NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Faculty of Communications, Media and Design

Alena Manuzina

Content-based Media Similarity and Clustering

MASTER’S THESIS
Field of study: 42.04.02. Journalism
Degree programme: Data Journalism

Reviewer
Junior Research Fellow
____
Ilya Karpov

Supervisor
Senior Lecturer, Ph.D
____
Ilya Makarov

Moscow 2020

Abstract

Algorithms for evaluating the similarity between texts or for clustering media outlets based on their content are included in many methods for studying media landscape or individual media outlets. They can help compare the coverage of several media, detect fake news and media bias, and therefore can be useful for researchers as well as practitioners: advertisers, media managers, and public relations specialists.

In this paper, three algorithms were applied to a dataset of 38 Russian-language media. The first one uses topic modelling to assess which share of coverage different topics get in media outlets; the second method is based on comparison of sets of people cited by media most often, and the third algorithm leverages information about the most typical words, i.e. those that are used in the media outlet much more often than in general language. All the three methods examine news articles at different levels (topics, quotations as an element of the text structure, and words).

The results showed that the third algorithm provided the most accurate division of media outlets depending on their niche, region of publishing, or position on the political spectrum. Along with comparison of the findings, the study provides new insights based on the data itself, suggesting potential areas of further research.

Table of Contents

1. Introduction

1.1. Significance of the Subject

The year 2020 will be long remembered for the coronavirus pandemic and economic crisis that is expected to be worse than that of 2008 and the others in almost a century[1]. Media industry was one of the first to suffer from the new crisis[2]as advertisers began to cut expenses and prevent their content from being shown on the pages dedicated to the pandemic. Since most media outlets live on advertising revenue, this may strongly influence the balance of power on the media market. When the crisis ends, advertisers will be likely to revise their contacts and thus will need tools for objective evaluation of media content and other characteristics.

At the same time, the pandemic caused a flood of rumors and fake news. Symptoms of the disease and its origin, actions of governments, miraculous medicines, forecasts – almost any information about the virus has been distorted or complemented with something that has proved to be untrue. Fake news is not a new problem, but during the pandemic its ability to influence people's behavior can be more dangerous than ever. Governments and activists as well as developers of social media, messengers, and search engines take measures to prevent the distribution of misinformation (e.g. Russian government introduced penalties, community service, or imprisonment as punishment for citizens and organizations[3]). Better understanding of the reasons behind the creation of fake news, patterns of its content, and properties of media outlets involved in its dissemination can help effectively struggle with misinformation.

One more tendency that demands thorough study of the rules governing the creation and distribution of media content is media bias. Not so fast-acting as fake news, bias is even harder to detect and counteract. Bias may affect the choice of events to cover and speakers to cite, the tone of voice, and the selection of details

[1] IMF, World Economic Outlook, April 2020 https://www.imf.org/en/Publications/WEO/Issues/2020/04/14/weo-april-2020, 'Great Recession showed countries can't fight the coronavirus economic crisis alone' https://www.weforum.org/agenda/2020/04/covid-19-coronavirus-economic-crisis-great-recession/

[2] https://wwd.com/business-news/media/coronavirus-business-impact-advertising-media-recession-1203559395/

[3] https://rg.ru/2020/03/31/vvoditsia-nakazanie-za-rasprostranenie-fejkov-o-koronaviruse.html

and issues for the story. It can occur unwittingly as a reflection of the author's beliefs but still change the readers' perception of the event.

Computational media analysis offers tools for accomplishing tasks in all three areas mentioned above. An important step in many of the algorithms is calculating similarity between texts and entire media. For advertisers it can help find new media outlets to collaborate with, particularly in the case of native advertising that strongly depends on the media content. Various strategies for identification of bias and fake news use similarity to group suspect media or articles.

1.2. Aim, Objectives, and Research Question

The object of the research is text clustering methodology. The subject is media content clustering and similarity search. Both conventional methods for evaluating similarity between texts and algorithms created specially for media analysis will be considered.

The aim of the study is to find an algorithm for dividing media outlets into meaningful clusters. The research questions is what algorithm(s) can be considered as the best for media clustering.

To achieve the goal and to answer the question the following objectives should be achieved:

- to examine existing methods of text clustering
- to list approaches suitable for clustering media outlets
- to test the algorithms on collected data
- to compare the methods and select the best one
- to interpret the results of such clustering and its applicability to similarity-based search.

Since clustering is an unsupervised method, we need a way to evaluate its performance. First, we expect that all clustering methods will be able to group niche media and separate them from socio-political ones. Second, we will check if any of the algorithms can group media of one type, e.g. news agencies or TV channels, into clusters and thus differentiate media types. Third, it is interesting to see if these algorithms can tell media from the opposite sides of the political spectrum (in Russia, however, media can not be so easily divided into parties as, for instance, in the USA). Fourth, we would like them to isolate two regional media, covering news of Moscow and Saint-Petersburg.

Code for all the algorithms is available at Github[4].

1.3. Overview of the Thesis

The paper is organized as follows: the second section presents previous studies describing algorithms that will be used further, the third one is devoted to preprocessing methods, and the last section examines in more detail the known algorithms for analyzing relations between media. The third and the fourth sections include technical details about how to implement the algorithms.

[4] https://github.com/manalyona/Content-based_Media_Similarity_and_Clustering

2. Related Works

Text analysis is actively used in a wide range of areas: apart from studying media and online networks as such they include medicine (Denecke & van Harmelen, 2019, Wang et al., 2019) and city management (Liu & Jansson, 2017), monitoring of emergency situations (Rogstadius, 2013) and financial event detection (Qian et al., 2019); it is applied for comparing patents (Helmers et al., 2019, Shahmirzadi et al., 2019) and medical histories (Kim & Meystre, 2020).

Methods suitable for media texts analysis can be divided into two groups: the first one includes general purpose algorithms, while the second one – methods designed specifically for media studies.

2.1. Methods of Text Clustering

A number of algorithms for computational media analysis make extensive use of Natural Language Processing methods, particular topic modelling, keyword extraction and named entities recognition. Detailed catalogue of methods used for text classification is available in (Mirończuk & Protasiewicz, 2018). Apart from listing techniques for textual data representation, they also provide an overview of methods for dimensionality reduction, training classifiers and evaluating the results, that one may find useful while implementing the algorithms. The authors of one more review (Prasetya et al., 2018) grouped algorithms into string-based, corpus-based, knowledge-based, and hybrid similarities, and lexical and semantic similarity approaches.

The comparison of several widely used methods for assessing texts similarity (tf-idf with two modifications – Latent Semantic Indexing, LSI and Document to Vector, D2V) is given in (Shahmirzadi et al., 2019) with the conclusion that in most cases computational complexity of the modified algorithms was not justified by the results. One more comparison with the same conclusion (except for very small datasets) was made in (Dzisevic and Sesok, 2019).

Almost the same results were obtained in (Sitikhu et al., 2019) on the dataset of short news articles. After comparing cosine similarity with tf-idf vectors, cosine similarity with word2vec vectors, and soft cosine similarity with word2vec vectors they have found that all three algorithms gave almost the same result (with the highest accuracy achieved with tf-idf vectors). Taking into account the relative complexity of the methods, the simplest of them seems like a reasonable choice.

The review of the complicated methods using neural networks is given in (Zhou et al., 2020). It covers commonly known models like word2vec, RNN and BERT as well as not so widely used and more specialized architectures.

All these methods can be applied to any set of texts, so they do not consider typical features of news texts or patterns of distribution, citation or consuming of media content. They are rather useful on their own and, in addition, formed the basis for more specialized methods. However, algorithms developed for studying media can add a lot of new information about selected media outlets as well as the whole system of national or global media.

2.2. Algorithms for Calculating Media Similarity

Studies that propose algorithms for media analysis can be grouped based on the type of the methods and characteristics of media they pay attention to. Works of the first group describe links between media outlets and their influence on each other, so they are focused on any information about the media except its content.

In (Aires et al., 2019) a new method was proposed to evaluate media bias based on the closeness between media outlet in question and the one known as biased. Similarity in this work can be considered as a belonging to one of the clusters in link-based graph. However, this study uses a database of manually checked news, and this limits the approach: only 'hot' topics (like politics) and a few regions (those having a fact-checking community) can be examined.

One more research of this group (Vargo et al, 2018) studies links between media which are understood as influence: in terms of graph theory, media outlet A has a directed edge to media outlet B if A caused B to cover particular topic (no matter in the same tone as A or not). So media can be grouped based on how influential they are or whom they are influenced by.

The study (Álvarez-Carmona et al., 2018) is dedicated to the algorithms for detection of plagiarism, especially its complex cases, such as paraphrase plagiarism. They can be useful for dividing media outlets into groups of influencers and followers, where the latter are likely to reprint news of the former.

One more approach is widely used in recommender systems (Chaudhary & Anupama, 2020, Joris et al., 2020, Babanejad et al., 2020): two media outlets are similar if they are read by the same group of people.

The second group of studies includes research focused on the content of articles as the main feature of the media and considers it at three levels. At the macro level articles can be represented as a list of topics or events they cover.

The method for topic modelling based on k-means clustering algorithm was proposed in (Rashid et al., 2020). It is claimed to show better results than LDA (Latent Dirichlet Allocation) and LSA (Latent Semantic Analysis) on the dataset of Reuters and BBC news. Interest in the same set of topics and close to equal share of the topics in several media outlets coverage may be applied for assessing similarity between them. In the work (Nimark and Pitschner, 2019) topic modelling was used to examine specialization of media outlets.

The study (Rappaz et al., 2019) suggests that temporal dimension can significantly enrich our vision of relations between media. In this paper similarity between two media outlets is considered as a share of event that both of them, or neither of them, cover. Dynamic embedding method lets the authors trace purchases of media outlet based on the changes in its content. The research (Zhou et al., 2018) suggested a neural network based method for grouping articles into storylines.

The methods for plagiarism search proposed in (Álvarez-Carmona et al., 2018) can also be used at this level to identify media outlets that often reprint texts without proper citation, to compare media by the level of uniqueness of their content or to assess the influence of press-releases.

The studies of the middle level examine text components like quotations.

The authors of the paper (Niculae et al., 2015) paid attention to quotes added to new articles. The idea is that the choice of persons, particular politicians, to cite can reflect party affiliation of the media outlet and its bias. Thus, media can be called similar if their articles contain quotations of the same group of people more often than all the others.

At the micro level the main features are frequencies of words and phrases, including keywords. These studies rely heavily on algorithms for comparing texts in general, regardless of their source (i.e. mentioned in the beginning of the section).

An approach described in (Mikhina & Trifalenkov, 2018) combines conventional tf-idf and cosine similarity with graph methods to measure the closeness of the articles. Using the information about words contained in the text the authors constructed a graph and detected clusters. The study of breaking news and influence of media outlets on each other (Varlamis & Hilliard, 2017) suggests that measuring closeness between texts (using the same tf-idf and cosine similarity) can help to trace dissemination of the content, at least breaking news that is usually reprinted by a number of media before any new information is available.

The algorithm used in (Potthast et al., 2018) applied stylometry to the search for hyperpartisan news. The authors found that frequency of the words can help distinguish left- and right-wing news from mainstream ones and both of them from satirical articles. Similar approach was suggested in (Barrón-Cedeño et al., 2019) for assessing level of propagandistic content in articles based on writing style and readability.

A large set of studies describe algorithms for keyword extraction, and they can be adopted for measuring news similarity. A news recommender system using keywords is described in (Wang et al., 2018). Keywords extraction based on graph methods discussed in (Anjali et al., 2019, Vega-Oliveros et al., 2019). Comparison of several methods is given in (Zhang, 2020, Nasar et al., 2019, Yang et al., 2018).

For the study we selected three algorithms examining texts at different levels. One of them uses simple topic modelling, the second one compares sets of people cited most often, and the third algorithm is based on word frequencies.

3. Data Gathering and Preprocessing

3.1. Collecting Texts

The data used in the present study include news and articles from 38 Russian-language media outlets listed as the most cited by ‘Medialogia’[5]. The sample includes news agencies, printed newspapers, TV and radio channels, and online media (the type of media was defined based on Medialogia classification). Most of media outlets cover mainly social and political news, so we included three groups of niche media writing about science, sport, and auto. Media covering mostly regional problems were excluded from the sample, except for two media writing about news of Moscow and St. Petersburg – they were added to test algorithms ability to detect such differences in topics. Both state-owned and opposition media (or, at least, considered as that) were studied.

The time period covers the year 2019 and the first two month of 2020.

Data were collected using Python libraries `Requests`, `BeautifulSoup` and `Selenium`, choice was made based on the media site structure and functionality of its ‘Search’ page. For example, links on articles of Rossiyskaya Gazeta were gathered with `Selenium`, while for Gazeta.Ru it was enough to apply `Requests` and `BeautifulSoup`. When available (like in the cases of Vedomosti and Komsomol’skaya Pravda), API of the GDELT project[6] was used to collect links. Undocumented API of media websites (via `.get` or `.post` queries in `Requests`) was used where possible (Meduza, Postnauka, The Bell). In some cases (Novaya Gazeta, Izvestiya) search was only allowed for meaningful words, so a set of words included in the majority of news (like ‘told’, ‘said’ or ‘reported’) or names of the months was used, and search was repeated for each of them.

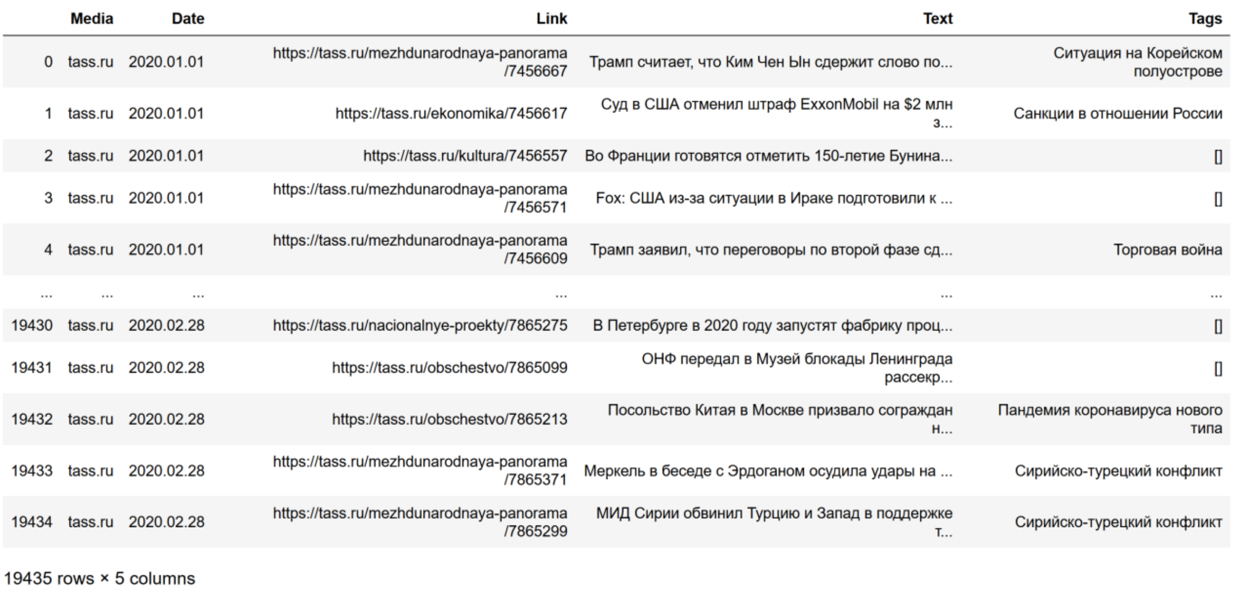
[5] https://www.mlg.ru/ratings/media/federal/7130/

[6] https://gdelt.github.io/#api=doc&query=&contentmode=ArtList&maxrecords=75&timespan=1d

For most media outlets two functions were written: the first one collected links on the articles published on the given date, the second one downloaded title, text, tags and so on from the given link.

For every article, we saved the name of the media outlet, link, date of publication, and text of the article. Some media outlets also provide a topic (like ‘politics’ or ‘sport’), while others – a set of keywords, usually reflecting events or naming actors and places. Collected data were saved in files of two types. Links were collected in dictionaries like `‘date’:[list_of_links]` and saved in `.json` format, while dates, links, texts and other information were organized in `DataFrames` (`Pandas` library) and saved in `.csv` format.

Table 1. An example of `DataFrame` with the texts of the articles (TASS.ru, for the year 2020)



	Media	Date	Link	Text	Tags
0	tass.ru	2020.01.01	https://tass.ru/mezhdunarodnaya-panorama/7456667	Трамп считает, что Ким Чен Ын сдержит слово по...	Ситуация на Корейском полуострове
1	tass.ru	2020.01.01	https://tass.ru/ekonomika/7456617	Суд в США отменил штраф ExxonMobil на $2 млн з...	Санкции в отношении России
2	tass.ru	2020.01.01	https://tass.ru/kultura/7456557	Во Франции готовятся отметить 150-летие Бунина...	[]
3	tass.ru	2020.01.01	https://tass.ru/mezhdunarodnaya-panorama/7456571	Fox: США из-за ситуации в Ираке подготовили к ...	[]
4	tass.ru	2020.01.01	https://tass.ru/mezhdunarodnaya-panorama/7456609	Трамп заявил, что переговоры по второй фазе сд...	Торговая война
...	...	...	...	...	...
19430	tass.ru	2020.02.28	https://tass.ru/nacionalnye-proekty/7865275	В Петербурге в 2020 году запустят фабрику проц...	[]
19431	tass.ru	2020.02.28	https://tass.ru/obschestvo/7865099	ОНФ передал в Музей блокады Ленинграда рассекр...	[]
19432	tass.ru	2020.02.28	https://tass.ru/obschestvo/7865213	Посольство Китая в Москве призвало сограждан н...	Пандемия коронавируса нового типа
19433	tass.ru	2020.02.28	https://tass.ru/mezhdunarodnaya-panorama/7865371	Меркель в беседе с Эрдоганом осудила удары на ...	Сирийско-турецкий конфликт
19434	tass.ru	2020.02.28	https://tass.ru/mezhdunarodnaya-panorama/7865299	МИД Сирии обвинил Турцию и Запад в поддержке т...	Сирийско-турецкий конфликт

19435 rows × 5 columns

3.2. Data Preprocessing

Two Python libraries were used on the stage of data preprocessing: `re` and `pymorphy2`. The former allows to split text into words and exclude punctuation at a time, the letter is a popular tool for morphological analysis.

For the method based on analysis of quotations two queries (using `re` module) were written to gather small fragments of texts that contain quote and one or two sentences around it (depending on the position of the quote). Names that can potentially refer to the speaker were detected and their morphological characteristics were examined with the help of `MorphAnalyser` from `pymorphy2` library. This method is not very precise since it does not include coreference resolution and the resulting rating of speakers includes both surnames and position (like ‘president’ or ‘head’) of the people. However, this information may still give valuable insights, as it will be discussed in the fourth section.

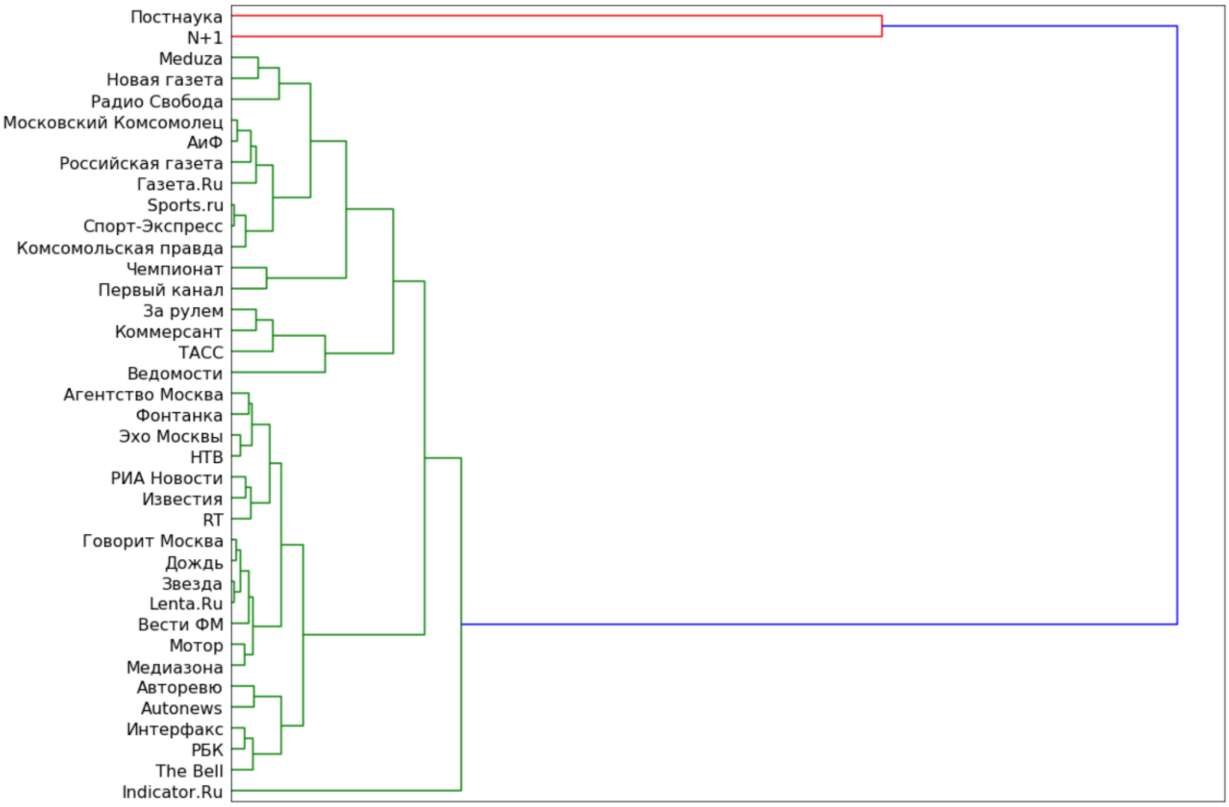
For methods that use word frequencies, articles of each media were united into one `.txt` file, then divided into tokens using `re` module. Punctuation as well as stop-words (like ‘we’, ‘that’, ‘off’ and ‘the’, the list is available in `nltk` library) were excluded from the list. Then words were counted using `Counter` from `collections` module and the rest of preprocessing was conducted with these unique words. Prepositions, conjunctions, particles, interjections and pronouns were identified using `MorphAnalyser` and excluded on this stage. All the other tokens were put in their initial form and counted once again. These numbers were divided by the length of the whole text. As a result, we got the list of words in initial form with their frequencies in particular media outlets.

4. Media Similarity and Clustering

For the comparison we selected three methods that are at the different levels (according to classification, proposed in Related Works section). They are relatively simple and do not require large sets of (manually) labelled data, that may be too time-consuming and labour-intensive to collect.

The first one applies clustering twice: to group articles into topics and to cluster media outlets based on the proportion of the topics in their coverage. To detect topics, texts of the articles were transformed into vectors with `CountVectorizer` and grouped using `AgglomerativeClustering` from `sklearn` library. To reduce the weight of the model 1000 articles from each media were taken randomly and the number of features was limited to 1000. The resulting matrix provided data for dendrogram from `scipy` library (Figure 1).

Figure 1. Dendrogram built based on the proportion of different topics in media outlets coverage



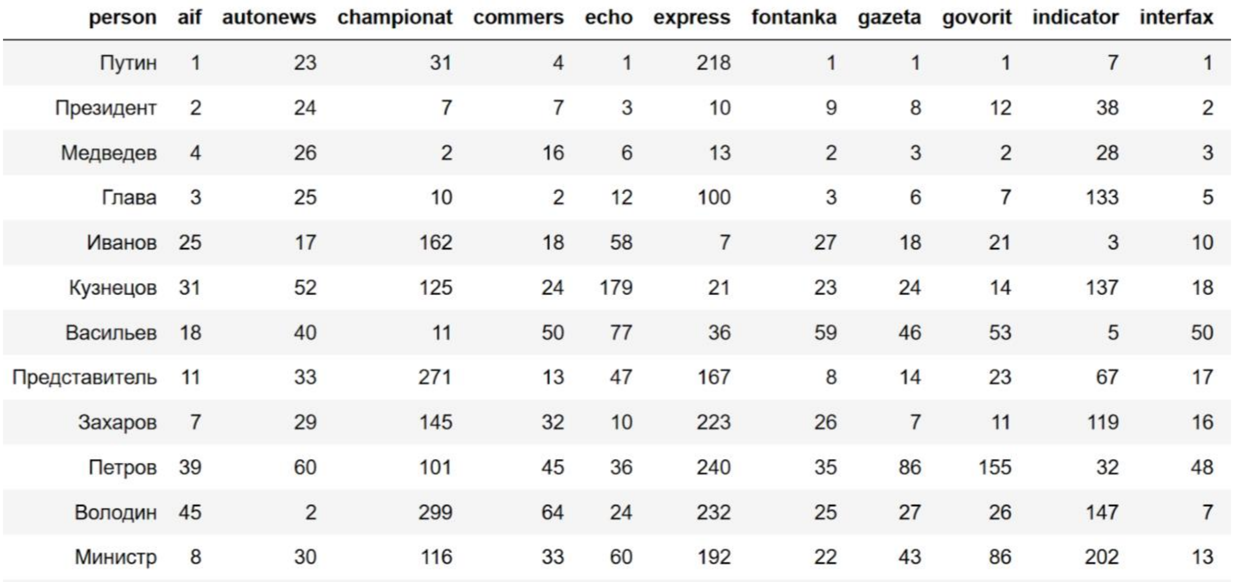
The results are not as good as expected: science media are isolated, but separated into two groups, sports and auto media are dispersed throughout the dendrogram as well as news agencies. Absolutely different media outlets (like Mediazona and Motor, or Championat and the First Channel) are placed next to each other.

The second method was inspired by the work (Niculae et al., 2015). The study is devoted to how often, how quickly and how extensively media outlets cite speeches of Barack Obama. We decided to check whether the choice of whom to cite (evaluated as a frequency of quotations of various people) can give us any information about media, e.g. it's position on the political spectrum.

List of cited people obtained after preprocessing was put into `Counter` and the share for each person in each media outlet was calculated. All data about people who entered top-100 in at least one media (n = 1102) were gathered in one table.

Due to the simplicity of preprocessing, both surnames and positions of cited people were included, and surnames were put into initial form, so male and female surnames were not separated. There are solutions for both these issues, however, they fall outside the scope of the present study.

Table 2. A list of people cited most often in all selected media with their rank for several media outlets. Note: all calculation were done with frequencies, ranks are given for illustrative purposes

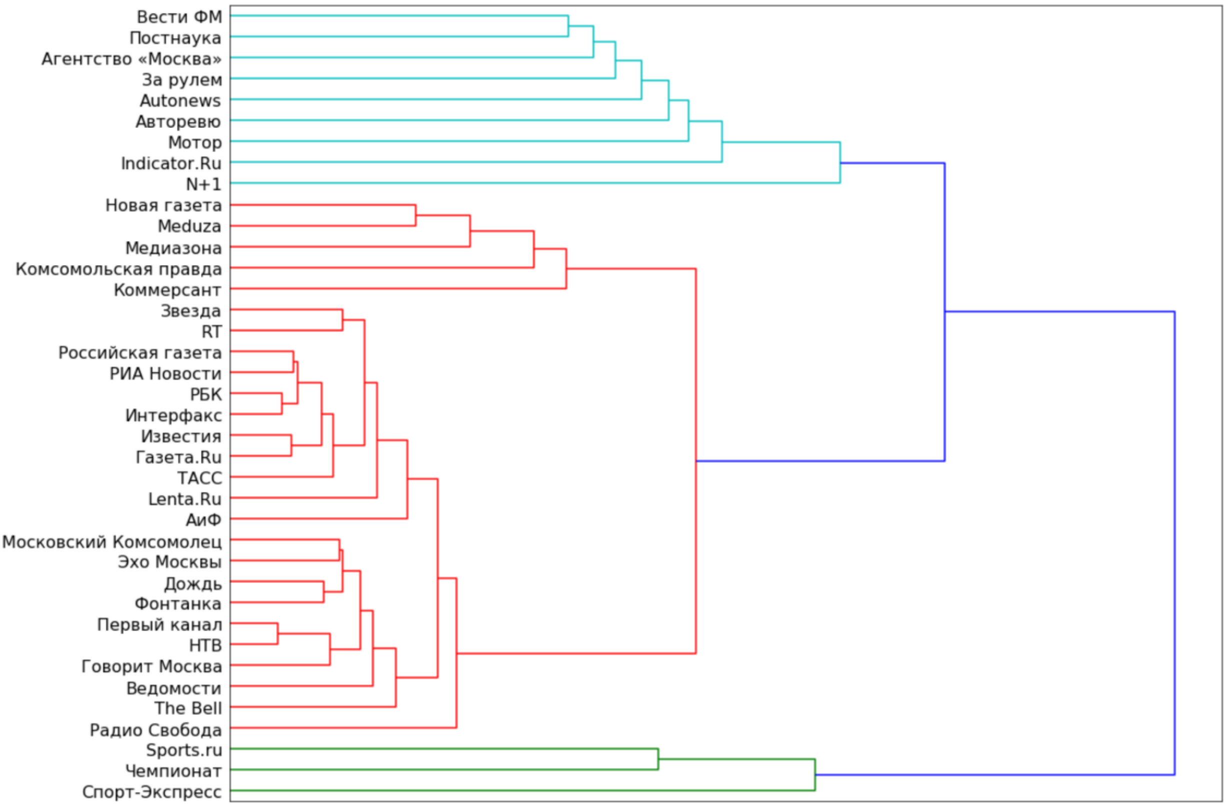


person	aif	autonews	championat	commers	echo	express	fontanka	gazeta	govorit	indicator	interfax
Путин	1	23	31	4	1	218	1	1	1	7	1
Президент	2	24	7	7	3	10	9	8	12	38	2
Медведев	4	26	2	16	6	13	2	3	2	28	3
Глава	3	25	10	2	12	100	3	6	7	133	5
Иванов	25	17	162	18	58	7	27	18	21	3	10
Кузнецов	31	52	125	24	179	21	23	24	14	137	18
Васильев	18	40	11	50	77	36	59	46	53	5	50
Представитель	11	33	271	13	47	167	8	14	23	67	17
Захаров	7	29	145	32	10	223	26	7	11	119	16
Петров	39	60	101	45	36	240	35	86	155	32	48
Володин	45	2	299	64	24	232	25	27	26	147	7
Министр	8	30	116	33	60	192	22	43	86	202	13

Indeed, even the ratings themselves (Table 2) can give some information about the media. For example, in Novaya Gazeta, which often covers high-profile legal cases and violation of prisoners' rights, second most cited ‘person’ is a lawyer (адвокат), and the seventh – a prosecutor (прокурор). In the rating of The Bell, that writes mostly about business and finances, three of the five most cited people are influential economists (former Minister for Economic Development, Minister of Finance and Minister of Economics and Trade). Putin is #1 in 24 out of 38 media outlets.

Before clustering all values were normalized, i.e. divided by the highest number for the observation, so that values stayed within the range from zero to one. Dendrogram showing relative similarity between groups of media was drawn with the help of `dendrogram` from `scipy` library (Figure 2).

Figure 2. Dendrogram of studied media constructed based on the list of most cited people

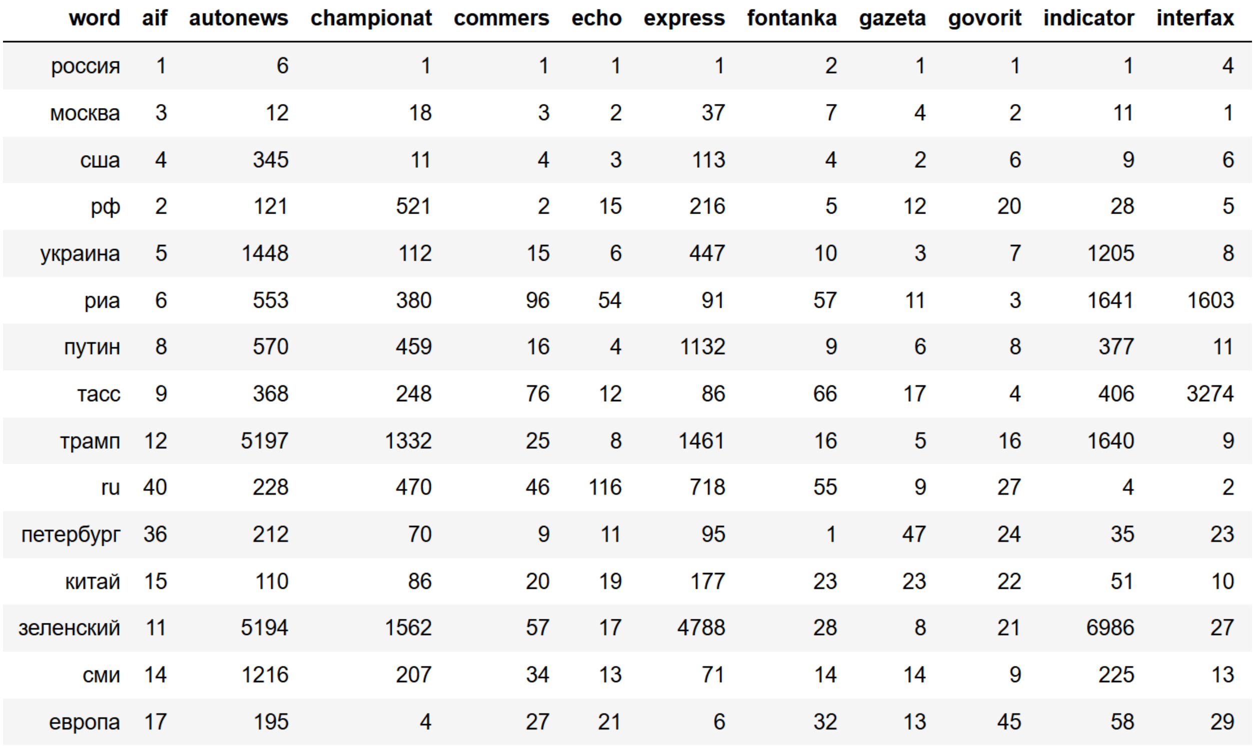


The first two divisions are predictable: the largest distance to the rest of media outlets have sports media, after them – a combined group of science and auto media. The algorithm correctly isolated niche media from all the rest and sports media from other thematic, however, it could not distinguish media writing about science and auto, and added Vesti FM and ‘Moscow’ Agency to them. Novaya Gazeta, Meduza and Mediazona have a reputation of liberal media, but they are in the same cluster as Komsomol’skaya Pravda and Kommersant. Three news agencies are quite close to each other, but Fontanka and ‘Moscow’ Agency with their regional content were not isolated. The same situation with TV channels: the First Channel, NTV and Dozhd’ are in the same group, but Zvezda lies rather far from them.

The third algorithm replicate method described in (Mikhina & Trifalenkov, 2018). The authors represented each text as a list of words, selected by their relative frequency (defined as tf-idf), then calculated cosine similarity between media outlets, constructed a graph and searched for communities, i.e. groups of similar media.

After the preprocessing stage we obtained a list of words and their frequencies for each media outlet. To select words that can represent the media we compared these frequencies with data from frequency dictionary of the Russian language[7]. As the total number of words contained in the texts is large, for further analysis we selected words that enter top-1000 at least in one media outlet. These words (n = 9098) with frequencies for all media were gathered in a table (Table 3).

Table 3. The most ‘typical’ words among all the studied media with their ranks. Note: all calculation were done with frequencies, ranks are given for illustrative purposes



word	aif	autonews	championat	commers	echo	express	fontanka	gazeta	govorit	indicator	interfax
россия	1	6	1	1	1	1	2	1	1	1	4
москва	3	12	18	3	2	37	7	4	2	11	1
сша	4	345	11	4	3	113	4	2	6	9	6
рф	2	121	521	2	15	216	5	12	20	28	5
украина	5	1448	112	15	6	447	10	3	7	1205	8
риа	6	553	380	96	54	91	57	11	3	1641	1603
путин	8	570	459	16	4	1132	9	6	8	377	11
тасс	9	368	248	76	12	86	66	17	4	406	3274
трамп	12	5197	1332	25	8	1461	16	5	16	1640	9
ru	40	228	470	46	116	718	55	9	27	4	2
петербург	36	212	70	9	11	95	1	47	24	35	23
китай	15	110	86	20	19	177	23	23	22	51	10
зеленский	11	5194	1562	57	17	4788	28	8	21	6986	27
сми	14	1216	207	34	13	71	14	14	9	225	13
европа	17	195	4	27	21	6	32	13	45	58	29

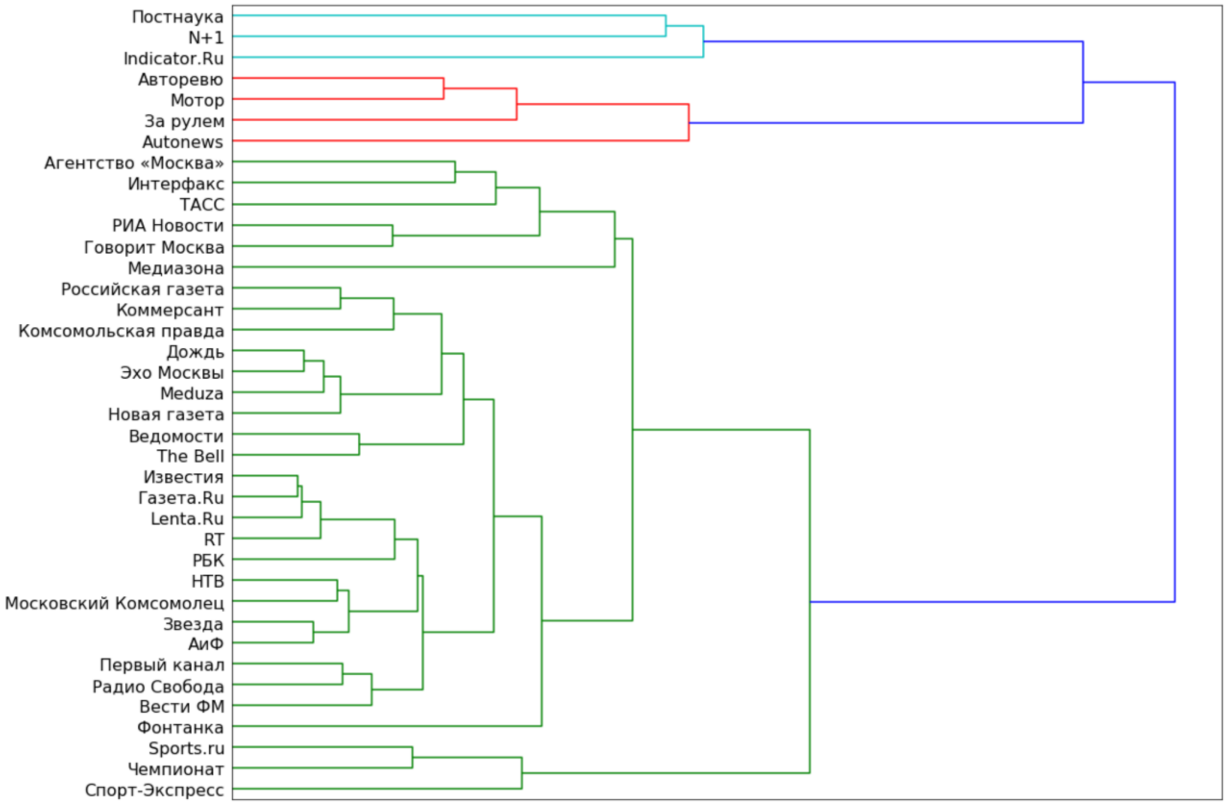
The majority of words in this rating are geographical names and names of nation leaders. We can also notice that among these ‘typical’ words there are two

[7] Frequency Dictionary of the Modern Russian Language (based on the materials of the Russian National Corpus). Moscow: Azbukovnik, 2009. http://dict.ruslang.ru/freq.php

titles of news agencies – RIA and TASS, they are often cited as the source of information.

Based on this matrix a dendrogram (Figure 3) was constructed.

Figure 3. Dendrogram built based on the most ‘typical’ words for each media



This method showed the best results in separating thematic media, all three groups are clearly visible on the dendrogram. News agencies successfully gathered together. Liberal/opposition media (Dozhd’, Echo, Meduza and Novaya Gazeta) were grouped together too. Two of three media writing about business and finances (Vedomosti and The Bell) were placed near each other.

The algorithm isolated the regional media, Fontanka, but it put ‘Moscow’ Agency in the same group as federal media. One more outlier, Mediazona, was isolated even earlier than Fontanka. However, this method did not group radio and TV channels.

5. Conclusion

In our study, three methods for clustering media outlets were tested on the collected dataset of Russian-language news. These methods consider media content at three levels: macro (topics and events covered), medium (quotations), and micro (words). The first one compares the space devoted to each of the selected topics. The second method uses information about the people most often cited by the media. And the third algorithm leverages the list of the most ‘typical’ words (compared with general language).

Four criteria were suggested to evaluate the quality of the algorithms – we assessed their ability to a) divide media into the general-purpose media and the thematic ones, b) separate TV and radio channels, news agencies, and online media from each other, c) distinguish between the media from the opposite sides of the political spectrum, and d) isolate two regional media.

Three methods showed nearly the same quality of clustering. Two of them (except for the first one) were able to isolate niche media from the rest, but the second algorithm did not distinguish science media from the sports ones. The third algorithm managed to separate one of the two regional media (Fontanka); however, none of the methods could single out the other, ‘Moscow’ Agency (this may be explained by the sustained attention of the Russian media to the Moscow news in general). TV and radio channels were not separated by any of the algorithms; however, the third method managed to cluster news agencies. It was also the most accurate in isolating groups of opposite and business media.

The third method proved to be the best for clustering media outlets, and in contrast to the second one it can be applied to various types of articles, including those without quotations. One more advantage of the method is that it is computationally simple and does not require any labelled data.

The fact that the algorithms were not perfect for clustering media outlets leaves room for improvement. Some of them may be incorporated in preprocessing methods to clean the data more thoroughly, the others – in more complicated clustering algorithms. Recent and future advances in Natural Language Processing can increase the accuracy of the algorithms and contribute to academic and industry media studies.

References

Aires, V. P., Nakamura, F. G., & Nakamura, E. F. (2019). A link-based approach to detect media bias in news websites. The Web Conference 2019 – Companion of the World Wide Web Conference, WWW 2019. https://doi.org/10.1145/3308560.3316460

Álvarez-Carmona, M. A., Franco-Salvador, M., Villatoro-Tello, E., Montes-Y-Gómez, M., Rosso, P., & Villaseñor-Pineda, L. (2018). Semantically-informed distance and similarity measures for paraphrase plagiarism identification. *Journal of Intelligent and Fuzzy Systems*, *34*(5), 2983–2990. https://doi.org/10.3233/JIFS-169483

Anjali, S., Meera Nair, M., & Thushara, M. G. (2019). A graph based approach for keyword extraction from documents. *2019 2nd International Conference on Advanced Computational and Communication Paradigms, ICACCP 2019*. https://doi.org/10.1109/ICACCP.2019.8882946

Babanejad, N., Agrawal, A., Davoudi, H., An, A., & Papagelis, M. (2020). Leveraging emotion features in news recommendations. *CEUR Workshop Proceedings*.

Barrón-Cedeño, A., Jaradat, I., Da San Martino, G., & Nakov, P. (2019). Proppy: Organizing the news based on their propagandistic content. *Information Processing and Management*, *56*(5), 1849–1864. https://doi.org/10.1016/j.ipm.2019.03.005

Chaudhary, S., & Anupama, C. G. (2020). Recommendation System for Big Data Software Using Popularity Model and Collaborative Filtering. *Advances in Intelligent Systems and Computing*. https://doi.org/10.1007/978-981-15-0199-9_47

Denecke, K., & van Harmelen, F. (2019). Recent advances in extracting and processing rich semantics from medical texts. In *Artificial Intelligence in Medicine*. https://doi.org/10.1016/j.artmed.2018.07.004

Dzisevic, R., & Sesok, D. (2019). Text Classification using Different Feature Extraction Approaches. *2019 Open Conference of Electrical, Electronic and Information Sciences, EStream 2019 – Proceedings*. https://doi.org/10.1109/eStream.2019.8732167

Helmers, L., Horn, F., Biegler, F., Oppermann, T., & Müller, K. R. (2019). Automating the search for a patent's prior art with a full text similarity search. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0212103

Joris, G., Colruyt, C., Vermeulen, J., Vercoutere, S., De Grove, F., Van Damme, K., De Clercq, O., Van Hee, C., De Marez, L., Hoste, V., Lievens, E., De Pessemier, T., & Martens, L. (2020). News diversity and recommendation systems: Setting the interdisciplinary scene. *IFIP Advances in Information and Communication Technology*. https://doi.org/10.1007/978-3-030-42504-3_7

Kim, Y., & Meystre, S. M. (2020). Ensemble method-based extraction of medication and related information from clinical texts. *Journal of the American Medical Informatics Association : JAMIA*. https://doi.org/10.1093/jamia/ocz100

Liu, S., & Jansson, P. (2017). City event detection from social media with neural embeddings and topic model visualization. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*. https://doi.org/10.1109/BigData.2017.8258430

Mikhina, E. K., & Trifalenkov, V. I. (2018). Text clustering as graph community detection. Procedia Computer Science, 123, 271–277. https://doi.org/10.1016/j.procs.2018.01.042

Mirończuk, M. M., & Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications*, *106*, 36–54. https://doi.org/10.1016/j.eswa.2018.03.058

Nasar, Z., Jaffry, S. W., & Malik, M. K. (2019). Textual keyword extraction and summarization: State-of-the-art. *Information Processing & Management*, *56*(6), 102088. https://doi.org/https://doi.org/10.1016/j.ipm.2019.102088

Niculae, V., Suen, C., Zhang, J., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2015). QUOTUS: The structure of political media coverage as revealed by quoting patterns. WWW 2015 – Proceedings of the 24th International Conference on World Wide Web. https://doi.org/10.1145/2736277.2741688

Nimark, K. P., & Pitschner, S. (2019). News media and delegated information choice. *Journal of Economic Theory*. https://doi.org/10.1016/j.jet.2019.02.001

Qian, Y., Deng, X., Ye, Q., Ma, B., & Yuan, H. (2019). On detecting business event from the headlines and leads of massive online news articles. *Information Processing and Management*. https://doi.org/10.1016/j.ipm.2019.102086

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. ACL 2018 – 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers). https://doi.org/10.18653/v1/p18-1022

Prasetya, D. D., Wibawa, A. P., & Hirashima, T. (2018). The performance of text similarity algorithms. *International Journal of Advances in Intelligent Informatics*, *4*(1), 63–69. https://doi.org/10.26555/ijain.v4i1.152

Rappaz, J., Bourgeois, D., & Aberer, K. (2019). A dynamic embedding model of the media landscape. The Web Conference 2019 – Proceedings of the World

Wide Web Conference, WWW 2019. https://doi.org/10.1145/3308558.3313526

Rashid, J., Shah, S. M. A., & Irtaza, A. (2020). An Efficient Topic Modeling Approach for Text Mining and Information Retrieval through K-means Clustering. *Mehran University Research Journal of Engineering and Technology*. https://doi.org/10.22581/muet1982.2001.20

Rogstadius, J., Vukovic, M., Teixeira, C. A., Kostakos, V., Karapanos, E., & Laredo, J. A. (2013). CrisisTracker: Crowdsourced social media curation for disaster awareness. *IBM Journal of Research and Development*. https://doi.org/10.1147/JRD.2013.2260692

Shahmirzadi, O., Lugowski, A., & Younge, K. (2019). Text similarity in vector space models: A comparative study. *Proceedings – 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019*, 659–666. https://doi.org/10.1109/ICMLA.2019.00120

Sitikhu, P., Pahi, K., Thapa, P., & Shakya, S. (2019). A Comparison of Semantic Similarity Methods for Maximum Human Interpretability. *International Conference on Artificial Intelligence for Transforming Business and Society, AITB 2019*. https://doi.org/10.1109/AITB48515.2019.8947433

Vargo, C. J., Guo, L., & Amazeen, M. A. (2018). The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. New Media and Society. https://doi.org/10.1177/1461444817712086

Varlamis, I., & Hilliard, D. F. (2017). Finding influential sources and breaking news in news media using graph analysis techniques. *International Journal of Web Engineering and Technology*. https://doi.org/10.1504/IJWET.2017.086449

Vega-Oliveros, D. A., Gomes, P. S., E. Milios, E., & Berton, L. (2019). A multi-centrality index for graph-based keyword extraction. *Information Processing and Management*. https://doi.org/10.1016/j.ipm.2019.102063

Wang, X., Wang, R., Bao, Z., Liang, J., & Lu, W. (2019). Effective medical archives processing using knowledge graphs. *SIGIR 2019 - Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://doi.org/10.1145/3331184.3331350

Wang, Z., Hahn, K., Kim, Y., Song, S., & Seo, J. M. (2018). A news-topic recommender system based on keywords extraction. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-017-5513-0

Yang, L., Li, K., & Huang, H. (2018). A new network model for extracting text keywords. *Scientometrics*. https://doi.org/10.1007/s11192-018-2743-5

Zhang, Y., Tuo, M., Yin, Q., Qi, L., Wang, X., & Liu, T. (2020). Keywords extraction with deep neural network model. *Neurocomputing*, *383*, 113–121. https://doi.org/https://doi.org/10.1016/j.neucom.2019.11.083

Zhou, M., Duan, N., Liu, S., & Shum, H. Y. (2020). Progress in Neural NLP: Modeling, Learning, and Reasoning. *Engineering*, *6*(3), 275–290. https://doi.org/10.1016/j.eng.2019.12.014

Zhou, D., Guo, L., & He, Y. (2018). *Neural Storyline Extraction Model for Storyline Generation from News Articles*. https://doi.org/10.18653/v1/n18-1156