A Text Similarity Study: Understanding How Differently Greek News Media Describe News Events

Nikos Kapellas[a] and Sarantos Kapidakis[b]

Department of Archival, Library and Information Studies, University of West Attica,
Ag. Spyridonos 28 12243, Athens, Greece

Keywords: Greek News Media, News Events, News Articles, Similarity Analysis, Cosine Similarity.

Abstract: Online news media usually cover the same news events. It would be interesting to understand, how similar are the descriptions of these events. In other words to explore how a distinct news event, is described by different news media. Can we safely conclude when two or more event descriptions refer to the same event? This research aims to investigate similarities in Greek news articles and provide new data on the field, as there is no relevant research. To do so, news article's text is extracted, processed and analyzed with automated methods. To establish an understanding of similarity, three similarity settings are explored and special focus is given in examining the degree of similarity between news articles that cover the same event. In some cases results are inconclusive, while in others results show that groups of news media share the exact same articles with little or no modifications at all. The similarity analysis of news articles is a complex task and relies on many different factors that need to be addressed.

1 INTRODUCTION

In the era of social media and online news one major issue is the degrading quality of the information that is being shared. More than once fake news have gone viral and biased authors, unintentionally spread misinformation. News media controlled by different interests, write articles aiming to convince individuals on a certain topic.

Additionally, a phenomenon that is observed on a daily basis, is that of duplicate or near duplicate articles that circulate, among online newspapers. And while it is easy to address this duplication task via automated methods, it is difficult for readers to understand how similar specific news articles are. Often there are minor transformations to an original article: information is added or removed, the title is altered, a different source is credited, the structure is shuffled, and different images are posted.

Initially we are not interested in proposing a novel methodology for near duplicate detection, as the goal of this study is to put similarity in context. Thus, we construct three different controlled similarity settings to explore. We aim in analyzing news articles in Greek, as there are no relevant studies. Also, Greek can be considered a low-resource language, in terms of datasets and processing tools. This is part of the motivation to research local news articles, instead of using an English dataset for example. Moreover, the monetization paradigm in this particular news media landscape, relies primarily on ads. This fact, disables the involved parties in producing original journalistic material, as they struggle to keep up with competition; It is practically impossible for many online news media to cover every national and global news event. Thus, editors prefer to copy or plagiarise news articles and make minor modifications (or no modifications at all), and post them to their channels.

We assume that this can be problematic, since the information contained in these articles is not always cross-checked or validated. This publication practice allows many forms of misinformation to manifest and to be shared with ease and in an incredible speed. Readers on the other hand, are presented with the same information, rather than different views and details of a specific news event. By exploring the degree of similarity between different news events, and focusing on examining whether we can safely conclude if two or more event descriptions refer to the same event, we aim in understanding how text similarity is propagated.

In section 2, we are going to present selected re-

[a] https://orcid.org/0000-0002-2767-3956
[b] https://orcid.org/0000-0002-8723-0276

Kapellas, N. and Kapidakis, S.
A Text Similarity Study: Understanding How Differently Greek News Media Describe News Events.
DOI: 10.5220/0011589700003335
In *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2022) - Volume 2: KEOD*, pages 245-252
ISBN: 978-989-758-614-9; ISSN: 2184-3228

search that touches on the various subjects of our topic and provide helpful terminology regarding our study. In section 3, we explain the methodology followed in creating three small news articles datasets and the initial processing of this text corpus. In section 4, we present the results obtained from the similarity analysis and we present some additional information from the clustering algorithms. In section 5, we provide some thoughts and conclusion regarding the overall workflow and future developments.

2 RELATED WORK

In literature an event is something that happened at a specific time, in a specific place. There is a difference though between an event and a generic theme. For example, the plane crash of the Ukraine International Airlines Flight 752 is considered an event, while “plane crashes” is a theme. It is observed that “bursts” of news articles follow each event in the media (Yang et al., 1999).

At the same time, a big event may comprise smaller ones. For example, the national elections in Greece cover a broad range of other events or themes, such as election campaigns, opinion polls, turnout of voters, results and so on. In general, events are more precise than themes since the factor of time is rather crucial. Say, the different invasions of Afghanistan by the Soviets and the US, are different events that repeated over the years (Po et al., 2017).

Regarding the trustworthiness of online news, it is proven that they are often biased, something that can lead users to adopt biased opinions as well (Spinde et al., 2021), (Hamborg et al., 2019). Another concern is populism or the audience agenda; prior to online news, editors had only a general idea regarding their reading audience. Nowadays, editors have detailed data, regarding the popularity of published articles. This may lead to a journalistic culture where important but non-appealing stories make space for news that grab attention more easily (Bright and Nicholls, 2014).

In (Nicholls and Bright, 2018) authors propose “news stories”, a second observational level for news events. News stories are defined as a collection of articles that approach a specific theme from different perspectives, or a collection of articles that contains news articles on an event of continuous interest, the initial news article, and its updates.

Another concerning phenomenon of the online news industry, is the fact that if a news event makes a headline to a big or international news media, it is more likely to become a headline to smaller, local news media as well. These smaller media will effectively change their content to match the coverage offered by prominent news media (Papadopoulou et al., 2021).

New age journalism shifts focus from the old fashioned editors-readers dynamic, to promoting the involvement of people in the field, finding new publishing procedures. Technology, financial crisis and distrust in news media, created circumstances for the so-called “citizen journalism”. Since people are not obliged to follow standard journalistic practices, it is claimed, yet uncertain, that citizen journalism gave a helping hand to the propagation of rumors, fake and unverified news (Antonopoulos et al., 2020a), (Rubin et al., 2015).

Greece’s online news industry is also steadily changing. In the last 10 years, more than 50 news media, 9 radio stations and 3 TV channels have closed, while 4 news groups have been sold or closed as well. According to data, average circulation of newspapers has dropped from around 220.000 in 2016 to 98.000 in 2015 (Leandros and Papadopoulou, 2020). It is also noted that 46% of international media follows a subscription model for giving access to their content, while in Greece only 9% follows a similar practice (Antonopoulos et al., 2020b).

In journalism, there is a wide range of practices: transformations in prototype text, summaries, shrinking and reinforcement, reconstruction of pieces of information to write a concise news article (Tenenboim-Weinblatt and Baden, 2018).

News articles can be described by a set of terms that represent the main event or theme within the document. It is reasonable for documents that describe relevant, similar or the same events, to contain common terms. Thus, to identify news events useful key terms must be extracted from the text. However, as it happens key terms lists are often dominated by non-representative words, such as stop-words or corpus-specific terms (Fan et al., 2017).

A basic attribute of news pages, is that they use a common structure (template) to present content that can potentially aid in automatically extracting text (Varlamis et al., 2014). However in many occasions, authors fail to declare the appropriate information inside the corresponding HTML tags, making the process of identifying useful content harder (Hu et al., 2005), (Mohammadzadeh et al., 2012).

There are two major practices in extracting text from news pages: the first approach uses sets of rules, while the second uses parser that try to make an educated guess regarding the location of a page’s content, extracting only specific parts (Ibrahim et al.,). As (Yi et al.,) mentions noise on the web falls under two categories, depending on the level of detail: global noises a website, such as ads, navigation bars, GDPR and copyright statements, of the web, that include duplicate or obsolete web pages and local noise, that concern various parts of links to social media. As mentioned, templates may prove useful for the extraction process. On the other hand, some argue that the wide usage of templates is problematic.They have a negative impact on the efficiency of tools related to web pages processing and they represent 40-50% of data on the web and this value increases by 6% each year (Vieira et al., 2006).

For text, determining the similarity degree between words is the starting point for phrase, paragraph, and document similarity. A similarity measure based on a text corpus, quantifies the similarity between words according to information obtained by big corpora. A similarity measure based on knowledge, quantifies the degree of that similarity between words, using information from semantic networks (H.Gomaa and A. Fahmy, 2013), (Dhyani et al., 2002). This came as a solution in the ever-growing volume of data and information overload. The amount of information users can handle given a specific time frame is limited, making it more difficult to retrieve and access information within their interest (Ahmed and Hanif, 2020).

There numerous applications were similarity analysis lies on their core. Recommendation systems are helping users locate, filter information, and make decisions, according to their needs, when there is no prior knowledge to evaluate a product or a piece of information (Feng et al., 2020). Automatic summarization is also a research field relevant to similarity analysis. The idea is that units of text, such as sentences, can represent whole documents (Llewellyn et al.,). Paraphrase identification has similarity in its core as well. Paraphrase identification studies in which ways systems can identify alternative linguistic expressions of the same meaning, in different texts and levels (documents, paragraphs, sentences) (AL-Smadi et al., 2017). Classification applications contain the elimination of duplicates from databases, spam email filtering, plagiarism detection, among other (Soloshenko et al., 2015). Classification may be supervised or unsupervised. In the first case, prior knowledge is needed on the data, while in unsupervised there is no such need (Prasad et al., 2018). In general, defining a pair of similar items is not always clear. In news articles, research is often focused on classifying the content into categories such as politics, economy, sports (Huang, 2008). Specifically for text, since the number and composition of the elements involved are not well defined, documents are considered unstructured data. Therefore, a structure is applied by using a text representation method (Mozer et al., 2020).

2.1 Terminology

To avoid confusion, some basic terminology is provided in the context of this research. An *event*, that is a news event, can be of the following three categories: *same event* (news describing the same event), *related event* (news describing events that have something in common, ex. a common theme), and *unrelated event* (news describing different events).

Also, regarding the similarity of news *news articles*, there are 4 distinct categories: articles can be *identical*, *similar*, *dissimilar* or *discrete*. Moreover, articles can be *similar*, if they are *near-duplicates*, *reduced* (containing less information) and *modified* (additional changes in text and structure). *discrete* articles, concern only *related event* and *unrelated event* news, while the rest categories (identical, similar and dissimilar) concern *same event* news.

The term *news media* is referring to news sites, primarily online newspapers. Lastly, in some cases, the terms *documents* and *news articles* are used to denote the same thing.

In the following section, the rationale of formulating the methodology is explained.

3 METHODOLOGY

3.1 Creating the News Articles Datasets

To better describe the similarity between articles, three different data sets are formed by extracting text from news media sites. Each set contains 24 news articles from 24 different news media; one news article from each news media website. All of the articles included in the three mentioned sets were selected and checked manually. The first set contains *unrelated event* news articles. These belong to various categories (sports, politics, economy) and each presents an event that has no connections with any other event within the set. The second set contains *related event* news articles that are retrieved based on the common theme “accident”, but report on different events; that is, articles that present accidents that happened in different places or time, involving different actors. The third set contains *same event* news articles, all reporting on a specific event that was published by different media.

This distinct event is related to the economy of Greece. Specifically, the story is about how the European leaders agreed to support Greece by providing financial aid, due to the economic recession caused by the pandemic of COVID-19. In total, the entire dataset numbers 72 news documents. The article selection process was done manually, by visiting the news media websites and inspecting their content. Later, the article's text was extracted and stored into separate documents, one for each article.

The objective here is threefold: First, to establish a quantified meaning of similarity, by exploring similarity limits for each set. Second, to compare similarity observations among sets, to better understand how similarity changes according to the different event categories. Third, to specifically examine the range of similarities between articles of the third set, containing news articles of the *same event* and evaluate those findings.

The underlying goal is to understand how differently the media describe the news events. It is outside of the scope of this study to provide a competitive methodology for text extraction or a novel duplicate detection algorithm. To analyze the similarity news articles, that describe the *same event*, is a low starting point for a task as such.

In the following section text processing is described.

3.2 How to Perceive Similarity

There are several ways to represent a text document. Here the three distinct sets of documents are transformed into TF-IDF matrices. Each word represents a dimension in the resulting space and each document becomes a vector. Term frequency is used as a weight, meaning that the terms that appear more frequently are more descriptive for the document. In practice though, this is not always true. Words such as articles, stop-words, and function words, are not that important for the document subject. With documents as vectors, the degree of similarity between two documents is measured as the correlation between their assigned vectors. This can be quantified as the cosine of the angle between those vectors, as it can be seen in 1.

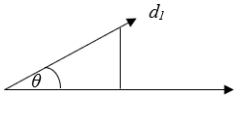

Figure 1: Angle between documents.

To remove stop-words from the corpus, the text was tokenized into words, annotated with part-of-speech (POS) information, and lemmatized. This way each document is transformed into a list of triplets. Each row contains the initial term, its POS tag based on its grammatical category, and its word stem. For example the triplet: *banks NoCmFePlNm bank*. In some cases, words were miss-tokenized, and their lemma was marked as *unknown*. These words were removed from the corpus as well, since they would map into one another, falsely increasing similarity results. Three grammatical categories were qualified as valuable for the scope of our research. These are *nouns*, *adjectives*, and *verbs*. Each of the three categories, has several subcategories specifying their type, such as proper nouns, superlative adjectives, past tense verbs and so on. Words with POS tags other than *(No)* for nouns, *(Aj)* for adjectives, and *(Vb)* for verbs, were removed from each document. To further improve accuracy, too infrequent words were removed from the corpus. Terms that appeared in a few documents for each set, were ignored by setting a minimum threshold to term document frequency.

To further understand how similar the news articles of the third set are, *k-means* and *dendrogram* clustering algorithms are applied. Both have a common application in this case, since they provide a view of the data at different level of abstraction and granularity. For k-means clustering, to determine an appropriate number of clusters, the Elbow method was used. As Fig. 2 shows, the distortion value decreases at a steady rate for our data, creating a rather smooth curve, signifying that this method is inconclusive. Since the elbow-point could not be identified, after some trial and error attempts, the *k* was empirically set to *k=4*.

Results for all three similarity settings are presented in the following section, along with the evaluation of these results and the clustering algorithms outputs.

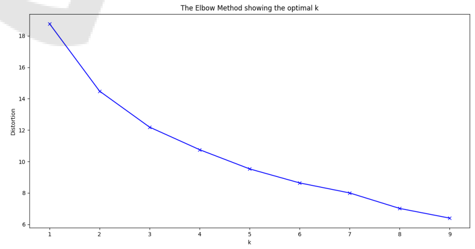

Figure 2: The elbow method showing the optimal k.

4 RESULTS

4.1 Unrelated and Related News Articles Similarity Limits

The similarity analysis between news articles of the first set, that of unrelated event news articles, revealed that there are 48 article pairs with similarity score above 0.384 (similarity threshold), while 2 observations are equal to 0.03 (minimum) and another 2 are equal to 0.61 (maximum).

For the second set, that contains related event news, drawn from the common news theme ”accident”, there are 67 observations above 0.384 and the lowest similarity observed is equal to 0.04. On the other hand, there are 16 observations between 0.5-0.56 (maximum).

In contrast, in the third set, that of same event news, there are 20 pairings of high similarity scores, ranging from 0.69 to 0.91, 87 similarity values from 0.50 to 0.69, whereas the rest pairs have scores from 0.49 to 0.15 (lowest).

These results indicate that for the first two sets, the similarity minimum and maximum are approximately the same, even though the number of news articles above the similarity threshold differs.

The similarity minimum of these two sets is of different order, compared to the minimum value of the third set. Accordingly, the maximum value of the third set, is significantly higher than the maximum of the two other sets.

Also, the similarity limits of the first two sets, were expected to increase or decrease more gradually. The similarity maximum of the unrelated event news set (0.61), was expected to be lower. Likewise, the similarity minimum of the related event news set (0.04), was expected to be higher.

In the third set, that of same event news articles, the similarity limits differ substantially, observing a higher minimum and maximum, while many values range well-above the similarity threshold.

Based on the above, a context of similarity is beginning to emerge, even though not all observations follow reasonable patterns. By measuring similarity, an understanding of what similar means can be formed, to evaluate the descriptions for the different categories of news events.

4.2 Same Event News Articles Similarity Analysis

Special focus is given on the *same event* news set. In it, some news articles appear *identical* or *similar* to one another, while others seem to be *dissimilar*, which is reasonable even though they describe the same event.

In Fig. 3 we can see the document similarity matrix of the same event news set, with a color scale on the right, indicating the degree of that similarity, with 1 being similar (close to yellow) and 0 being different (close to blue). In both x, y axis there are labels for each one of the 24 documents. Each document is paired with every other document of the dataset, forming a matrix of 576 squares (288 pairings). The observations running the matrix diagonally, represent the same document pairings.

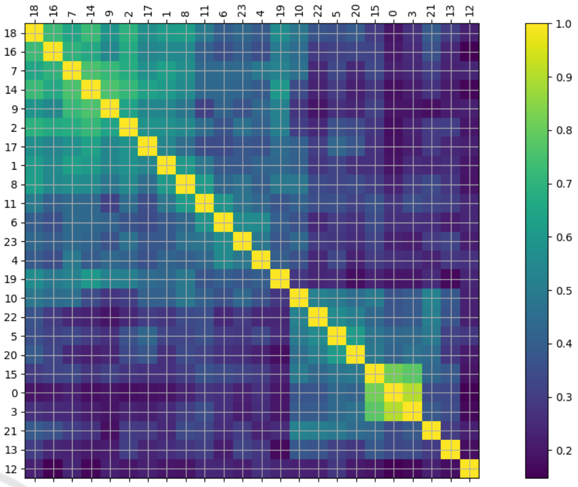

Figure 3: Pair-wise similarity matrix.

Table 1 shows the highest similarity score values. Columns Document A and Document B represent pairs of documents and cosine similarity is represented as the similarity score. Based on the results, document 0 and document 3 could be classified as *identical* with a score of 0.91, the pairs of documents 0, 15 and 15, 3 are *near duplicates* with scores of 0.82 and 0.78 accordingly, while document 14 is 0.72 *reduced* comparing to document 2, and so on and so forth.

Table 1 indicates that there are two groups of *identical* or *similar* documents, within the set of articles. The first consists of documents 0, 3, 15, whereas the second consists of documents 14, 2, 7, 9, 16, 18. However, not all documents of the second group are equally similar to each other. For example, even though the pair of documents 7, 16 and 16, 18 are highly similar, their intersection pair of documents 7 and 18, has a lower similarity score of 0.65. Moreover, out of the 276 similarity pairings (excluding the 24 same document pairings), 20 pairings have high similarity scores, from 0.69 to 0.91 (highest), 87 have similarity scores from 0.50 to 0.69, whereas the rest pairs have scores from 0.49 to 0.15 (lowest). Interestingly enough, document 19, taken from the foreign news agency Reuters.com and was automatically translated from English to Greek, is similar to a degree of 0.50 or greater, with documents 2, 7, 9, 14, 16 and 18.

Table 1: Highest similarity score pairings.

Document A	Document B	Similarity score
0	3	0.91
0	15	0.82
15	3	0.78
14	2	0.72
7	9	0.72
7	14	0.76
7	16	0.70
9	16	0.76
14	9	0.76
14	18	0.76
16	18	0.73

4.3 Evaluation and Clustering

Regarding the evaluation of the methodology followed, the observed similarity scores were matched against a similarity threshold. That threshold was produced based on the average over all similarities observed. This cut-off value expresses the degree of similarity that two documents must have to be considered as similar. The formula for calculating that a threshold derives from the physical properties of its parts:

$$T = avg(sim) + a \cdot s(sim)$$

where a is a parameter and s is the standard deviation of the similarity values.

The $avg(sim) = 0.246$, is the average over all similarity pairs within the dataset and to distance the T value from that average, parameter a was set to, $a = 0.75$ that is equal to 1.5 standard deviations. The calculated value $T = 0.384$, seems reasonable when examining the pairwise similarity scores. In other words, document pairs with similarity equal or greater to 0.384 qualify as *identical* or *similar*, while pairs with a score below 0.384 appear to be *dissimilar* or *discrete*, depending on the set. The similarity findings reveal 122 observations with a score over 0.384.

Fig. 4 shows how documents of the set are clustered by the k-means algorithm. Each point on the diagram represents a different document. Colors are used to indicate groups and grey circles represent the center of each cluster. The clustering reveals 4 separate clusters. Documents 0, 3, 15 form the first cluster, documents 5, 10, 13, 20, 21, 22 form a second cluster, documents 4, 6, 8, 11, 12, 23 form a third cluster, whereas documents 1, 2, 7, 9, 14, 16, 17, 18, 19 form the biggest cluster of similar documents.

Fig. 5 shows how the dendrogram algorithm clustered the news articles. Horizontally the 24 documents and their clusters are found, colored accordingly and vertically the distance or dissimilarity between those clusters is shown. Each joining of the two clusters is represented on the graph by the split of one vertical line in two. The position of the split is shown by the horizontal bar, which also gives the distance between the two clusters. As expected, documents 0, 3, 15 form a cluster. Documents 1, 2, 7, 9, 14, 17, 18, 19 form a second cluster. Documents 4, 6, 11, 23 form a third cluster and documents 5, 10, 20, 21, 22 form a fourth one.

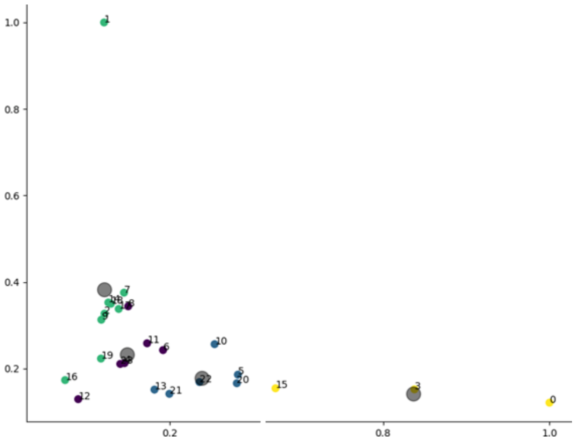

Figure 4: Scatter plot of k-means clustering.

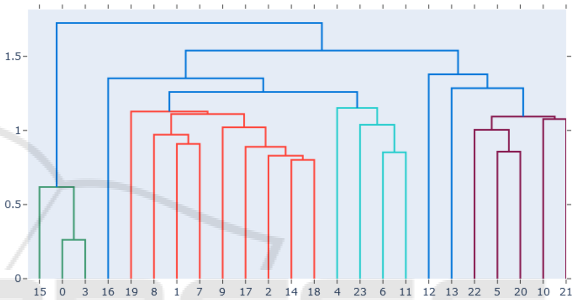

Figure 5: Dendrogram of the hierarchical clustering.

An overview of the results suggest that there is a certain pattern that follows the publication of the *same event* news. Even though it is reasonable to expect high similarity between text presenting a specific event, it can be safely assumed that certain news media shared the exact same text. These small groups of media may have based their articles in information drawn from the same sources. Surely, one of them posted the article first and the others followed, by simply copying the article. It is not certain regarding who copied from who, or if the credited sources are the true sources of the articles text.

The observations represent only an instance of a practice, that it is believed to be widely followed by many news media. Considering the impact this practice might have on quality news, the misinformation pitfall here is twofold: some pieces of information can be circulated intact to avoid miscommunicating them to readers. But what happens if the source of the article failed to validate facts or if the author of the article is biased? Surely the news media would not want their brand name to be harmed, by transferring misinformative articles. Readers on the other hand, can easily fall for propaganda or different forms of fake news, something that can greatly affect social coherence.

5 DISCUSSION AND FUTURE WORK

In this research the degree of similarity between news articles, published by different news media, was explored. In other words, we tried to inspect how differently online news media describe news events. To establish a quantified understanding of that similarity, three different datasets were created. The first contains unrelated event news, the second related event news, while the third, same event news articles.

The underlying goal of this task, is to define similarity limits for each set, to be able to put findings in context and finally compare them. Special focus is given in examining the similarity, between same events news articles. Even though results are diverse, in some cases they support our argument; high similarity scores reveal that many news articles, are *identical* and/or *similar*.

This fact suggests that news media fail to investigate events and describe them in an original way. Instead, groups of media are formed that circulate nearly identical piece of information to their readers.

The methodology followed can surely be improved. Even though *tfidf* and *cosine* similarity perform well, there are numerous representations and measurement algorithms to consider for feature tasks. Also, we would like expand this study by collecting more data, and including additional factors such as topic models, named entities, and comparison with other languages. This research contributes in current literature as there is no relevant research, that examines the similarities of news articles in Greek, which is a low-resource language.

ACKNOWLEDGEMENTS

This research is supported by the University of West Attica.

REFERENCES

Ahmed, S. S. and Hanif, U. (2020). News Recommendation Algorithm Based on Deep Learning. 06:9.

AL-Smadi, M., Jaradat, Z., AL-Ayyoub, M., and Jararweh, Y. (2017). Paraphrase identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic features. *Information Processing & Management*, 53(3):640–652.

Antonopoulos, N., Konidaris, A., Polykalas, S., and Lamprou, E. (2020a). Online Journalism: Crowdsourcing, and Media Websites in an Era of Participation. *Studies in Media and Communication*, 8(1):25.

Antonopoulos, N., Lamprou, E., Kiourexidou, M., Konidaris, A., and Polykalas, S. (2020b). Media Websites Services and Users Subscription Models for Online Journalism. *Media Watch*, 11(2).

Bright, J. and Nicholls, T. (2014). The Life and Death of Political News: Measuring the Impact of the Audience Agenda Using Online Data. *Social Science Computer Review*, 32(2):170–181. Publisher: SAGE Publications Inc.

Dhyani, D., Ng, W. K., and Bhowmick, S. S. (2002). A survey of Web metrics. *ACM Computing Surveys*, 34(4):469–503.

Fan, A., Doshi-Velez, F., and Miratrix, L. (2017). Prior matters: simple and general methods for evaluating and improving topic quality in topic modeling. Technical Report arXiv:1701.03227, arXiv. arXiv:1701.03227 [cs] type: article.

Feng, C., Khan, M., Rahman, A. U., and Ahmad, A. (2020). News recommendation systems - accomplishments, challenges & future directions. *IEEE Access*, 8:16702–16725.

Hamborg, F., Donnay, K., and Gipp, B. (2019). Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):391–415.

H.Gomaa, W. and A. Fahmy, A. (2013). A Survey of Text Similarity Approaches. *International Journal of Computer Applications*, 68(13):13–18.

Hu, Y., Xin, G., Song, R., Hu, G., Shi, S., Cao, Y., and Li, H. (2005). Title extraction from bodies of HTML documents and its application to web page retrieval. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '05*, page 250, Salvador, Brazil. ACM Press.

Huang, A.-L. (2008). Similarity measures for text document clustering.

Ibrahim, H., Darwish, K., and Abdel-sabor, A.-R. Automatic Extraction of Textual Elements from News Web Pages. page 5.

Leandros, N. and Papadopoulou, L. (2020). *Creative destruction in the Greek media landscape: New and alternative business models. In Vovou, I., Andonova, Y. and Kogan, A.F. (Eds), Proceedings of The Creative Contagion. The creative contagion. Media, industries, storytelling, communities, pp. 89-97.*, pages 89–97.

Llewellyn, C., Grover, C., and Oberlander, J. Summarizing Newspaper Comments. page 4.

Mohammadzadeh, H., Gottron, T., Schweiggert, F., and Heyer, G. (2012). TitleFinder: extracting the headline of news web pages based on cosine similarity and overlap scoring similarity. In *Proceedings of the twelfth international workshop on Web information and data management - WIDM '12*, page 65, Maui, Hawaii, USA. ACM Press.

Mozer, R., Miratrix, L., Kaufman, A. R., and Jason Anastasopoulos, L. (2020). Matching with Text Data: An Experimental Evaluation of Methods for Matching Documents and of Measuring Match Quality. *Political Analysis*, 28(4):445–468.

Nicholls, T. and Bright, J. (2018). Understanding news story chains using information retrieval and network clustering techniques. Technical Report arXiv:1801.07988, arXiv. arXiv:1801.07988 [cs] type: article.

Papadopoulou, L., Kavoulakos, K., and Avramidis, C. (2021). Intermedia Agenda Setting and Grassroots Collectives: Assessing Global Media's Influence on Greek News Outlets. *Studies in Media and Communication*, 9(2):12.

Po, L., Rollo, F., and Trillo Lado, R. (2017). Topic Detection in Multichannel Italian Newspapers. In Calì, A., Gorgan, D., and Ugarte, M., editors, *Semantic Keyword-Based Search on Structured Data Sources*, volume 10151, pages 62–75. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Prasad, D., Bisandu, D., and Liman, M. (2018). Clustering news articles using efficient similarity measure and n-grams. *International Journal of Knowledge Engineering and Data Mining*, 5:1.

Rubin, V., Conroy, N., and Chen, Y. (2015). *Towards News Verification: Deception Detection Methods for News Discourse*.

Soloshenko, A. N., Orlova, Y. A., Rozaliev, V. L., and Zaboleeva-Zotova, A. V. (2015). Establishing Semantic Similarity of the Cluster Documents and Extracting Key Entities in the Problem of the Semantic Analysis of News Texts. *Modern Applied Science*, 9(5):p246.

Spinde, T., Rudnitckaia, L., Mitrović, J., Hamborg, F., Granitzer, M., Gipp, B., and Donnay, K. (2021). Automated identification of bias inducing words in news articles using linguistic and context-oriented features. *Information Processing & Management*, 58(3):102505.

Tenenboim-Weinblatt, K. and Baden, C. (2018). Journalistic transformation: How source texts are turned into news stories. *Journalism*, 19(4):481–499.

Varlamis, I., Tsirakis, N., Poulopoulos, V., and Tsantilas, P. (2014). An automatic wrapper generation process for large scale crawling of news websites. In *Proceedings of the 18th Panhellenic Conference on Informatics - PCI '14*, pages 1–6, Athens, Greece. ACM Press.

Vieira, K., da Silva, A. S., Pinto, N., de Moura, E. S., Cavalcanti, J. M. B., and Freire, J. (2006). A fast and robust method for web page template detection and removal. In *Proceedings of the 15th ACM international conference on Information and knowledge management - CIKM '06*, page 258, Arlington, Virginia, USA. ACM Press.

Yang, Y., Carbonell, J., Brown, R., Pierce, T., Archibald, B., and Liu, X. (1999). Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems*, 14(4):32–43.

Yi, L., Liu, B., and Li, X. Eliminating Noisy Information in Web Pages for Data Mining. page 10.