*, *image*, *author*, *date*, *body-text*, and *author*. A total of 20 feature-specific similarity metrics are developed for this study.

Following the method put forward by Trattner and Jannach [53], a survey is conducted to collect human similarity judgments. Crowd-workers from Amazon Mechanical Turk participated in the survey. Interestingly, the obtained similarity judgments exhibited low correlations with the metrics across all aspects, with an average Spearman correlation coefficient of 0.092. Among the metrics, the highest correlating one was TF-IDF when applied to body-text, demonstrating a correlation coefficient of 0.29.

The low correlations observed in the study may be attributed to the fact that the mean similarity judgments provided by the users were low, with an average rating of 1.8 out of 5. This indicates that a significant portion of the article pairs presented to the users were perceived as dissimilar.

News Recommender Criteria

In his study, Solberg [48] addresses two primary problems. The first problem focuses on defining the criteria for news recommendation, while the second problem aims to explore the differences between specific news categories, namely *Sports* and *Recent Events*. The thesis is divided into two separate studies, each addressing one of these questions.

The first study is a qualitative survey asking the participants three questions. The first question asks the participants' opinion on the factors a news recommender should consider when suggesting the next articles for a reader to view after they've finished an article. The second question asks the participant to describe a similar article, real or hypothetical, to an article that is presented to the participant. The third and final question asks the participant for the single biggest factor they consider as the determinant for whether two news articles are similar.

Similar to Yao and Harper [59], the answers to the first question show that 26 of the 45 participants in the study listed item similarity as a factor. While this was the most common response, it still shows that item similarity isn't everything a recommender should take into account when recommending the next articles to read [48].

The next question either showed an article about Boris Johnson and Covid, or about Sadio Mané and Liverpool. In the description of the similar news articles, there were key differences in responses between the two articles. The participants who were shown the first article, all proposed to present covid-related news articles. While the participants who were

shown the second article mostly suggested articles about the football player Mané or an entity related to him. A majority of the responses to the final question pointed out *shared topic* to be the single biggest factor that determined the similarity between articles with 29 out of 39 responses [48].

Comparing Categories

The second study that Solberg [48] performs is a survey similar to those of Trattner and Jannach [53] and Starke et al. [49], where pairs of articles are presented to the participants, who are then asked to rate their similarity. The articles used for the survey were collected from the British Newspaper The Guardian[3] [48]. A total of 385 articles were manually collected based on specific criteria. The criteria for collection are based on *familiarity*, *recency*, and *covid-19*. To summarize, the criteria for selecting articles can be described as choosing topics that are generally familiar[4] and were published between 2019 and 2021, while excluding any articles related to COVID-19.

The dataset of Solberg [48] have the following features available: *Subcategory*, *title*, *subheading*, *image*, *author*, *date*, *body-text*, and *author bio*. A total of 17 *feature-specific similarity metrics* are used. As Solberg [48] is attempting to analyze differences between categories, the sampling strategy is different from that of Trattner and Jannach [53] and Starke et al. [49]. Instead of using the biased stratified sampling strategy, Solberg [48] instead creates pairs based on specific features and the articles' affiliation to the category that is analyzed.

Solberg [48] recruited participants for his second survey from Prolific[5]. His study shows higher correlations on the various metrics than Starke et al. [49]. The two highest correlating metrics in Solberg [48] are TF-IDF performed on the main Text (0.52) and the Jaccard similarity of the Tags (0.45). Finally Solberg [48] explores the differences between similarity ratings of the two categories he analyzed, where he shows some minor differences between the performance of specific similarity metrics across the two domains.

2.5 Summary and Key Differences

The news recommender domain faces several domain-specific challenges that are yet to be overcome. Several of these challenges obstruct recommender approaches that are successful in other domains. One promising class of recommenders for the news domain, are those

[3] https://wwww.theguardian.com
[4] As opposed to obscure or unknown.
[5] https://www.prolific.co/

of *similar item* recommenders. However, such recommenders are not without their own challenges.

When only evaluating items, similar item recommenders have to rely on the characteristics of the items that are being recommended. This is done by calculating a similarity metric. However, when developing such recommenders, little work is done to estimate how well these similarity metrics represent human judgments of similarity. When these metrics are evaluated against human similarity judgment, finding good similarity metrics has proven challenging. Much of the research is done on limited data, which in itself can be a problem.

This study diverges from previous work by expanding its focus beyond specific categories within single news publications. It undertakes a comprehensive evaluation of news articles across multiple publications, and also across different geographical domains, contributing to a broader understanding of the subject matter.

Uniquely, this study investigates feature-specific similarity functions using human judgments for Norwegian language news, a first in this domain where previous investigations have been conducted primarily for English language news. In addition, it extends the understanding of similarity functions used in recommender systems by contrasting these findings with similar work in other domains. This detailed analysis includes not only national-level news, as previous studies have done, but also local-level news, allowing for a more nuanced view of different publication levels.

An advancement of this study lies in its application of recent developments in Natural Language Processing (NLP) to evaluate their effectiveness in representing human similarity judgments. This provides novel insights into the capabilities of current state-of-the-art NLP methods, an aspect overlooked in previous work.

Despite the limitation of restricting the articles to only those from 2022, this study compensates by conducting a comprehensive analysis that encompasses diverse news sources and geographical domains. By incorporating these significant variations, the study expands on existing knowledge and offers a fresh perspective to the field.

Chapter 3

Methodology

In this chapter, the data and methods used in this thesis are described. The chapter is split into four sections: Section 3.1 describes the datasets used in the thesis, the process of cleaning them, and the specific features that are used for calculating the similarity metrics. Section 3.2 describes all the similarity metrics used. Section 3.3 describes the process of collecting the human similarity judgments. Finally, Section 3.4 lists the statistical methods used in the analysis of the results.

3.1 Dataset

The dataset used for this thesis is a combination of data from four separate publications from two separate media organizations. The datasets were obtained through the MediaFutures research institute[1] and consist of publications from two of the MediaFutures industry partners, Amedia[2] and Schibsted[3].

The datasets were selected based on the following criteria:

Contain Local and National news. The main research question of this thesis is to find any differences between Human Similarity Judgments between the National and Local news domains. Available large-scale datasets were considered, but none were found to have the sufficient geographical granularity to isolate a clear *local* news domain. Because of this, it was decided that a specific dataset would have to be obtained or created.

Participant availability. One challenge identified early on was the potential struggle of obtaining participants for the Human Similarity Judgment survey. Considering that a local

[1] https://mediafutures.no/about/
[2] https://www.amedia.no/english
[3] https://schibsted.com/about/we-are-schibsted/news-media/

Table 3.1: Features available in the Amedia raw dataset

Feature	Description
Content ID	The ID of article, internally used by Amedia
Publication	The publication that published the article (BA or Nettavisen)
Date	Date of publication
URL	The URL to the article
Authors	Hashed author names
Title	Title of the article
Lead Text	Lead text of the article
Body	Raw HTML of the article text
Processed Text	Article text ready for tokenization
Tags	Manual tags of the article
Predicted Category	The automatically predicted article category
Top Image URL	URL to the main image of the article
Top Image Caption	The caption of the main image

news domain would also require local participants for the survey, overly restricting the definition of *local*, or restricting it to an area where potential participants are difficult to contact, could create unwanted challenges. Because of this, the local domain was chosen to be the Bergen area. As a result of this, the national domain is Norway.

Recency. In the news domain time is a very important factor. The lifespan of breaking news is generally very short, down to a few hours [13, 12]. Conducting an experiment to collect Human Judgments in such a timespan would, while interesting, not be feasible for this thesis. To avoid the problem of recency affecting the similarity ratings, the most recent articles should therefore be avoided. At the same time, news articles may risk losing their relevance entirely if they become too old. Considering these limitations, it was decided that the dataset should only contain articles from 2022.

Comparable Features. Since this thesis builds directly upon the work done in Trattner and Jannach [53], Starke et al. [49] and Solberg [48], this work should be able to be directly compared to those works. This necessitates some level of comparability of the features. The specific process of the selection of the features is detailed in section 3.1.3

3.1.1 Publications

Amedia

The Amedia dataset consists of the two publications *Bergensavisen (BA)* and *Nettavisen*. The dataset was obtained from Amedia directly and was tailored based on the criteria above. The full list of data available for each article can be seen in Table 3.1. Some notable aspects of

Table 3.2: Q4 2022 ranks and daily readership[6]for online versions of publications used for the thesis, as well as the number of articles in raw datasets.

Publication	Rank	Readership	Raw Articles
VG	# 1	1 957 961	17 686
Nettavisen	# 7	529 582	20 051
BT	# 16	184 514	17 444
BA	# 22	97 658	8 653

this dataset are the inclusion of both raw HTML article text, as well as processed text. The processed text field contains text that is ready for tokenization. This text differs slightly from a grammatically correct text in that there are spaces surrounding most punctuation, except punctuation that is part of words, for example, acronyms. In addition, it features a field with hashed author names. The hashing is due to strict GDPR rules within Amedia.

BA. The local newspaper in the Amedia dataset is *Bergensavisen* (BA). BA was founded in 1927 as the labor movement's newspaper in Bergen. BA is the second largest newspaper in Bergen, after Bergens Tidende (BT), and is primarily a pure local newspaper that journalistically covers the city and its immediate surroundings more closely and in more detail than BT [51]. Its coverage is *Bergen, Askøy, Fjell and Os*[4]. BA is the smallest newspaper across the datasets in both daily readerships, with 97 658 daily readers, and the number of articles, with 8 653 articles available in the dataset.

Nettavisen. The national newspaper in the Amedia dataset is *Nettavisen.* It was founded in 1996 as the first Norwegian online newspaper that did not have a print edition [42]. It is Amedias largest newspaper in terms of readership and their only general national news outlet. Its daily readership in Q4 2022 was 529 582. The raw dataset contains 20 051 Nettavisen articles[5] from 2022.

Schibsted

The Schibsted dataset is comprised of the publications *Bergens Tidende (BT)* and *Verdens Gang (VG)*. The dataset was obtained through a corpus project at MediaFutures. Because of this the dataset is not specifically tailored for the study. The corpus contained most of the online news articles from Schibsted publications dating back to 1994. For this thesis, however, we are only using articles from 2022 from BT and VG. The full list of features available in the dataset can be seen in Table 3.3.

[4] https://www.amedia.no/aviser/amedias-aviser/bergensavisen
[5] During cleaning it was discovered that a large amount of the Nettavisen articles might be duplicates
[6] https://www.medietall.no/index.php?liste=persontall&r=PERSONTALL

Table 3.3: Features available in the Schibsted Data

Feature	Description
UUID	The ID of article, internally used by Schibsted
Title	The title of the article
Newsroom	The Newsroom that published the article (BT or VG)
Creation Date	Date the article was originally published
Last Modified Date	Date of which the article was last modified
Tags	Semantic tags manually annotated by the newsroom
Lead Text	Lead text of the article
Body Text	Crudely cleaned HTML text of the article
Section	The section of the Article

BT. The local newspaper in the Schibsted dataset is Bergens Tidende (BT). Founded in 1868 it is one of the oldest newspapers in Norway that is still being published. It is the largest Norwegian newspaper outside of Oslo and the dominating media outlet in Western Norway [16]. Its daily readership in Q4 2022 was 184 514, and the raw dataset contains 17 444 news articles published in 2022 by BT.

VG. The national newspaper in the Schibsted dataset is Verdens Gang (VG). VG is the largest online newspaper in Norway measured in readership, with a daily readership in Q4 2022 of 1 957 961. The raw dataset includes 17 686 VG news articles from 2022.

3.1.2 Dataset Cleaning

Before applying the metrics some dataset cleaning was performed. The main motivation behind this were findings in Starke et al. [49] and Solberg [48]. In particular, the pre-study in Solberg [48] found that for large topics, in his case Covid-19, the reader would be particularly focused on the large topic, rather than other contents in the article. The intuition is that for major topics like this, the articles will be considered similar based on the topic alone, and not other similarity features. Some of these topics will also have weekly or daily summaries that are likely to be highly similar. For 2022 there were three major topics to be removed: *Covid-19*, *War in Ukraine*, and *Power crisis*.

In order to remove these topics a tag-based filtering strategy was employed. Since the tags are manually annotated by the journalist, there is a probability that some articles are not properly tagged. However, after manually reviewing the dataset it was considered that sufficient articles were removed to mitigate the problem of some topics being too prevalent. In addition to removing articles of prevalent topics, the same tagged-based strategy was utilized to remove periodical articles as well as tag-groups with a high grade of similarity between articles. The full list of tags removed can be seen in Table 3.4.

Table 3.4: Full list of tags removed from each dataset

Dataset	**Tags**
Schibsted	Russland, Ukraina, Krigen i Ukraina, Coronaviruset, invasjonen av Ukraina, Strømpriser, Koronaviruset, eavissalg, Volodymyr Zelenskyj, Kryssord, Minneord, Frode Thuen
Amedia	oddstips, travtips, tippetips, ukraina, gratis-travtips, russland, strømpriser, summetonen, debatt, norsk-tipping, korona, strømpris, vladimir-putin, koronaviruset, norsk-rikstoto, ukraina-krigen, galopptips, strømkrise, russisk-invasjon, vikinglotto, salg, nettavisen-nettbutikk, sparetips, v75, debatt, meninger, korona, leder, eurojackpot, lotto, søndagskupongen, polsk, stalltips, sexologen-svarer, finere-fanafruers-forening

The text column in the Schibsted dataset seemed to have undergone a basic HTML cleaning process, but it was not done thoroughly. Specifically, there were numerous missing spaces, causing the tokenizers used for various metrics to generate vocabulary entries that combined symbols and characters that should have been separated by spaces. Because of this, a cleaning strategy using regular expressions to manipulate the text was employed. The goal of this was to have the format of the Schibsted text data match the format of the Amedia *processed text* data as much as possible. For each regular expression made, a manual review was taken of article texts to find edge cases that were added to the list of expressions until no more edge cases could be found.

Since one of the features that are considered is *Images*, all articles that did not contain a URL to an image were removed. In a later stage, all articles where the image could not be retrieved or could not be processed when calculating the metrics were also removed. In addition, articles that did not contain one of the other features listed in Table 3.6 were removed. Articles that were very long or short were also removed. The specific method for the length-based removal was to remove articles where the string length was less than 1000 characters or more than 10000 characters. This is equal to roughly 3% shortest and longest articles in the dataset. Finally, all articles were divided into separate datasets for each publication, and articles that had duplicate title and text fields within each publication were removed. Curiously, this last step removed nearly half of the remaining Nettavisen articles, indicating that the raw dataset contained a large amount of duplicate Nettavisen articles.

Key figures of the datasets after cleaning can be seen in Table 3.5. The distribution of the articles across the different sections of the publications can be seen in Figure 3.1. In Starke et al. [49] and Solberg [48] a subset of categories was used. In this study however, all sections of the newspaper are used, and the separation between the local and national news domains is instead done by a publication-level separation of the dataset.

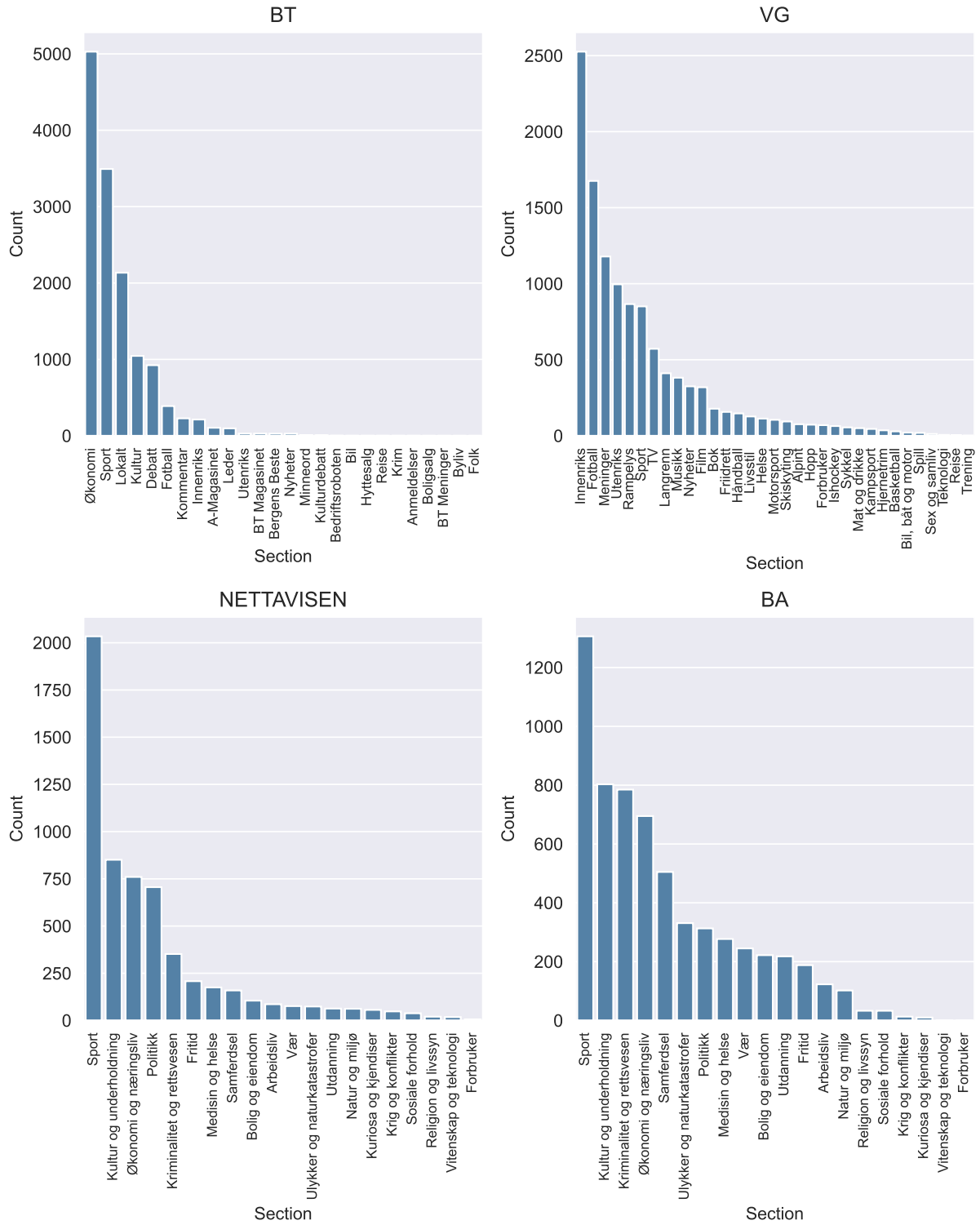

Figure 3.1: Distribution of articles per section for each of the publications

Table 3.5: Cleaned dataset main numbers. The numbers for the Text and Title features are based on word counts.

Feature		**BA**	**Nettavisen**	**BT**	**VG**
Articles		5,865	5,468	13,808	11,587
Sections		20	20	26	33
Tags	Unique	3,538	5,028	7,118	6,154
	Min	3	1	1	1
	Max	18	23	32	35
	Mean	3.99	3.46	4.51	4.05
Text	Min	171	173	149	156
	Max	2,010	2,070	1,970	2,095
	Mean	663	721	654	701
Title	Min	1	1	1	1
	Max	30	33	30	23
	Mean	10.75	10.33	9.67	9.18

3.1.3 Features

During the process of cleaning the datasets, the candidate features were also selected. The selection of features is based on the features used in Trattner and Jannach [53], Starke et al. [49], and Solberg [48]. In Starke et al. [49] the following features were used: *Title*, *Image*, *Body Text*, *Subcategory*, *Date*, *Author* and *Author Bio*. In Solberg [48] a selection of the previous features was used, as well as *Subheading*. He also evaluated *Named Entities*, but it is unclear if it was considered a feature or a metric. In this thesis, it is defined as a metric used on the text feature and is explored in section 3.2.2.

The features in the datasets are listed in Tables 3.1 and 3.3. Based on the features available, the *Author Bio* feature was immediately ruled out as it was not contained in any of the datasets, and could not be mapped to any of the features. While the Amedia dataset did contain hashed *author names*, it too was ruled out. This was primarily because it was not available in the Schibsted dataset. The fact that the author names is hashed does also pose a challenge in terms of how they could be displayed during the survey to collect the human judgments, but mapping the hashes to strings that appear like names, essentially pseudonyms, could have been a way to mitigate this.

In both the Amedia and Schibsted datasets there is a feature named *lead-text*. However, a lot of articles, especially in the Schibsted dataset, did not utilize this feature. During manual review of the usage of the feature, it also appeared that the usage was highly inconsistent, only rarely being used as a traditional lead-text paragraph. It was more commonly used as a

Table 3.6: News article features used in study

Feature	**Description**
Date	The UNIX-time of the publication date
Section	List containing Section or Sections
Tags	List containing the tags
Title	Title text
Text	Tokenized article main text
Image	The main image

subheading or a summary, and other times only contained a single word or quote. Because of this, it was not used as a feature for the thesis. Another alternative could be to merge the lead-text with the main text of the articles. While reviewing this however, it was found that in most cases this resulted in a jarring experience for the reader, as it was clearly distinct from the rest of the text. Because of this, the *lead-text* was not used. The *subheading* feature used in Solberg [48] was also scrapped as no other features could be mapped to it.

Another feature present in the previous studies and both of the datasets is the category. In Starke et al. [49] the category feature used is the *subcategory*, while in Solberg [48] a feature named *topic* had similar properties. The Schibsted dataset contains a feature named *section* that shares the same properties, while the same can be said for the *predicted category* feature of the Amedia dataset. As the name implies, Amedia utilizes a category prediction model for the placement of articles in their respective categories. This sometimes leads to several categories for a single article. As a result of this, the predicted category of the Amedia dataset may not be directly comparable to the comparable features in the other datasets. However, since the only metric that is used on this feature is Jaccard similarity, this effect is estimated to be small.

For the Schibsted dataset, the *section* feature is similar to the *subcategory* feature of Starke et al. [49], in that it features a higher granularity than simple categories, and in most cases can be mapped to a parent category. This procedure was not deemed necessary for this thesis, but the practical approach would be to infer it from the URL of the article or to use a dictionary of the category to section mappings.

Finally, both the Amedia and Schibsted datasets contain a *tags* feature. These are manually added tags that describe the content or some of the content of the article. The usage of this feature is prevalent throughout the dataset, and it was therefore decided to include the tags as a feature in the study. In addition the *title*, *text*, and *date* features are included. The full list of features can be seen in Table 3.6.

3.2 Metrics

3.2.1 Metrics used in previous studies

As this thesis builds directly on top of the work done in [53] [49] and [48], several of the metrics used are shared with them. In this section, each of them are explained. A full list of the similarity metrics and the features they are used on can be seen in Table 3.8.

Jaccard Similarity

One of the most common and intuitive ways to measure the similarity between two items is the Jaccard Similarity. Any item that can be split into several sub-items or features can be evaluated using Jaccard Similarity. The method is simply to take the sets of unique sub-items of the items and divide the intersect of the sets on the union of the sets. The result of this equation is the Jaccard Coefficient, referred to as Jaccard Similarity in this thesis. It is expressed in equation 3.1.

$$sim_{JACC}(s,t) = \frac{|s \cap t|}{|s \cup t|} \text{ [53]} \tag{3.1}$$

The Jaccard Similarity is used to calculate the similarity of the Section and Tags features of the news articles. In addition, it is used to compare the similarity between the Named Entities in the Text feature.

Jaro-Winkler

An intuitive way to approach the problem of measuring the difference between text strings is to simply look at the character difference between the strings. *Jaro-Winkler*, *Levensthein*, *Longest Common Subsequence*, and *Kondrak's BiGram*, are all metrics that use this approach.

The *Jaro-Winkler distance* starts with the *Jaro-similarity* function, which takes into account the number of matching characters and the order in which they appear in the strings. A simplified definition can be seen in equation 3.2.

$$JARO(s,t) = \begin{cases} 0, & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s|} + \frac{m}{|t|} + \frac{m-t}{m}\right), & \text{otherwise} \end{cases} \text{ [19]} \tag{3.2}$$

where m is the number of matching characters between the two strings, t is the number of transpositions between the two strings (i.e., the number of matching characters that are not

in the same position), and $|s|$ and $|t|$ are the lengths of the two strings.

The Winkler variation extends Jaro-similarity by giving a higher score to strings where up to the first 4 characters are equal. It is defined in equation 3.3.

$$dist_{JW}(s,t) = JARO(s,t) + l \cdot p \cdot (1 - JARO(s,t)) \text{ [19]} \tag{3.3}$$

where $JARO(s,t)$ is the Jaro similarity between strings s and t, l is the length of the common prefix between the two strings (up to a maximum of 4 characters), and p is a constant scaling factor (usually 0.1). In itself, Jaro-Winkler is a distance metric, and to get the similarity metric we need to subtract the Jaro-Winkler distance from 1. The definition of the final similarity equation can be seen in equation 3.4. Jaro-Winkler similarity is used in the Title feature of the news articles.

$$sim_{JW}(s,t) = 1 - |dist_{JW}(s,t)| \text{ [53]} \tag{3.4}$$

Levenshtein

While Jaro-Winkler is a popular algorithm for measuring string similarity, it may not always be sufficient for certain use cases. One reason for this is that Jaro-Winkler only considers the number of matching characters and the number of transpositions between two strings. Levenshtein distance calculates the minimum number of insertions, deletions, and substitutions required to transform one string into another. As such it is a more comprehensive way of measuring the difference between two strings. It is defined in equation 3.5.

$$LD(s,t) = \begin{cases} \max(|s|,|t|) & \text{if } \min(|s|,|t|) = 0 \\ \min \begin{cases} \text{lev}_{s,t}(|s|-1,|t|)+1 \\ \text{lev}_{s,t}(|s|,|t|-1)+1 \\ \text{lev}_{s,t}(|s|-1,|t|-1)+[s_{|s|} \neq t_{|t|}] \end{cases} & \text{otherwise} \end{cases} \text{ [60]} \tag{3.5}$$

To use this as a similarity metric, it also needs to be normalized. Yujian and Bo have developed a normalization strategy that simply divides the Levensthein distance by the length of the longest string [60]. It is expressed in equation 3.6.

$$dist_{LV}(s,t) = \frac{LD(s,t)}{\max(|s|,|t|)} \text{ [60]} \tag{3.6}$$

Finally we subtract the normalized distance metric from 1 in order to get the similarity met-

ric. It is expressed in equation 3.7. Levenshtein similarity is used on the Title feature of the news articles.

$$sim_{LV}(s,t) = 1 - |dist_{LV}(s,t)| \text{ [53]} \tag{3.7}$$

Longest Common Subsequence

A third approach to comparing the characters in a string is to look at the number of characters that appear following each other. There are two main ways to approach this: *Longest Common Substring*, where we compare the longest continuous string of characters common in both strings that are compared, and *Longest Common Subsequence* (LCS) [2], where we compare the longest sequence of common characters that can be constructed from the ordered sequence of characters of the two strings to be compared.

LCS is typically solved using a 2-dimensional matrix where the axis is the characters of the strings to be compared. We can define a matrix C using the strings s and t as the axis. The cell $C_{i,j}$ in the matrix C represents the length of the LCS of the position s_i and t_j. We can then fill the matrix using the formula in equation 3.8

$$C_{i,j} = \begin{cases} 0 & \text{if } i = 0 \text{ or } j = 0 \\ C_{i-1,j-1} + 1 & \text{if } s_i = t_j \\ \max(C_{i,j-1}, C_{i-1,j}) & \text{if } s_i \neq t_j \end{cases} \text{ [2]} \tag{3.8}$$

The final solution for the LCS problem will be the value in cell $C_{|s|,|t|}$.

In order to use this as a similarity metric it needs to be normalized. In Trattner and Jannach [53] the metric is normalized by subtracting the LCS from the length of the longest string, and dividing it by the length of the longest string. This gives us a distance metric expressed in equation 3.9.

$$dist_{LCS} = \frac{\max(|s|,|t|) - LCS}{\max(|s|,|t|)} \text{ [53]} \tag{3.9}$$

In order to use it as a similarity metric, the distance metric needs to be subtracted from 1. The final similarity metric is expressed in equation 3.10. In this thesis, LCS is used in the Title feature of the news articles.

$$sim_{LCS}(s,t) = 1 - |dist_{LCS}(s,t)| \text{ [53]} \tag{3.10}$$

Kondrak's Bi-Gram Distance

As previously mentioned, another way to compare the string similarity can be to use *Longest Common Substring*. *Kondrak's N-gram* [30] is an adaptation of this, where we take a substring of length N and calculate the edit distance of the substring. Kondrak's Bi-Gram is the approach of dividing the string into substrings of two characters, referred to as *bigrams*, and calculating the edit distance between them, following the same approach as the Levensthtein metric.

Using N-grams, however, also introduces the problem of partial matches. To account for this, the *positional* N-gram distance metric that Kondrak developed is used. This approach simply checks if there are characters in the same position in the substrings that are being compared. The positional n-gram distance is expressed in equation 3.11.

$$d_n(\Gamma^n_{i,j}) = \frac{1}{n}\sum_{u=1}^{n} d_1(x_{i+u}, y_{j+u}) \text{ [30]} \tag{3.11}$$

Where Γ is the non-empty sets of n-grams for strings s and t, n is the length of the n-gram and i and j are the positions of the n-grams in the two strings.

For bigrams this does not differ from the *comprehensive* approach also proposed in the same paper. It is expressed in equation 3.12.

$$d_n(\Gamma^n_{i,j}) = \frac{1}{n} d_1(\Gamma^n_{i,j}) \text{ [30]} \tag{3.12}$$

In order to use it as a similarity metric Kondrak's Bigram Distance needs to be subtracted from 1. The final similarity metric is expressed in equation 3.13. In this thesis, Kondrak's Bi-Gram Distance is used in the Title feature of the news articles.

$$sim_{BI}(s,t) = 1 - |dist_{BI}(s,t)| \text{ [53]} \tag{3.13}$$

While Trattner and Jannach [53] Starke et al. [49] and Solberg [48] all employ a metric referred to as *BiGram similarity*, it was discovered during the work with this thesis that different metrics were used. Trattner and Jannach [53] used the same Kondrak's Bigram metric as used in this thesis, however in Starke et al. [49] a different bi-gram similarity metric is used. Essentially the version in Starke et al. [49] uses a metric where a set of bigrams for the string is created, and then the similarity is calculated using the Jaccard-similarity of the two sets. Because of this, the BiGram similarity used in this thesis is not comparable to Starke et al. [49] but is comparable to Trattner and Jannach [53].

Latent Dirichlet Allocation

A common way of measuring the similarity between news articles is by creating a *Topic Model* and comparing the news articles by their weights or probabilities calculated based on the topic model.

One popular model to use is the *Latent Dirichlet Allocation* (LDA). LDA is a generative probabilistic model of a corpus [7]. The particular variant of LDA used is the *online variational Bayes* algorithm put forward in [23].

LDA topic modeling works by taking a number of topics k, a number of documents N each containing a collection of words. The total number of words in the corpus is represented by n. While training the model, the following process is used:

Algorithm 1: A variational inference algorithm for LDA [7]

Initialize ϕ_0: $n_i = 1/k$ for all i and n
Initialize γ: $\gamma_i = \alpha_i + N/k$ for all i
repeat
  for $n = 1$ **to** N **do**
    for $i = 1$ **to** k **do**
      Update ϕ_{t+1}: $n_i \propto \beta_i w_n \exp(\Psi(\gamma_{t,i}))$
    Normalize ϕ_{t+1}: $\phi_{t+1,n} \leftarrow \frac{\phi_{t+1,n}}{\sum_i \phi_{t+1,i}}$
  Update γ_{t+1}: $\gamma_{t+1} = \alpha + \sum_{n=1}^{N} \phi_{t+1,n}$
until *convergence*;

In the update step, the algorithm loops over all words in the corpus, and for each word, updates the corresponding topic distribution in ϕ. The update is performed using the formula $n_i \propto \beta_i w_n \exp(\Psi(\gamma_{t,i}))$, where β_i is the probability of observing word n given topic i, w_n is the count of word n in the corpus, and Ψ is the digamma function, the first derivative of the log Gamma function [7]

Once the model is converged, it can be used to predict a vector of topic weights for each document. To maintain consistency with Trattner and Jannach [53], Solberg [48] and, Starke et al. [49], the number of topics is set to 100. As was done in the previous studies, the vectors of weights are then compared using cosine similarity to get the LDA similarity between two documents. The final similarity metric is expressed in equation 3.14. In this thesis, LDA similarity is used on the Title and Text features.

$$sim_{LDA}(s,t) = \frac{LDA(s) \cdot LDA(t)}{||LDA(s)||\,||LDA(t)||} \text{ [53]} \tag{3.14}$$

TF-IDF

Another way to compare the text of similar documents is using a *vector model*. One of the most commonly used vector models for text classification is *Term-Frequency-Inverse-Document-Frequency* (TF-IDF). Inverse document frequency was originally introduced by Luhn [36], and later coupled with Term Frequency. A survey conducted in 2015 reported that 70% of text-based recommender systems in digital libraries used TF-IDF [4].

The approach of using TF-IDF starts with counting the *term frequency*, that is the number of times a term occurs in a document, divided by the total amount of words in the document. It is expressed in equation 3.15.

$$tf(t,d) = \frac{\sum(t,d)}{|d|} \tag{3.15}$$

Next, we take the *inverse document frequency*, which is the logarithm of the total number of documents, divided by the number of documents that contain the term. It is expressed in equation 3.16.

$$idf(t,D) = log\frac{|D|}{|\{d \in D : t \in d\}|} \text{ [36]} \tag{3.16}$$

The reason we use the logarithm is to provide a heavier weight to documents that are in fewer of the documents. Finally, we multiply the term frequency with the inverse document frequency to get the TF-IDF score of the specific word for the specific document as seen in equation 3.17

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D) \tag{3.17}$$

When applying this to a corpus, each document will be assigned a vector of the length of the vocabulary in the corpus, and each component will have the TF-IDF score for one of the words in the vocabulary. For document similarity, these vectors can then be compared using cosine similarity. The final similarity metric is expressed in equation 3.18

$$sim_{TFIDF}(s,t) = \frac{TF-IDF(s) \cdot TF-IDF(t)}{||TF-IDF(s)||\,||TF-IDF(t)||} \text{ [53]} \tag{3.18}$$

In this thesis, TF-IDF is used on the Text feature of the news articles. There are two variants of TF-IDF used in Starke et al. and in Solberg. One that considers the entire document, and one that only considers the 50 first words of the document.

Image Brightness

The brightness of an image is the subjective visual perception of the energy output of a light source [46]. The average brightness can be computed by using default parameters and the NTSC weighting scheme as follows:

$$avg\_brightness = \frac{1}{N}\sum_{x,y} Y_{x,y}, \text{ with} \tag{3.19}$$

$$Y_{x,y} = (0.299 \cdot R_{x,y} + 0.587 \cdot G_{x,y} + 0.114 \cdot B_{x,y})[53].$$

In the luminance algorithm, $Y_{x,y}$ denotes the luminance value, and N is the size of the image. R, G, and B correspond to the RGB color space channels of pixels x, y [53].

Image Sharpness

The sharpness of an image can be computed by using the Laplacian L of an image, then divided by the locale average luminance (μ_{xy}) around pixel (x, y):

$$avg\_sharpness = \sum_{x,y} \frac{L(x,y)}{\mu_{xy}}, \text{ with} \tag{3.20}$$

$$L(x,y) = \frac{\delta^2 I_{xy}}{\delta x^2} + \frac{\delta^2 I_{xy}}{\delta y^2}[53]$$

where I_{xy} denotes the intensity of a pixel [53]

Image Contrast

The intensity of each pixel in an image can be used to compute the relative difference luminance, i.e. the contrast. The root-mean-square contrast (RMS contrast) approach is defined as follows:

$$avg\_contrast = \frac{1}{N}\sum_{x,y}(I_{xy} - \bar{I})[53] \tag{3.21}$$

I_{xy} denotes the intensity of a pixel, $\bar{I}$ the arithmetic mean of the pixel intensity, and N the number of pixels [53].

Image Colorfulness

The colorfulness of an image can be computed by using the individual color distance of the pixels in an image [46]. To do this, the image needs to be transferred to an sRGB color space defined as $rg_{xy} = R_{xy} - G_{xy}$ and $yb_{xy} = 1/2(R_{xy} + G_{xy}) - B_{xy}$ where R_{xy}, G_{xy} and B_{xy} the color channels of the pixels, and subsequently measure colorfulness, as follows:

$$COL = \sigma_{rgyb} + 0.3 \cdot \mu_{rgyb}, \text{ with} \tag{3.22}$$
$$\sigma_{rgyb} = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2}, \mu_{rgyb} = \sqrt{\mu_{rg}^2 + \mu_{yb}^2} [53]$$

where σ and μ stand for the standard deviation and the arithmetic mean, and 0.3 is a predefined parameter in OpenIMAJ [53].

Image Entropy

The entropy of an image can be described as the amount of information observed. In this work, the Shannon entropy is used to compare two images. First, the images are converted to grayscale, resulting in each pixel containing exactly one intensity value. Second, the occurrence of each distinct value is counted. The entropy can then be computed as follows:

$$avg\_entropy = -\sum_{x \epsilon |0..255|} p_x \cdot log_2(p_x) [53] \tag{3.23}$$

Here, p_x denotes the probability of finding the gray-scale value x among all pixels in the image [53].

In order to use the low-level image features above as a similarity metric, the Manhattan distance is used. This gives us a final similarity metric as expressed in equation 3.24. As with [53] and [49], the low-level image features are extracted using the OpenIMAJ ilbary[7] as proposed by San Pedro and Siersdorfer [46] [53].

$$sim_{IM}(s,t) = 1 - |IM(s) - IM(t)| \text{ [53]} \tag{3.24}$$

Image Embeddings

In addition to the low-level features, *image embeddings* were extracted from the images. The embeddings were extracted using a pre-trained (ImageNet) VGG-16 model, identically

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

to Solberg [48], Starke et al. [49], and Trattner and Jannach [53]. Similar models have also been used in several other recommendation scenarios such as Eksombatchai et al. [17] and Messina et al. [39]. As with Trattner and Jannach [53] and Starke et al. [49], the first fully connected layer is used as the output. The first fully connected layer of the VGG-16 model features a 4096-element vector embedding [47]. The vectors are then compared using cosine similarity. Using the Keras[8] framework for the computations, the images were all automatically downsampled to fit the input layers [53]. The final similarity metric is expressed in equation 3.25.

$$sim_{EMB}(s,t) = \frac{EMB(s) \cdot EMB(t)}{||EMB(s)||\,||EMB(t)||} \text{ [53]} \tag{3.25}$$

Days Distance

The days distance is a simple metric that calculates a similarity score between two news articles based on their publication dates. It takes the absolute difference of days between two articles and divides this difference by the maximum difference possible across the dataset. This gives us the relative distance between the days which is then subtracted from 1 in order to get a similarity metric. It is expressed in equation 3.26:

$$sim_{DAYS}(s,t) = 1 - \left| \frac{s_d - t_d}{max(D) - min(D)} \right| \tag{3.26}$$

where s_d and t_d are the publication days of articles s and t, and D is the set of publication days in the dataset.

3.2.2 Additional metrics evaluated in this thesis

In addition to the metrics above, which were all implemented and evaluated in Starke et al. [49], some additional metrics are evaluated in this thesis. In Starke et al. [49] and Solberg [48] it was found that the Text and Title features were the most representative of human similarity judgments. The metrics introduced are therefore aimed at further exploring ways to measure similarity based on these features.

Because of the popularity of BERT-based models in recent works, two BERT-based metrics are introduced: One utilizing Sentence Transformers (SBERT) embeddings, and one utilizing BERTopic topic modeling. In Starke et al. [49] the metric that best represented human similarity judgments was TF-IDF used on the body text feature. In Starke et al. [49] the words are

[8] https://keras.io/

stemmed using a *snowball stemmer*, however, in some cases stemming words using lemmatization techniques may provide better results [29]. Because of this, a metric using TF-IDF with lemmatized tokens is used.

In Solberg [48] a pre-study was done in order to explore possible differences in human similarity judgments between the Recent Events and Sports categories. In the pre-study, it was suggested that Named Entities could be relevant in the Sports category. While the findings in the pre-study were not supported in the survey to collect human similarity judgments, Named-Entities may still be worth exploring. I therefore employ a strategy to automatically extract the named entities from the Text feature.

SBERT

Sentence Transformers, commonly referred to as Sentence-BERT or SBERT, is a modification of a pre-trained BERT network that uses siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity [44]. As the name implies, SBERT is intended to extract embeddings from shorter texts and the transformer design cannot extract embeddings from texts that are longer than 512 tokens. Because of this, only the 512 first words in each article text are used to extract the embeddings. This is slightly lower than the word count median of the articles, and the embeddings will therefore not be representative of the semantic contents of the entire article text for a majority of the article texts. Nonetheless, the first 512 tokens should still be sufficient to extract sufficient semantic information to produce comparable embedding vectors.

The SBERT model used is the *nb-sbert-base*[9] model, trained by The National Library of Norway (NLN) for the Norwegian Language. It is built upon the *nb-bert-base* [31] model using a machine-translated version of the Multi-Genre Natural Language Inference (MNLI) [56] dataset. In the training of nb-sbert-base NLI triplets were used as suggested by Reimers and Gurevych [44]. The nb-bert-base model, that nb-sbert-base is trained on top on, follows the same method as put forward by Vaswani et al. [55]. It is trained using the Norwegian Colossal Corpus [31] also maintained by NLN, featuring a large collection of primarily Norwegian texts dating back over 200 years, totaling over 18 billion tokens.

When SBERT with the model *nb-sbert-base* is used to encode text, it returns a 768-dimensional dense vector of values from -1 to 1. Each of these dimensions represents an evaluation of the document against an unknown semantic factor that has been learned during the training of the model. We can then perform cosine similarity compare vectors of pairs of documents to find a similarity score. This gives us a similarity metric as expressed in equation 3.27. SBERT

[9] https://huggingface.co/NbAiLab/nb-sbert-base

similarity is used on the Title and Text features of the news articles.

$$sim_{SBERT}(s,t) = \frac{SBERT(s) \cdot SBERT(t)}{||SBERT(s)||\,||SBERT(t)||} \tag{3.27}$$

BERTopic

While using cosine similarity directly on the SBERT embeddings is a simple way of finding the similarity between documents, another approach is to use the embeddings as part of a topic modeling process. BERTopic [21] is a method to train a Topic Model using BERT embeddings. It is highly modular, and the default setup, which is also used in the survey, takes SBERT embeddings as the input. It then performs a dimensionality reduction on the embeddings using UMAP [38] and clustering using HDBSCAN [9] before performing a class-based TF-IDF method on the clusters to assign words to the topics. A illustration of the BERTopic default setup can be seen in Figure 3.2.

This *c-TF-IDF* method is a slight variation of the TF-IDF method mentioned in the previous subsection. Here the inverse document frequency is replaced by the inverse class frequency to measure how much information a term provides to a class. It is calculated by taking the logarithm of the average number of words per class A divided by the frequency of term t across all classes [21]. The formal expression can be seen in equation 3.28.

$$W_{t,c} = tf_{t,c} \cdot log(1 + \frac{A}{tf_t})[21] \tag{3.28}$$

Once the topic model has been trained, it can be used to find the probability of a document belonging to a specific topic. By default, BERTopic only returns the probability of the assigned topic of the document, but it also provides the option of calculating the probabilities for all topics. A vector of these probabilities can then be compared using cosine similarity to find the similarity of the two documents. This gives us a similarity metric as expressed in equation 3.29.BERTopic similarity is used on the Title and Text features of the news articles.

$$sim_{BERTopic}(s,t) = \frac{BERTopic(s) \cdot BERTopic(t)}{||BERTopic(s)||\,||BERTopic(t)||} \tag{3.29}$$

Lemmatized TF-IDF

When using TF-IDF, what we are doing is comparing the specific words as they appear in the text. However, using TF-IDF without any pre-processing leads to situations where different inflections of the same word will be handled as completely separate words. Because of this,

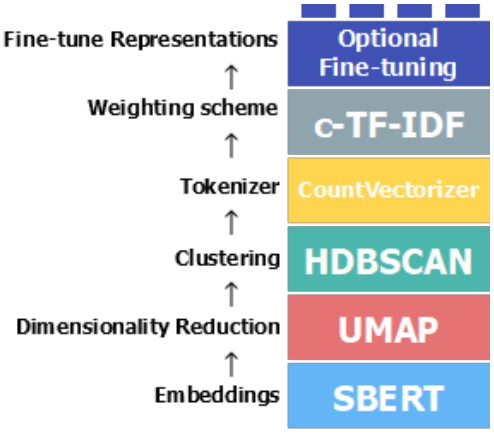

Figure 3.2: Default BERTopic setup [21]

it is common to use morphological techniques in order to normalize words that are based on the same root word [1]. There are two main approaches to do the normalization: *stemming* and *lemmatization*.

Stemming is the raw heuristic process of removing the end of words, in the hope to remove common morphological and inflectional endings from the words [6]. *Lemmatization* is the process of reducing a word to its lemma. A words lemma is the canonical form, dictionary form, or citation form of a lexeme. A lexeme is a unit of lexical meaning that exists regardless of the number of inflectional endings it may have or the number of words it may contain. The lemmas of a dictionary are all lexemes [6].

In Starke et al. [49] a stemming method using the Snowball Algorithm [43] is used. The Snowball stemmer aims to remove *regions* at the end of a word. The remaining word after the region(s) are removed, is the word *stem* For Norwegian the stemmer has one region definition, R1: *R1 is the region after the first non-vowel following a vowel, or is the null region at the end of the word if there is no such non-vowel. But then R1 is adjusted so that the region before it contains at least three letters.*[10]

While stemming increases the performance of TF-IDF in information retrieval, lemmatized TF-IDF has been shown to surpass the performance of stemmed TF-IDF [3]. When TF-IDF was also found to be the best-performing metric in Starke et al. [49], it could be interesting to see if this performance can be improved using lemmatized TF-IDF. Because of this a lemmatized TF-IDF metric utilizing the lemmatizer for Norwegian Language implemented with SpaCy[11] was used. For practical reasons the model *nb_core_news_lg*[12] was used due to its higher performance on Named Entity Recognition.

nb_core_news_lg features a rule-based lemmatizer with lemmas extracted from Norsk Ord-

[10] https://www.nltk.org/_modules/nltk/stem/snowball.html
[11] https://spacy.io/
[12] https://spacy.io/models/nb#nb_core_news_lg

bank[13]. The lemmatization done on top of a Part-Of-Speech tagger that follows the guidelines of Universal Dependencies [40]. In addition to using a lemmatization approach, a further reduction of the TF-IDF vectorizers vocabulary is performed by removing words that have a document frequency of above 99% or below 0.1%. This is to remove corpus-specific words that would otherwise be considered stop-words, as well as words that are likely to be misspellings or other formatting errors.

The TF-IDF equation itself does not change and still follows the definition in 3.17. As with the other TF-IDF metrics, cosine similarity is used on the TF-IDF vectors of two documents to find their similarity, and the similarity equation is expressed in equation 3.18. The Lemmatized TF-IDF metric is used on the Title and Text features.

Named Entities

In Solberg [48] a pre-study was conducted that showed that Named Entities could in some situations be a factor in human similarity judgments of news articles. In his thesis i smaller dataset was used and named entities were manually extracted for each article. In this thesis, this is not viable due to the large dataset. However, an automated Named Entity Recognition (NER) method is used. For this *nb_core_news_lg* is again utilized, using the SpaCy pipeline.

The NER model in nb_core_news_lg is trained on the NorNe dataset [25]. Released in 2019, NorNe is the first public dataset for named entity recognition for Norwegian. It initially provided NER annotations on top of the Norwegian Dependency Treebank (NDT) but has since been translated into Universal Dependencies along with the NDT [40]. The dataset features approx. 300 000 tokens for bokmål and nynorsk respectivly. For bokmål, the dataset features 16 309 sentences with 14 369 named entities. The named entities are divided into the following types: *Person*, *organization*, *location*, *product*, *derived* and *geo-political* entities. Geo-political entities are further split into two subtypes with a *locative* sense and an *organization* sense. See Table 3.7 for definitions of the types.

When using the NER module of nb_core_news_lg to extract entities, it returns a tuple of an entity and its label. In order to compare the similarity between the documents the entities are extracted and put into a set for each document. These sets are then compared using Jaccard Similarity. The final similarity metric is expressed in equation 3.30. The metric is used on the Text feature.

$$sim_{NENTS}(s,t) = \frac{|NENTS(s) \cap NENTS(t)|}{|NENTS(s) \cup NENTS(t)|} \tag{3.30}$$

[13] https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-5/

Table 3.7: Definitions of select named entity labels. [50]

Entities	Label	Explanation
Person	PER	Real of fictional characters and animals
Organization	ORG	Any collection of people, such as firms, institutions, and organizations.
Location	LOC	Places, buildings, facilities, etc
Geo-political entity	GPE	Geographical regions defined by political and/or social groups
	GPE_LOC	GPE with a locative sense
	GPE_ORG	GPE with an organization sense
Product	PROD	Artificially produced entities are regarded as products
Event	EVT	Festivals, cultural events, sports events, weather phenomena, wars, etc
Derived	DRV	Words that are derived from a name, but are not a name in themselves
Miscellaneous	MISC	Other named entities

3.3 Human Similarity Judgments

3.3.1 Survey design

The main survey follows a similar design to that in Trattner and Jannach [53], Starke et al. [49], and Solberg [48]. The main task of the participant is to rate the similarity between two items, in our case a pair of news articles, on a 5-point Likert scale. As in Starke et al. [49] and Solberg [48], the users were also asked about their familiarity with the articles and their confidence in their similarity rating. The phrasing of the familiarity question was changed a little in the translation to Norwegian. Where the familiarity question in the previous studies was phrased as *How familiar are you with Article n*, in this study it was changed to *How familiar are you with the subject in Article n*. This was done so as to not give the user the impression that they were being tested in their ability to gain familiarity with the article itself, but rather a question about their familiarity with the article's general subject.

Before starting the survey, the participants were instructed that they would be asked to rate 10 random pairs of news articles according to the mentioned questions. They were also instructed that they did not need to read the entire articles but were expected to form a general impression of the similarity between the articles. They were asked to ignore potential formatting errors, and that there would be an attention check during the survey. The instructions also included that the survey should take 5-10 minutes, in line with the time spent on the surveys in the previous studies. This was also done in order to help participants understand that fully reading the articles was not required of them.

Table 3.8: Full list of similarity metrics and the features they are applied to. Metrics used in [53] or [49] are denoted by *

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

Figure 3.3: Presentation of articles in the survey

When the participant started the survey, they were first randomly assigned to a group of either Amedia context or Schibsted context. This was done because of the minor differences between the datasets. Once assigned, 10 article pairs were selected. 5 from the local publication and 5 from the national publication. Each pair belong to a specific sample group outlined in section 3.3.2 below. A screenshot of the presentation of a pair can be seen in Figure 3.3. As in the previous work, the participants were subjected to an attention check. In this survey, the attention check was set to a random pair rating between steps 4 and 8.

The final part of the survey is a questionnaire to collect some basic demographics as well as the participant's usage of information cues. The demographics collected consist of age, gender, news usage, and location. The location question is phrased to measure their location with reference to Bergen, and encompasses the choices: *Bergen Municipality*, *Bergen Area*, *Vestland County*, *Norway* and *Outside Norway*. The Bergen Area was specified as a set of

municipalities surrounding Bergen, excluding Bergen Municipality itself[14].

The information cues are considered the article features displayed to the user. The users were asked to rate their usage of information cues using a 5-point Likert scale. The participants were asked to give their usage evaluation of the following information cues: *Category*, *Title*, *Image*, *Date*, *Text* and *Tags*.

3.3.2 Sampling strategy

To construct the set of pairs for the survey a similar strategy as to Trattner and Jannach [53] and Starke et al. [49] was used. As mentioned in section 3.1 the dataset is split into separate subsets for each publication. On each of these subsets, the 25 metrics listed in Table 3.8 were calculated. The scores were then calculated using equal weights. The result of this was 4 matrices of similarity scores for all articles in each of the publications.

A challenge in Starke et al. [49] and Solberg [48] is that both the metrics and the participants in the surveys rate the similarity of the articles in the pairs as low. This effect is likely to be further exaggerated in this study as we are looking at all articles across all categories, instead of looking at specific sections. Because of this, the biased stratified sampling strategy used in Trattner and Jannach [53] and Starke et al. [49] was not used. Instead, a strategy utilizing the standard deviation of the pairwise similarity scores was used. Essentially the strategy can be considered to be biased towards the tails of the pairwise similarity score distribution.

The strategy works by calculating the standard deviation of the pairwise similarity scores and then dividing the pairs into the following sample groups:

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

The results of applying this grouping strategy to the pairwise similarity scores of each of the publications can be seen in Table 3.9. The pairwise similarity scores and their distributions can be seen in Figure 3.4.

[14]Specifically: Samnanger, Bjørnafjorden, Austevoll, Øygarden, Askøy, Vaksdal, Modalen, Osterøy, Alver, Austreim, Fedje or Masfjorden municipalities.

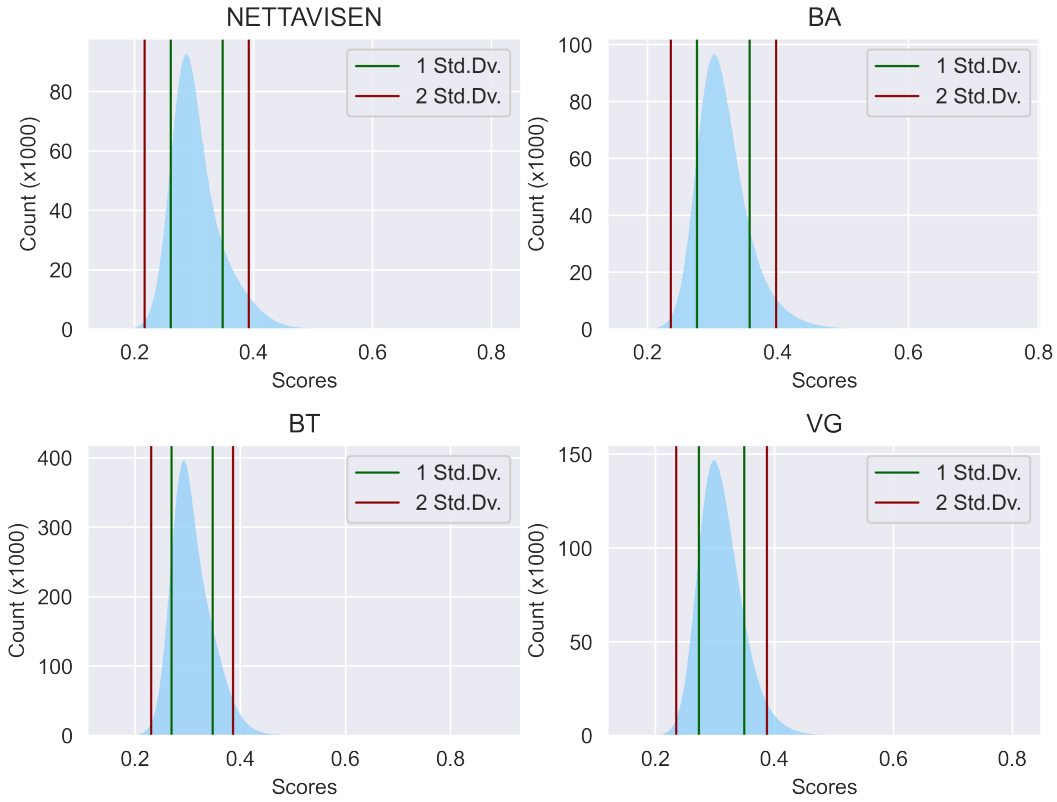

Figure 3.4: Mean pair rating distributions with standard deviations marked in order to show the standard-deviation group-based segmentation.

Once the scores were divided into groups, 1 000 pairs were randomly sampled from each group for each publication and added to the survey database. This resulted in 5 000 pairs for each publication and 20 000 pairs available in total. In the survey, the participant is presented with either one pair from each of the sampling groups from BA and Nettavisen, or from BT and VG.

3.3.3 Participant recruitment

As mentioned in section 3.1, a challenge early identified with this thesis would be to recruit participants for the survey. While Trattner and Jannach [53], Starke et al. [49] and Solberg [48] all used crowd-working platforms for participant recruitment, the popular crowd-working platforms do not have a large number of participants from Norway and especially specific areas of Norway. This was the main reason why Bergen was chosen as the local domain for this thesis, as it was expected it would ease participant recruitment.

Several options for local participant recruitment were examined, primarily commercial sur-

Table 3.9: Amount of pairs and percentages per sample group. Group 1 is least similar and group 5 is most similar

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

vey platforms and academic research panels. These options were either too costly or had too long waiting times for the scope of this thesis. In the end, a Snowball recruitment strategy applied to social media [32] was chosen. All participants who completed the survey were offered to partake in a drawing of 500 NOK gift cards, one gift card was drawn for every 25 participants, bringing the expected return of completing the survey to 20 NOK.

The URL to the online survey was shared across several social media platforms with a text informing about the main task as well as the availability of a reward. The primary sharing was done through Twitter, Instagram, several Facebook groups, select Discord channels and the /r/Bergen subreddit on Reddit. Users were encouraged to share the link with the information that a higher number of participants would result in more gift-card drawings.

3.3.4 Participants and Pairs

In total 329 participants started the survey with 143 completions. The low completion number is assumed to stem from issues related to a clunky mobile interface which was later improved[15]. 2 of the participants were below 18 years old and were removed from the results, bringing the total number of participants to 141. 73 of the participants completed the Schibsted context, giving ratings to BT and VG, while 68 completed the Amedia context, giving ratings to BA and Nettavisen.

119 of the 141, or 84.4%, passed the attention check, which compared to the previous work using crowd-working platforms is quite high. After accounting for the attention check, a total of 1071 pair ratings were available from users who passed the attention check. The final figures for all segmentations of participants and pairs can be seen in Table 3.10. Throughout the rest of the section, only the pariticipants and pairs that passed the attention check will be used. In addition, the pairs that had the attention check are removed.

[15]Initially, the survey required mobile users to scroll through the entire articles in order to get to the rating schema. This was mitigated by adding article scroll toggle buttons.

Table 3.10: Segmentations of the participants and pairs for the analysis in the chapter. The pairs in the *pass* groups include the removal of the attention check ratings. Participants are divided into *Local* and *National* groups depending on their reported place of residence. *Bergen* and *Bergen Area* are considered *Local*.

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

3.3.5 Demographics

The main research question of this thesis is to evaluate how local readers rate the similarity of news at the local and national levels. It was therefore important to recruit local participants for the thesis. The recruitment strategy proved successful in this with a total of 95 participants stating their place of residence as *Bergen* and another 13 stating their place of residence as the *Bergen Area*. This gave a total amount of 108 participants that are considered as *Local*. The remaining 33 participants are considered *National*. The choices included an option *Outside Norway* of which no participants stated as their place of residence.

When selecting which region to consider as *Local* the main question was whether to include *Bergen Area* and *Vestland County*. Both of these options were put in to serve as a potential limitation for *Local*, should the participants from Bergen end up as low. However the opposite ended up being the case. Considering the amount of *Bergen* participants in the survey it would be more feasible to limit the *Local* participant group to *Bergen* only, however, the *Bergen Area* is so closely connected to Bergen[16] that excluding it from *Local* is likely to provide problems when comparing ratings of *Local* and *National* users.

A total of 112 participants, 79.4%, reported their frequency of news reading to be *approximately every day*. This is higher than in the previous work, and somewhat higher than expected. The previous work also included a question about how many days per week the participant read online news, but this was excluded from this survey to avoid confusion related to the included online news read frequency question. Charts showing the results from the demographic questions can be seen in Figure 3.5.

[16]Most of the municipalities in the *Bergen Area* are incorporated in the *Bergen Residential and Labor Market Region*. These regions are descriptive and based on travel time and commuting percentages. https://www.regjeringen.no/contentassets/735944a205424d14afef809bc039d76b/inndeling_ba-regioner_2020.pdf - Table 4.56 (Norwegian).

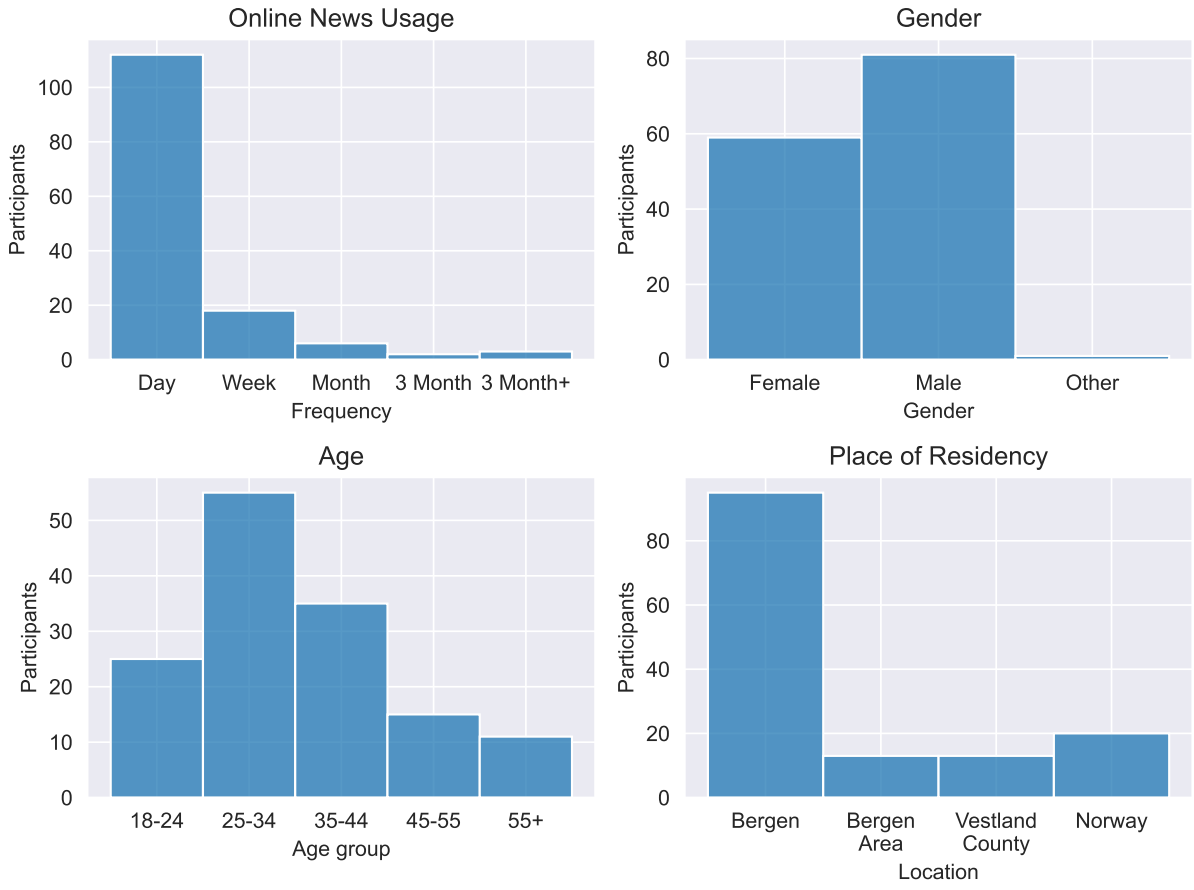

Figure 3.5: Participant demographic survey results

3.4 Methods of analysis

To answer some of the research questions, the following statistical analysis is done.

RQ1: The participants are asked for their usage of information cues (features). The mean and standard errors for the answers are calculated, a pairwise one-way Analysis of Variance (ANOVA) is then performed, and a Tukey's Honestly Significant Difference (HSD) is performed in order to see if the various cue usages are statistically different from each other.

RQ2: To investigate which feature-specific similarity metrics best represent human judgments, the scores of the various metrics are analyzed against the human similarity judgments by calculating the Spearman's rank correlation coefficients between them. This analysis is done using 6 subsets of the ratings. The *All* correlations are calculated using all ratings. The *National* correlations are calculated using the ratings for *VG* and *Nettavisen*. The *Local* correlations are calculated using the ratings for *BT* and *BA*. In addition, correlations are calculated for each publication individually.

RQ4.1: In the survey, participants rate 5 pairs from a local publication and 5 pairs from a national publication. For each of the publications, one pair is randomly selected from each of the sampling groups listed in Table 3.9. This is done in order to perform a paired t-test on the result, to see if there are differences in the similarity ratings on the national and local levels.

RQ4.2: For the investigation of the possible effects that differences in similarity ratings have on the correlations of similarity judgments and feature-specific metrics, the correlations are transformed to Z-values using Fishers-r-to-z, and then a Z-test is performed between the Z-values in order to find changes between the correlations between the feature-specific metrics and the human similarity judgments on the local and national domains.

Chapter 4

Results

In this chapter, I describe the results of the analysis of the data. It is divided into three main sections. Section 4.1 describes the results of the analysis of the reported information cue usage. In section 4.2 the results of the correlation analysis between the feature-specific metrics and human similarity (RQ2) is presented, as well as the comparison of the results to other domains (RQ3). Finally, in section 4.3 I show the results of the analysis of differences across the national and local domains (RQ4).

4.1 Information Cue Usage (RQ 1)

In order to answer RQ1 the participants were asked to state to what extent they used the different information cues when evaluating the similarity of the news articles on a scale from 1 to 5, where 1 was the lowest and 5 was the highest. Figure 4.1 shows the results. **A** shows the means and standard errors of the reported usage, while **B** shows the result of a one-way ANOVA and a Tukey HSD post hoc test. The test shows that all the differences in ratings are statistically significant, although only slightly for the Section-Text pair, which also are the two most used cues, with a mean reported usage of 4.3 for Text and 4.1 for Section. In addition, the Title has a reported usage of 3.78, Image has 3.08 and Tags has 2.84. Date is the least used information cue, with a value of 1.90.

The findings for Text and Title are consistent with Starke et al. [49] and Solberg [48], however the result for Section is above that of the previous work. This is likely due to using all sections or categories in this study, while subsections were used in the previous work. Image also shows similar usage as the previous work. While the previous studies showed a low usage of Date as an information cue, neither of the studies showed a result as low as shown by this study.

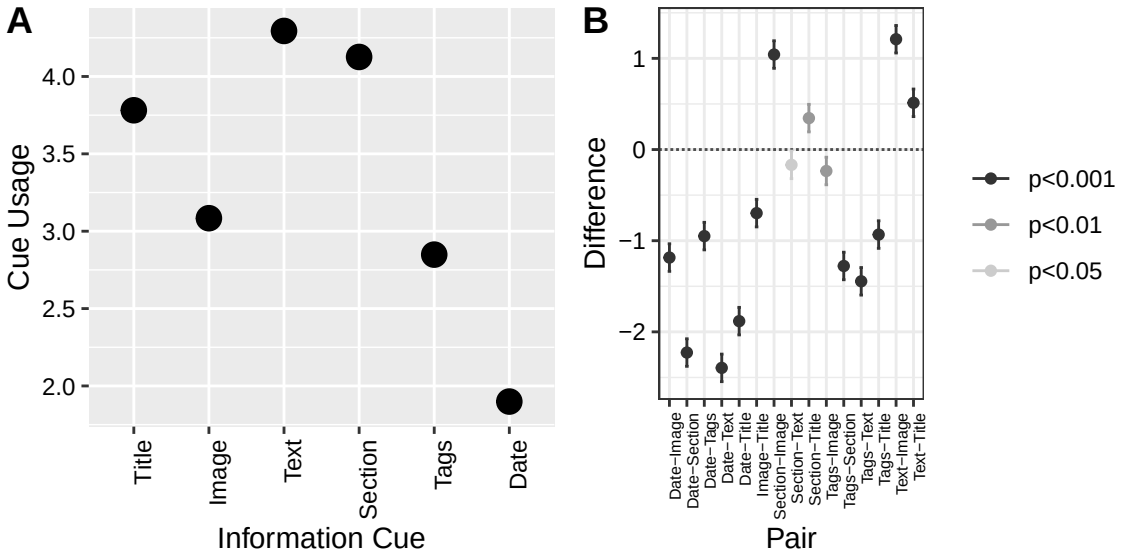

Figure 4.1: **A**: Information cue usage (means and std. errors). **B**: Tukey HSD Pairwise comparison after performing one-way ANOVA.

4.2 Evaluating Feature-Specific Metrics

4.2.1 Comparing Metrics to Human Judgments (RQ 2)

As with Trattner and Jannach [53], Starke et al. [49] and Solberg [48], one of the main questions for this thesis is how well *Feature-Specific Similarity Metrics* compare to *Human Similarity Judgments*. For this thesis, the scope is the *Norwegian News domain*. In order to compare the Similarity Metrics to the Similarity Judgments, Spearman correlations have been calculated between the metrics listed in Section 3.2 and the Human Similarity Judgments collected through the survey.

In this thesis, we are also looking at differences between the Local and National news domains, by using four different publications. Because of this, the correlations are calculated for several various subsets of data. In addition to the full list of pair ratings, it is calculated for the *National* and *Local* domains, as well as for *VG*, *BT*, *Nettavisen*, and *BA*. The *National* domain correlations are calculated using the pair ratings for *VG* and *Nettavisen*, while the *Local* domain is calculated using pair ratings for *BT* and *BA*. The various publication correlations are calculated using only the pair ratings for the specific publications. The amount of pairs can be seen in Table 3.10, while the full results of these calculations can be seen in Table 4.1

Text Based Metrics

In the current study, the *Text:SBERT* metric (0.60) presented the highest correlation across all divisions of the dataset. This outperformed the *Text:TF-IDF* metric (0.47), which was the

highest correlating metric in both the Starke et al. [49] (0.29) and Solberg [48] (0.53) studies. Furthermore, the *Text:TF-IDF-L* metric displayed high correlations (0.47), similar to the non-lemmatized TF-IDF metric.

The *Text:LDA* metric in this study (0.29) had a higher result than in Solberg [48] (0.17) and significantly higher than in Starke et al. [49] (0.03). The publications with larger datasets showed higher correlations with the *Text:LDA* metric, specifically VG (0.34) and BT (0.33), compared to those with smaller datasets like Nettavisen (0.29) and BA (0.26).

The *Text:BERTopic* metric (0.40) outperformed the *Text:LDA* metric (0.29). However, its score was lower than *Text:SBERT* (0.60) and *TF-IDF* (0.47). Finally, the *Text:NENTS* metric (0.21) had a wide range of correlations depending on the publication, with VG showing the lowest correlation (0.12) and Nettavisen the highest (0.36).

Title Based Metrics

The *Title:SBERT* metric demonstrated the highest correlation (0.38) among Title-based metrics, followed by *Title:BERTopic* (0.30). For *Title:SBERT*, BT showed a considerably higher score (0.45) than BA (0.33). Conversely, *Title:BERTopic* presented a higher score for BA (0.43) than for BT (0.24). *Title:TF-IDF* correlated similarly with *Title:BERTopic* but the variations were small. The *Title:LDA* metric displayed very low scores (0.07), with the exception of BT, which showed a slightly higher correlation (0.2).

Title:TF-IDF-L (0.17) showed slightly lower correlations than *Title:TF-IDF* (0.20). The Title-based edit distance metrics, *Title:BI*, *Title:JW*, *Title:LCS* and *Title:LV*, exhibited similar correlations of 0.18, 0.21, 0.22, and 0.18, respectively. Among these, *Title:LCS* showed a higher correlation with Nettavisen (0.35) and a lower one with BA (0.1).

Image Based Metrics

Among the Image-based metrics, *Image:EMB* demonstrated the highest correlation to Human Similarity Judgments, registering a correlation of 0.30. This correlation was especially high for Nettavisen (0.46). In terms of similarity metrics using low-level features, *Image:BR*, *Image:SH*, and *Image:EN* showed similar correlations of 0.24, 0.26, and 0.22, respectively. However, in the VG dataset, these metrics all demonstrated correlations too low to be statistically significant.

The remaining Image-based metrics, *Image:CO* and *Image:COL*, presented lower correlations across the datasets. Specifically, *Image:CO* displayed a correlation of 0.13 for all subsets and a correlation of 0.15 for the Local domain. For Nettavisen, this metric showed a corre-

Table 4.1: Similarity metric correlation (Spearman) with human similarity judgments. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Metrics are listed in the left column, with Spearman correlations for the various divisions of the datasets listed in the other columns. *All* combines the pair ratings of all publications. *National* combines VG & Nettavisen, *Local* combines BT & BA

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

lation of 0.10. On the other hand, *Image:COL*, with an overall correlation of 0.07, was only significant when considering all datasets.

Section, Tags and Time

Section:JACC showed high correlations of 0.49. The correlations were particularly high for the Amedia publications, with 0.59 for BA and 0.62 for Nettavisen, compared to lower correlations observed for the Schibsted publications, specifically 0.36 for VG. *Tags:JACC* demonstrated a correlation of 0.33 when considering all responses. A discrepancy was observed between the Schibsted publications, VG and BT (both 0.25), and the Amedia publications, Nettavisen (0.45) and BA (0.42). Lastly, the *Time:Days* metric revealed a correlation of 0.22 when compared with all similarity ratings.

4.2.2 Comparing Correlations to Other Domains (RQ 3)

In the previous subsection, some of the correlations were compared to the previous work done in the news domains. In this subsection, the correlations will be compared to the Movies and Recipe domains as well. All comparable correlations are listed in Table 4.2.

Several of the low-level image features, specifically *Image:BR*, *Image:SH* and *Image:EN*, have a considerably higher correlation to the human similarity judgments across the Norwegian news domains, than what we see from the WPO domain [49]. However, when comparing with the Recipe and Movie domain [53], they are more similar. The *Image:EMB* metric also have a higher correlation in the Norwegian news domains than the two other news domains. It does not surpass the correlation in the recipe domain, however.

The text metrics are performed somewhat differently across the domains. For the news domain it is calculated using the main text of the article, in the recipe domain, the directions are used for the calculations, while in the movie domain, the movie plot is used. Overall, we see slightly higher results for the common text-based metrics across the Norwegian news domains than the other news domains, aside from *Text:TF-IDF* where the UK news domain has a higher correlation. However, the observed difference is likely too small to be statistically significant. Compared to the Recipe and Movie domains, we see that *Text:TF-IDF* have correlations similar to the Recipe domain, but somewhat higher than the Movie domain. For *Text:LDA* the correlations in the Norwegian domains are significantly higher than other news domains, but lower than the Movie domain and significantly lower than in the Recipe domain.

The time-based metric shows a higher correlation in the Norwegian news domains than in

Table 4.2: Metric correlations for following domains: Norway (All), National (VG, Nettavisen), Local (BT, BA), UK (Guardian) [48], WPO (Washington Post) [49], Recipe [53] and Movie [53]. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. p-values were not available for UK domain. All correlations are calculated based on ratings from users passing the attention check.

Function	**Domain**						
	Norw.	**Nati.**	**Loca.**	**UK [48]**	**WPO [49]**	**Rec. [53]**	**Mov. [53]**
Image:BR	0.24***	0.16***	0.32***	-	0.10***	0.18**	0.22***
Image:SH	0.26***	0.24***	0.28***	-	0.06**	0.16*	0.10***
Image:CO	0.13***	0.11*	0.15***	-	0.05*	0.29***	0.03
Image:COL	0.07*	0.07	0.08	-	0.11	0.05*	0.15***
Image:EN	0.22***	0.15***	0.28***	-	0.07**	0.34***	0.15***
Image:EMB	0.30***	0.39***	0.23***	0.12	0.17***	0.44***	0.18***
Text:LDA	0.29***	0.29***	0.29***	0.17	0.03	0.54***	0.37***
Text:TF-IDF	0.47***	0.45***	0.48***	0.53	0.29***	0.50***	0.35***
Text:TF-IDF-50	0.17***	0.14**	0.2***	-	0.14***	-	-
Time:Days	0.22***	0.20***	0.24***	0.03	0.09***	-	0.37***
Section:JACC	0.49***	0.47***	0.50***	0.44	0.58***	-	0.56***
Title:LDA	0.07*	0.04	0.10	-0.03	0.02	0.22***	0.01
Title:BI	0.18***	0.19***	0.16***	0.30	0.08**	0.48***	0.17***
Title:JW	0.21***	0.2***	0.21***	0.13	0.05*	0.46***	0.16***
Title:LCS	0.22***	0.27***	0.17***	0.18	0.08***	0.50***	0.20***
Title:LV	0.18***	0.19***	0.16***	0.10	0.06**	0.48***	0.19***

the other news domains. However, it is lower than in the movie domain. In the movie domain, the metric is based on the release date of the movie. The *Section:Jaccard* shows relatively high correlations across all domains. They are slightly lower in the Norwegian news domains, but not likely to be statistically significant. It is important to note that the metric is calculated on slightly different data. In the Norwegian News domains, it is calculated on either the main category or the subsection, depending on the publication. For all 3 domains that are compared here, the data is mixed. In the UK and WPO domains, it is calculated using the subsections. While in the movie domain, it is calculated on the genre of the movie. As such it is difficult to make direct comparisons, but we see that the correlations are fairly similar.

Only in the Recipe domain does the *Title:LDA* metric provide a correlation that is statistically significant. In the other domains, it is all performing very low. The remaining Title-based metrics all have some positive correlations. For the recipe domains, the correlations are relatively high, while for the WPO domain, the correlation is low and barely significant.

While the movie domain also used tags, it is not included as the specific metrics used in the movie domain and the study in this thesis differed too mcuh.

4.3 The National and Local Domains

The main research question of this thesis is to explore if there are differences between the *National* and *Local* Norwegian news domains, and if these differences are large enough to warrant different approaches to similar item recommendation.

4.3.1 Comparing Local and National Similarity Ratings (RQ 4.1)

In order to compare the National and Local participants' similarity ratings of the National and Local publications, violin plots were made to look for potential differences, the plots can be seen in Figure 4.2. Looking at them, it becomes immediately apparent that there are no big differences between the various divisions. Some key findings can be seen. Most importantly: The overall ratings do not follow a normal distribution. From the plots, it can be seen that the participants have avoided giving a similarity rating of "neither nor" which is mapped to the similarity rating of "3", and instead selected choices of negative or positive similarity. This can potentially cause challenges in what methods to use for the analysis.

More pertinent to the research question, it can also be seen that the median for the local publications (left column) group 4 is at a level higher than the national publications (right column) group 4 across the plots. In addition, it can be seen that the median rating from the

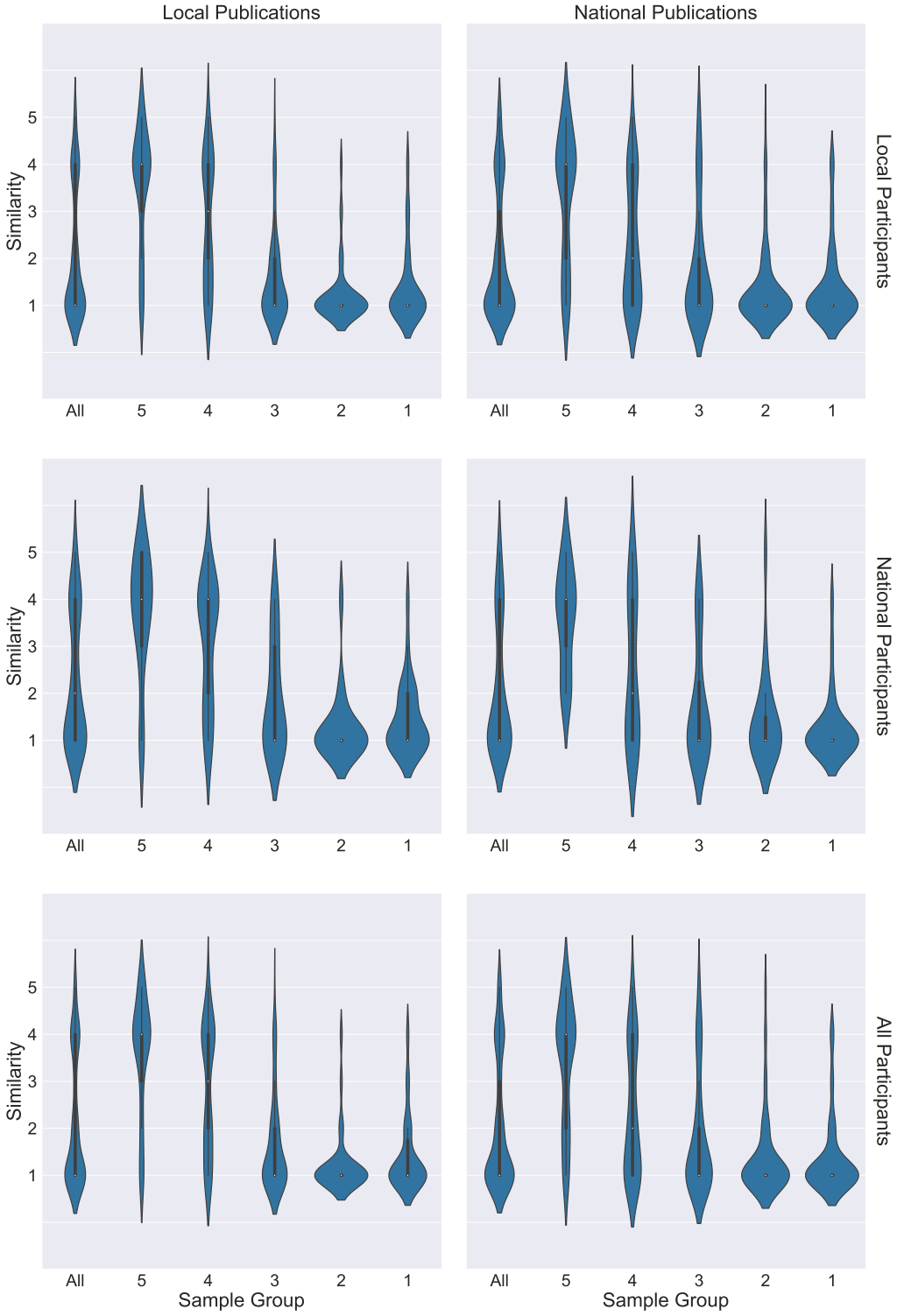

Figure 4.2: Violin Plots of Similarity Scores by Participant Location and Publication Level. 5 is the most similar group, 1 is the least similar group, see Section 3.3.2 for details of the groups.

Table 4.3: Results of running t-test on the local pair ratings vs national pair ratings of the participants. Group 1 is the least similar and 5 is the most similar articles. Note: In this table, National and Local denote the reported residence of the participants.

	All participants		**National participants**		**Local participants**	
Group	T	*p*	T	*p*	T	*p*
All	1.896	0.060	2.139	0.041	1.094	0.277
5	1.280	0.204	0.370	0.715	1.253	0.214
4	2.801	0.006	1.591	0.124	2.300	0.024
3	-1.313	0.118	0.418	0.681	-1.584	0.118
2	-1.682	0.480	-0.721	0.480	-1.554	0.124
1	0.575	0.566	0.926	0.363	0.0	1.0

national participants across all groups is one level higher for the local publication than the national publications. Similarly, when looking at the ratings for all participants, the inner quartile for the *all* group is higher for the local publications than for the national publications. Looking at the quartiles for the smaller groups will likely not point us to any results because of the low number of participants.

In the survey, each participant was asked to rate 10 pairs, one pair from each group, for both a national and local publication. This was done in order to perform a pairwise t-test on the results, to find differences in ratings between the local and national publications. Because of the possible multi-modality of the ratings, a Wilcoxon signed-rank test was also performed. Since the attention check replaced a random pair, the corresponding national or local pair in the same group were removed in addition to the attention check that was removed during the cleaning of the survey results. The results from the tests can be seen in Tables 4.3 and 4.4.

From the results of the t-test, there are a couple of statistically significant findings. The most significant is that the ratings for group 4 are higher for local publications than for national publications. This is significant both when considering all participants ($p<0.01$) and when considering local participants ($p>0.05$). The same findings can be seen in the Wilcoxon signed-rank test. The t-test also shows that national participants overall have given the local pairs higher similarity scores ($p<0.05$). However, when looking at the Wilcoxon test this significance is lost. Instead, the Wilcoxon signed-rank test shows the same result when considering all participants ($p<0.05$). The Wilcoxon signed-rank test is likely to be more appropriate for evaluating the significance when the groups are combined, due to the potential of multi-modality.

Table 4.4: Results of running Wilcoxon signed-rank test on the local pair ratings vs national pair ratings of the participants. Group 1 is the least similar and 5 is the most similar articles. Note: In this table, National and Local denote the reported residence of the participants.

	All participants		**National participants**		**Local participants**	
Group	W	p	W	p	W	p
All	2724.0	0.048	112.0	0.064	1757.0	0.242
5	608.5	0.169	50.5	0.582	316.0	0.201
4	834.5	0.006	56.5	0.116	464.5	0.024
3	339.5	0.229	15.5	0.719	199.0	0.138
2	100.0	0.080	9.0	0.380	50.5	0.114
1	175.0	0.507	17.5	0.289	84.0	0.946

4.3.2 Difference in Correlations Between National and Local Level. (RQ 4.2)

The final research question this thesis attempts to answer is whether differences in human similarity judgments on the local and national level in the Norwegian News Domain affect the Feature-Specific Similarity Metrics in a statistically significant way. In order to test this, Fisher r-to-z transformations were performed on a selection of the correlations calculated in Table 4.1. The Z-values were then pairwise compared by performing a Z-test. This was performed on various compositions of national and local publications.

The *All* Z-test is calculated using the *National* and *Local* correlations. The *Schibsted* Z-test is calculated using the correlations for *VG* and *BT*, the *Amedia* Z-test is calculated using the correlations for *Nettavisen* and *BA*, and the *VG vs BA* Z-test is calculated using the mentioned publications. The reasoning for the final group is to evaluate the most local publication against the most national publication. The Z-test calculations can be seen in Table 4.5.

From Table 4.1 we observe that the metrics on the metrics using low-level image features for *VG* show very low correlations. The difference of the *Image:BR*, *Image:EN* and *Image:SH* correlations are observed. The *Image:EMB* correlations do however seem like they differ across the national level. This difference is also significant when including all publications, as well as when looking at the Amedia composition.

The difference between the correlations of the *Section:JACC* metric, are significant only in the columns that consider *VG*, and while not significant, the Z-test result for the Amedia column holds the opposite direction than the others. There are also statistically significant differences between the *Text:NENTS*, *Title:LDA* and *Title:BERTopic* correlations, but they also do not manifest across all columns. The highest statistical significance can be seen for *Title:BERTopic*, but only in the *VG vs BA* column. Manually checking the *Title:BERTopic* score for *Nettavisen* and *VG* returns a score for -1.788 with a p-value of 0.074.

Table 4.5: Results of Z-test comparing national vs local news feature correlation after performing Fisher-r-to-z on the data in Table 4.1. All: VG and Nettavisen vs BT and BA. Schibsted: VG vs BT. Amedia: Nettavisen vs BA. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	All	**Schibsted**	**Amedia**	**VG vs BA**
Image:BR	-2.828**	-3.800***	-0.075	-2.451*
Image:COL	-0.018	-0.008	0.102	0.775
Image:CO	-0.605	-0.290	-0.599	-0.297
Image:EN	-2.269*	-2.497*	-0.728	-2.104*
Image:SH	-0.678	-2.439*	1.566	-2.311*
Image:EMB	2.819**	1.543	2.291*	0.416
Text:BERTopic	1.033	0.432	0.955	-0.054
Text:LDA	-0.096	0.131	0.395	1.030
Text:NENTS	0.380	-1.830	2.526*	-0.217
Text:SBERT	-0.949	2.159	1.032	-1.397
Text:TF-IDF	-0.675	-1.700	0.747	-1.228
Text:TF-IDF-L	-1.155	-1.704	0.010	-1.668
Text:TF-IDF-50	-1.065	0.190	-1.822	-0.732
Time:Days	-0.628	-0.990	0.020	-0.621
Section:JACC	-0.726	-3.377***	0.574	-3.442***
Tags:JACC	1.129	0.033	0.382	-2.180*
Title:JW	-0.223	-1.139	0.877	-0.515
Title:LCS	1.679	-0.377	2.856	1.044
Title:LDA	-1.038	-1.977*	1.281	1.234
Title:LV	0.452	0.472	-0.041	-0.755
Title:SBERT	-0.285	-1.504	1.056	0.262
Title:TF-IDF	-0.132	-0.858	0.383	-1.792
Title:TF-IDF-L	-0.443	-0.267	-0.536	-1.932
Title:BI	0.443	0.340	0.027	-0.606
Title:BERTopic	-0.760	-0.416	-1.100	-2.920**

Chapter 5

Conclusion and Future Work

5.1 Discussion

The main question of this thesis has been to find out whether there are differences in human similarity judgments of news articles between the local and national levels of Norwegian news. I was only able to find some minor differences, showing that similar local news is slightly more similar than similar national news. This difference was not enough to influence representations of the feature-specific similarity metrics in any meaningful way.

The main finding in this thesis are the representations of the BERT-based metrics, and in particular the SBERT metric. The SBERT metric turned out to be the metric that best represented the human similarity judgments on both features it was used on.

5.1.1 Usage of Information Cues (RQ1)

The results from the analysis of the participants' reported similarity cue usage show that the text feature along with the section are the most used features to determine similarity. The usage of body text is in line with previous work [49, 48], however, the reported usage of the section is considerably higher than in the previous work. The main reason for this might be that there is a much wider variety of articles in this thesis. The two main previous works both used subcategories within only two topics, which may be more difficult to discern for the user. While with the dataset used here, the user might be presented one article about "Sports" and another about "Politics" and be able to determine that the articles are different by the section alone. This may be more difficult with subcategories where both articles are within a main category. The lower usage of the title is somewhat surprising, and below that of Starke et al. [49] and Solberg [48], although slightly. The title should be the first thing that is read, but based on the result, users may not consider it descriptive of the similarity between

two articles. The usage of the date is reported to be somewhat low, and even lower than the previous work. While the concept of *recency* is important in the literature [26], it is more likely to show up in more general news recommendation contexts and not necessarily in an evaluation of news similarity as we are doing here.

5.1.2 Representativness of Feature-Specific Similarity Metrics (RQ2)

One of the main outputs of the methodology put forward by Trattner and Jannach [53] and used in Starke et al. [49] and Solberg [48] is to look at how well feature-specific similarity metrics represent human similarity judgments. In this thesis, such representations were calculated both for the entire set of ratings, but also for ratings divided by national and local publications, as well as for each of the four publications used in the dataset.

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

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

The high correlation of *Section:JACC* in this study (0.49) compared to Solberg [48] (0.44) and Starke et al. [49] (0.14) is notable. Moreover, the difference in correlations between Schibsted and Amedia publications may require additional analysis. The potential influence of the number of sections in each publication, and the distribution of articles across categories, particularly between VG and BT, are factors that might be evaluated in future studies.

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

The *Time:Days* metric's higher correlation in this study (0.22) compared to Starke et al. [49] (0.09) and Solberg [48] (0.03) raises questions about the relationship between self-reported information cues and their predictive ability for relevant features. This finding echoes the observations in [53].

The *Text:LDA* metric have higher correlations than in Starke et al. [49] and Solberg [48], however it is lower than in both domains explored in [53]. Curiously, the *Title:LDA* shows some weak correlation when looking at the BT pair ratings alone. In all other domains except for the recipe domain, *Title:LDA* have failed to show any correlations. This suggest that the training setup for LDA used in this study and the previous work is not optimal for when short texts like titles are evaluated.

The correlations for the *Text:NENTS* metric is lower than expected and show a wide range across the different publications. This suggests that it may be more effective in certain contexts. This aligns with findings from Solberg [48], where it was found to be more relevant for the *Sports* category than the *Recent Events* category. Further examination of this metric's performance across various categories could yield interesting insights.

The higher correlation of *Image:EMB* of 0.30 in the current study compared to Starke et al. [49] (0.17) and Solberg [48] (0.12) is noteworthy. Furthermore, the moderate correlation for Nettavisen (0.46) is particularly noteworthy, as it rivals some of the metrics with the highest correlations. One odd result is how *Image:BR*, *Image:SH*, and *Image:EN* shows correlations too weak to be significant when measured against the VG ratings. For the other publications, as well as with previous work, the metrics generally show some weak or very weak correlations. What may cause this is difficult to discern, but looking at image use across the publications could reveal some insights.

5.1.3 Comparison with Other Domains (RQ3)

When looking at the feature-specific metric representation across domains, which are displayed in Table 4.2, there are a couple of things that could be noted. First of all, we can quickly see how Starke et al. [49] shows overall correlations across all metrics when compared to the other domains. The reason for this is unclear but expected to be because of low amount of ratings of similar pairs.

Across all domains, we see that TF-IDF on the text feature is among the highest correlating metrics. It must be noted however that for the recipe and movie domain, the text feature represents the directions and plot respectively. Considering the high correlations of the SBERT-based metric it could be interesting to see if this performance could be reflected in the other domains as well.

All of the domains have used the same training parameters for the LDA model, while this generally seems to have a good effect when applied to the text, especially in the recipe domain, we see that for the title it is only in the recipe domain the metric returns even a weak correlation. The edit-distance-based metrics also show a much lower correlation outside of the recipe domain. Given their computational cost, it becomes difficult to defend their usage in a large-scale recommender setting.

Across all domains, we see that the Section:JACC metric shows moderate correlations with the similarity judgments. For the different domains the feature referred to as Section does differ slightly. In the Starke et al. [49] and Solberg [48] it refers to the subsection below a specific category. While in the movie domain, it is the genre of the movie. In the Norwegian domains, it is used both on the category feature for the Amdia publications and for the subsection feature on Schibsted publications. It is unsurprising that it offers high correlations as the purpose of categories is to group similar items.

Overall, we see that for the Norwegian news domain, a majority of the metrics show some level of correlation, even though a majority is weak. This is different from the other news domains where only a few metrics show more than a very weak correlation to the human similarity judgments. In that vein, the Norwegian news domains are closer to the movies and recipe domains.

5.1.4 Differences in Human Similarity Ratings Across National and Local Domains (RQ4.1)

The virgin plots, t-test, and Wilcoxon rank test did not show any major differences between the local and national levels. We do see a tendency that users may consider local news more similar overall. This is intuitive as well, assuming the readers would consider the same geographical area as a factor of similarity. However the results are just barely significant, and the t-test and Wilcoxon rank test varied on whether this should be considered statistically significant when considering the ratings from all participants or only participants from outside the Bergen Area. For local participants, this effect was not large enough statistically significant, giving an indication local users may not use the geography of their local area as a similarity indicator. However, that particular question was not specifically tested for in the study.

The one clear difference found is that participants rated the second most similar sample group, sample group 4, as more similar in the local publications than in the national publications. This indicates that articles in local publications have a slower reduction in similarity compared to national publications, at least when measured by the combined similarity metrics. This could also be indicative of the news being less varied, however, the distribution of

articles across categories seen in Figure 3.1 could be taken as an indicator against this.

While not statistically significant, we do also see the opposite effect for sample groups 2 and 3. Giving an indication that local news also has a higher human-judged dissimilarity between articles as the combined similarity scores of the metrics decline. However, as this is not statistically significant, this observation can be random.

Overall, we only see some very small differences between the way humans rate the similarity of articles across the national and local publications in the Norwegian news domain.

5.1.5 Feature-Specific Metric Representations Across Domains (RQ4.2)

Considering the results for RQ4.1, that the difference of human similarity judgments is small, the expectation would be that the results for RQ4.2 would also be small, or non-existent. This is also what we see from the results. In order to be confident that a result is due to differences in the local and national level, the results for the Z-test for a specific metric shown in Table 4.5 should fulfill the following criteria: All the values of the row should be of the same direction, that is, all should either be positive or all should be negative. In addition, the results should be statistically significant. Ideally, we should also see the largest value in the *VG vs BA* column as that is where we are comparing the *most national* to the *most local* publication. By holding these criteria, there are some metrics that warrant a closer look.

The *Image:BR* and *Image:EN* both generally fulfill the criteria listed above, although the difference is not statistically significant for the Amedia group. When we take a closer look at the correlations in Table 4.1, it also quickly becomes clear that the differences are a result of the low correlations the metrics have when evaluated against the VG ratings, rather than something inherent in the local or national domain. This underscores the influence of individual publications, such as *VG*, on the outcome of these correlations, which is a key point to consider when interpreting these results.

Additionally, *Section:JACC* and *Text:NENTS* exhibit statistically significant differences, although they do not meet all the criteria for consideration. Nevertheless, their presence provides valuable insight into the nuanced interaction between human similarity judgments and the local and national levels.

Image:EMB is also close to fulfilling the criteria. However, the smallest difference can be seen while looking at the *VG vs BA* column, indicating that it's not necessarily a difference between the local and national levels, as much as it's a difference between Nettavisen and the other publications. Considering that Nettavisen also is an outlier in terms of a high *Image:EMB* score, it is tempting to rule this metric out because of it. At the same time, we are also seeing a fairly large difference when looking at the Schibsted column, but it is still not large enough

to be statistically significant. It is difficult to reach a specific conclusion in regard to this metric.

Finally, *Title:BERTopic* also initially looks like a potential metric that correlates differently with human similarity judgments across the local and national domains. It is intuitive that titles may be something that differs across these domains as well. However, closer inspection reveals that the difference is primarily related to *VG* and not the local and national domains. When manually checking for the difference in correlation between Nettavisen and VG, it ends up being just short of being significant, and larger than the differences between the local and national publications within the Amedia and Schibsted datasets. This implies that the difference we are seeing is more likely to be at the publication level, rather than the geographical level.

All in all, we can conclude that the differences in ratings between the local and national levels only result in very minute differences in how the metrics represent the human similarity judgments. In addition, there are clear indications that the differences in ratings at *publication* level is larger than on the *geographical* level. Investigating this could yield interesting results.

Within the broader context of this field of research, these findings highlight the complexities involved in interpreting similarity judgments. As observed, individual publications like *VG* may significantly influence these judgments, which can challenge the assessment of similarities on a broader geographical scale, such as local and national domains. This emphasizes the need for a careful and multi-faceted interpretation of the results, as well as further research into the influence of individual publication biases and their impact on similarity judgments.

5.2 Limitations and Future Work

One of the main limitations of this study is the number of participants. Participant recruitment proved harder than expected, and because of this, it turned difficult to do in-depth statistical analysis across the different subdivisions of the pair ratings. Overall it still allowed for meaningful conclusions, but the data do indicate other findings that did not turn out to be statistically significant.

In the survey, the participants were asked for their place of residency. This question was meant to give the opportunity into dividing the group by their familiarity with the local domain. However, during recruitment, it became apparent that applying this strategy to students would not have this result, as many students who live in Bergen are from out-

side Bergen. Conversely, participants who report their residency outside Bergen may still be highly familiar with Bergen, perhaps even being from Bergen themselves. Formulating the correct question here might be difficult, but in reality, what needs to be known for the correct divisions is the participants' familiarity with the local news.

While Bergen was used as the local domain for this survey, Bergen itself is a moderately large city and within Bergen there are several publications targeting districts within Bergen. Bergen's size may also make it so that the local newspapers don't inhabit features that are generally associated with local news media. This would be especially true for BT, whose mission is to provide its readers with a full spectrum of news, including foreign affairs. At the same time, Amedia is a news organization that primarily handles local news, and Nettavisen is their only national newspaper. While Amedia has a "hands-off" approach to the various newsrooms in the organizations, because of the dominance of local news outlets in the organization, it can be speculated that characteristics typical of local news media may influence the entire organization, including Nettavisen. Combining these two hypothetical situations, this would end up giving us a situation where BT would be a "local national" publication, while Nettavisen would be a "national local" publication. If this should turn out to be the case, it would help explain the modest results of the analysis. Taking this into consideration, it could be interesting to attempt a similar study to this but look at publications at an even more local level than was done here. For example by including the publications targeting the districts within Bergen, or publications targeting smaller communities than Bergen.

An obvious form of future work for this thesis would be to test the findings in a recommendation scenario. In particular, attempting to develop the SBERT metric for news recommendation and testing different approaches on users. Other combinations of metrics could also be considered. As a direct next step, Trattner and Jannach [53] and Starke et al. [49] both used human similarity judgments to learn weights for the various metrics, and used these to develop a specific similarity function encompassing several metrics.

References

[1] Alkula, R. (2001). From plain character strings to meaningful words: Producing better full text databases for inflectional and compounding languages with morphological analysis software. *Information Retrieval 4*(3), 195–208.

[2] Allison, L. and T. I. Dix (1986). A bit-string longest-common-subsequence algorithm. *Information Processing Letters 23*(5), 305–310.

[3] Balakrishnan, V. and L.-Y. Ethel (2014). Stemming and lemmatization: A comparison of retrieval performances. *Lecture Notes on Software Engineering 2*, 262–267.

[4] Beel, J., B. Gipp, S. Langer, and C. Breitinger (2016). Research-paper recommender systems : a literature survey. *International Journal on Digital Libraries 17*(4), 305–338.

[5] Billsus, D. and M. J. Pazzani (2000). User modeling for adaptive news access. *User Modeling and User-Adapted Interaction 10*(2), 147–180.

[6] Bjerke-Lindstrøm, B. (2017). Teaching nltk norwegian.

[7] Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *J. Mach. Learn. Res. 3*, 993–1022.

[8] Bollegala, D., Y. Matsuo, and M. Ishizuka (2007). Measuring semantic similarity between words using web search engines. In *16th International World Wide Web Conference, WWW2007*, pp. 757–766.

[9] Campello, R. J. G. B., D. Moulavi, and J. Sander (2013). Density-based clustering based on hierarchical density estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu (Eds.), *Advances in Knowledge Discovery and Data Mining*, Berlin, Heidelberg, pp. 160–172. Springer Berlin Heidelberg.

[10] Cantador, I., P. Castells, and L. Gardens (2009). Semantic contextualisation in a news recommender system.

[11] Chamberlain, B. P., E. Rossi, D. Shiebler, S. Sedhain, and M. M. Bronstein (2020). Tuning word2vec for large scale recommendation systems. In *Fourteenth ACM Conference on Recommender Systems*. ACM.

[12] Chu, W., S.-T. Park, T. Beaupre, N. Motgi, A. Phadke, S. Chakraborty, and J. Zachariah (2009). A case study of behavior-driven conjoint analysis on yahoo! front page today module. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, New York, NY, USA, pp. 1097–1104. Association for Computing Machinery.

[13] Das, A. S., M. Datar, A. Garg, and S. Rajaram (2007). Google news personalization: Scalable online collaborative filtering. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, New York, NY, USA, pp. 271–280. Association for Computing Machinery.

[14] Deldjoo, Y., M. Elahi, M. Quadrana, and P. Cremonesi (2015). Toward building a content-based video recommendation system based on low-level features. In H. Stuckenschmidt and D. Jannach (Eds.), *E-Commerce and Web Technologies*, Cham, pp. 45–56. Springer International Publishing.

[15] dos Reis, J. C., F. Benevenuto, P. O. S. V. de Melo, R. O. Prates, H. Kwak, and J. An (2015). Breaking the news: First impressions matter on online news. In *International Conference on Web and Social Media*.

[16] Eide, M., T. Hetland, Allkunne, and O. Garvik. Bergens tidende. Store Norske Leksikon. Retrieved May 22. 2023 from https://snl.no/Bergens_Tidende.

[17] Eksombatchai, C., P. Jindal, J. Z. Liu, Y. Liu, R. Sharma, C. Sugnet, M. Ulrich, and J. Leskovec (2017). Pixie: A system for recommending 3+ billion items to 200+ million users in real-time. *CoRR abs/1711.07601*.

[18] Elahi, M., D. Jannach, L. Skjærven, E. Knudsen, H. Sjøvaag, K. Tolonen, Ø. Holmstad, I. Pipkin, E. Throndsen, A. Stenbom, E. Fiskerud, A. Oesch, L. Vredenberg, and C. Trattner (2022). Towards responsible media recommendation. *AI and Ethics 2*(1), 103–114.

[19] Friendly, F. (2019). Jaro–winkler distance improvement for approximate string search using indexing data for multiuser application. *Journal of Physics: Conference Series 1361*, 012080.

[20] Garcin, F., B. Faltings, O. Donatsch, A. Alazzawi, C. Bruttin, and A. Huber (2014). Offline and online evaluation of news recommender systems at swissinfo.ch. In *Proceedings of*

the 8th ACM Conference on Recommender Systems, RecSys '14, New York, NY, USA, pp. 169–176. Association for Computing Machinery.

[21] Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

[22] Hassan, H. A. M., G. Sansonetti, F. Gasparetti, A. Micarelli, and J. Beel (2019). Bert, elmo, use and infersent sentence encoders: The panacea for research-paper recommendation? In *ACM Conference on Recommender Systems*.

[23] Hoffman, M., F. Bach, and D. Blei (2010). Online learning for latent dirichlet allocation. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (Eds.), *Advances in Neural Information Processing Systems*, Volume 23. Curran Associates, Inc.

[24] Jannach, D., M. Zanker, A. Felfernig, and G. Friedrich (2010). *Recommender Systems an Introduction*. Leiden: Cambridge University Press.

[25] Jørgensen, F., T. Aasmoe, A.-S. R. Husevåg, L. Øvrelid, and E. Velldal (2020). Norne: Annotating named entities for norwegian.

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

[27] Kaviani, M. and H. Rahmani (2020). Emhash: Hashtag recommendation using neural network based on bert embedding. In *2020 6th International Conference on Web Research (ICWR)*, pp. 113–118.

[28] Kobayashi, M. and K. Takeda (2000). Information retrieval on the web. *ACM Comput. Surv. 32*(2), 144–173.

[29] Kompan, M. and M. Bieliková (2010). Content-based news recommendation. In F. Buccafurri and G. Semeraro (Eds.), *E-Commerce and Web Technologies*, Berlin, Heidelberg, pp. 61–72. Springer Berlin Heidelberg.

[30] Kondrak, G. (2005). N-gram similarity and distance. In M. Consens and G. Navarro (Eds.), *String Processing and Information Retrieval*, Berlin, Heidelberg, pp. 115–126. Springer Berlin Heidelberg.

[31] Kummervold, P. E., J. De la Rosa, F. Wetjen, and S. A. Brygfjeld (2021). Operationalizing a national digital library: The case for a Norwegian transformer model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, Reykjavik, Iceland (Online), pp. 20–29. Linköping University Electronic Press, Sweden.

[32] Leighton, K., S. Kardong-Edgren, T. Schneidereith, and C. Foisy-Doll (2021). Using social media and snowball sampling as an alternative recruitment strategy for research. *Clinical simulation in nursing 55*, 37–42.

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

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

[35] Lua, L., C. H. Yeung, Y.-C. Zhang, and Z.-K. Zhang (2018). Survey on collaborative filtering, content-based filtering and hybrid recommendation system.

[36] Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development 1*(4), 309–317.

[37] Lv, Y., T. Moon, P. Kolari, Z. Zheng, X. Wang, and Y. Chang (2011). Learning to model relatedness for news recommendation. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, New York, NY, USA, pp. 57–66. Association for Computing Machinery.

[38] McInnes, L., J. Healy, and J. Melville (2020). Umap: Uniform manifold approximation and projection for dimension reduction.

[39] Messina, P., V. Dominguez, D. Parra, C. Trattner, and A. Soto (2019). Content-based artwork recommendation: integrating painting metadata with neural and manually-engineered visual features. *User Modeling and User-Adapted Interaction 29*(2), 251–290.

[40] Øvrelid, L. and P. Hohle (2016). Universal Dependencies for Norwegian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, pp. 1579–1585. European Language Resources Association (ELRA).

[41] Özgöbek, Ö., J. A. Gulla, and R. C. Erdur (2014). A survey on challenges and methods in news recommendation. In *International Conference on Web Information Systems and Technologies*.

[42] Pettersen, Ø. B. (2023). Nettavisen. Store norske leksikon. Retrieved May 22. 2023 from https://snl.no/Nettavisen.

[43] Porter, M. F. (2001). Snowball: A language for stemming algorithms.

[44] Reimers, N. and I. Gurevych (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*.

[45] Ricci, F., L. Rokach, and B. Shapira (2011). *Introduction to Recommender Systems Handbook*, pp. 1–35. Boston, MA: Springer US.

[46] San Pedro, J. and S. Siersdorfer (2009). Ranking and classifying attractiveness of photos in folksonomies. pp. 771–780.

[47] Simonyan, K. and A. Zisserman (2015). Very deep convolutional networks for large-scale image recognition.

[48] Solberg, V. R. (2022). News recommendation based on human similarity judgment. Master thesis, The University of Bergen. Masteroppgave i informasjonsvitenskap, INFO390, MASV-INFO.

[49] Starke, A., S. Øverhaug Larsen, and C. Trattner (2021). Predicting feature-based similarity in the news domain using human judgments. In *Proceedings of the 9th International Workshop on News Recommendation and Analytics (INRA 2021)*.

[50] Sun, M., Q. Yang, H. Wang, M. Pasquine, and I. A. Hameed (2022). Learning the morphological and syntactic grammars for named entity recognition. *Information (Basel) 13*(2), 49.

[51] Terjesen, E. A. and O. Garvik (2023). Norsk presses historie. Bergensavisen in Store norske leksikon at snl.no. Retreived May 22. 2023 from https://snl.no/Bergensavisen.

[52] Tintarev, N. and J. Masthoff (2006). Similarity for news recommender systems. In G. Uchyigit (Ed.), *Workshop on Recommender Systems and Intelligent User Interfaces*. In conjunction with the International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, AH 2006, Dublin, Ireland, June 20-23, 2006.

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

[54] Unnikrishnan, R. and M. Hebert (2005). Measures of similarity. In *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05) - Volume 1*, Volume 1, pp. 394–394.

[55] Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017). Attention is all you need.

[56] Williams, A., N. Nangia, and S. Bowman (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics.

[57] Winecoff, A., F. Brasoveanu, B. Casavant, P. Washabaugh, and M. Graham (2019). Users in the loop: A psychologically-informed approach to similar item retrieval.

[58] Yang, N., J. Jo, M. Jeon, W. Kim, and J. Kang (2022). Semantic and explainable research-related recommendation system based on semi-supervised methodology using bert and lda models. *Expert Systems with Applications 190*, 116209.

[59] Yao, Y. and F. M. Harper (2018). Judging similarity: A user-centric study of related item recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, New York, NY, USA, pp. 288–296. Association for Computing Machinery.

[60] Yujian, L. and L. Bo (2007). A normalized levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence 29*(6), 1091–1095.

[61] Zhang, Q., J. Li, Q. Jia, C. Wang, J. Zhu, Z. Wang, and X. He (2021). Unbert: User-news matching bert for news recommendation. pp. 3356–3362.

[62] Zhou, K., Y. Zhou, W. X. Zhao, X. Wang, and J.-R. Wen (2020). Towards topic-guided conversational recommender system.