# Analyzing the Evolution of Disinformation Content on Facebook – a Pilot Study

Elena Tuparova[1,2], Andrey Tagarev[2], Nikola Tulechki[2], Svetla Boytcheva[2,3]

[1] *Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski", Sofia, Bulgaria*
[2] *Sirma AI (trading as Ontotext), Bulgaria*
[3] *Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Sofia, Bulgaria*

**Abstract**

Disinformation spread on social media generates a truly massive amount of content on a daily basis, much of it not quite duplicated but repetitive and related. In this paper, we present an approach for clustering social media posts based on topic modeling in order to identify and formalize an underlying structure in all the noise. This would be of great benefit for tracking evolving trends, analyzing large-scale campaigns, and focusing efforts on debunking or community outreach.

The steps we took in particular include harvesting through CrowdTangle huge collection of Facebook posts explicitly identified as containing disinformation by debunking experts, following those links back to the people, pages and groups where they were shared then collecting all posts shared on those channels over an extended period of time. This generated a very large textual dataset which was used in the topic modeling experiments attempting to identify the larger trends in the available data. Finally, the results were transformed and collected in a Knowledge Graph for further study and analysis. Our main goal is to investigate different trends and common patterns in disinformation campaigns, and whether there exist some correlations between some of them. For instance, for some of the most recent social media posts related to COVID-19 and political situation in Ukraine.

**Keywords**

Disinformation, Social networks, Natural Language Processing, Topic modelling, Big data management and analysis



## 1 Introduction

While the distribution of misleading information and propaganda has always been a challenge facing society, recent years have drawn a lot more attention to this issue thanks to the way social media has vastly increased the quantity and speed of transmission of such information. Newly popular terms such as "fake news" and "disinformation" aim to specify the unique features of this new format, but it remains an ever evolving and elusive problem. We aim to focus on a more narrowly defined but broadly applicable facet of the effort to track and monitor the spread of disinformation in online communities.

To understand in how technology can best be applied to this effort, it is necessary to first understand the specific challenges caused by social media as a medium of transmission. For one, the distribution of disinformation often isn't reliant on a centralized authority such as a respected news organization. It is much more common to use a combination of bots, trolls, and hacked accounts to initiate the spread and clever leveraging of the social media algorithm to ensure it reaches the appropriate real users that will give the message an organic boost in popularity. In addition, the actual content propagated is often

**Education and Research in the Information Society, October 13–14, 2022, Plovdiv, Bulgaria**
EMAIL: elena.tuparova@ontotext.com (E. Tuparova); andrey.tagarev@ontotext.com (A. Tagarev); nikola.tulechki@ontotext.com (N. Tulechki); svetla.boytcheva@ontotext.com (S. Boytcheva)
ORCID: 0000-0003-3222-4482 (E. Tuparova); 0000-0003-4262-7277 (A. Tagarev); 0000-0002-7318-1637 (N. Tulechki); 0000-0002-5542-9168 (S. Boytcheva)

very low effort and relies heavily on reuse of previous claims, ideas, images and videos- it is made different enough to not be a direct copy but enough of the message is reused that properly applied automated methods can be utilized to identify and track the reuse and evolution over multiple iterations and in different locations. These methods, however, need to be carefully selected and tailored to the task at hand.

In this paper, we will focus on the text aspect of reused messages which come with their own distinct set of challenges. Most obvious is that unlike image reuse or manipulation, rewriting text or only reusing parts of it requires minimal effort and is encountered quite often. This means that the algorithms used need to balance the identification of reused sentence fragments and terms with an awareness that those pieces of text need to be sufficiently unique and important to the topic of the text. In other words, a classical topic modeling task.

A high-quality and broadly applicable approach to topic modeling and clustering of disinformation content on a large scale is a crucial step in the analysis and understanding of its creation and spread. Once clusters of sufficiently good quality can be readily identified, it is possible to work on identifying clusters of compromised accounts, track large-scale disinformation contents that take part over months and even possibly identify the goals of the bad actors working on spreading this fake news.

## 2 Related Work

Boberg et al [1] present research for Facebook posts in German for 3-month period before the Corona crisis and different types of fake news and conspiracy theories. The authors apply various techniques, including some preprocessing, topic modelling based on Latent Dirichlet Allocation (LDA), co-occurrences study of top-ranked actors and insensitivity over time of the post. The study shows that it is difficult to distinguish disinformation related to COVID-19 from other social media posts, in contrast with alternative news media that show specific methods for spreading fabricated news. Chen et al [2] present an approach based on word embeddings over a large corpus of articles and Facebook posts collected by Media Cloud and CrowdTangle to study another socially important issue related to smoking and vaping. The authors show the limitations of BERT models for few-shots learning, as well as that the used data sources are not a representative for such data. Debnath et al [3] investigate another hot topic – electrical vehicles adoption in US and the reflections on this topic at social media. The authors also use a repository collected by CrowdTangle from Facebook posts. The methods for analysis include monitoring of the density and frequency of the related posts, topic modelling using LDA algorithm for different PESTLE (Political, Economic, Social, Technological, Legal and Environmental) categories. In this study LDA showed promising results helping the extract additional semantics for each of the PESTLE categories. Valensise et al [4] study similar research questions like our hypothesis and identify several correlations between vaccine attitudes and different infodemics indices and observe strong compatibility with a null model. Chatterje-Doody et al [5] presents research on social media trends of disinformation related to the other topic Russia's war on Ukraine. Broniatowski et al [6] presents a comparative study for disinformation in social media regarding various health related issues and argue that the disinformation about COVID-19 is significantly smaller than those for other health topics posts. Zhang et al [7] investigate in social media posts sentiment of users regarding the usefulness of COVID-19 vaccines.

## 3 Data

This paper focuses on studying the potentially disinformative content on social media, in particular Facebook. For that purpose, we have collected data about Facebook posts from 356 pages and 168 groups. These pages and groups were selected based on the fact that they have content in the Database of Known Fakes (DBKF), which "is a collection of debunking content from highly respected fact-checking organizations around the world extended with additional metadata related to said debunks" [8].

The data was collected via the CrowdTangle[1] platform, which is a public insights tool from Facebook and can be used to track public Facebook pages with more than 25K Page Likes or Followers, all public Facebook groups with 95k+ members, all US-based public groups with 2K+ members, and all verified profiles[2]. For each post CrowdTangle provides metadata such as page or group, time of creation of the page or group, time of creation of the post, number of reactions to the post, text of the post, image, link, video (if any), etc. For our study we have used the text of the post and the text from the corresponding link (if any). In compliance with CrowdTangle's policies we can make publicly available only the results from the study and not the raw data.

As the problem with disinformation and fake news has been becoming more serious in the past couple of years, the time period of the collected posts is between January 1st 2020 and June 6th 2022. For the time being the study focuses on English data, so the extracted posts were

preprocessed using SpaCy[3] in order to automatically detect their language. The text of a total of 1 184 944 posts from the pages and 3 641 224 from the groups was detected as English. It should be noted that it could be possible that part of the proven as disinformative content had already been taken down from Facebook prior to the moment of collecting the data. Nevertheless, the extracted English data is very large and for the purposes of the pilot study the time period has been narrowed to the time between September 1st, 2021 and March 31st 2022 and only the posts from the Facebook pages were considered, resulting in a total of 278 179 posts.

# 4 Methods

For the purpose of the study two topic modeling approaches were applied to the text of the extracted posts – Latent Dirichlet Allocation (LDA) and BERTopic. As topic modeling is generally an unsupervised machine learning technique, both methods are unsupervised, but handle the task in a different manner. Before we discuss the details of each method, we should clarify that in the area of text analysis a topic can be defined as a set of key words or phrases, which are related to each other and often occur together and by that define a given topic.

## 4.1 LDA

Latent Dirichlet Allocation (LDA) is a generative statistical classic model for topic modeling, discussed in [9]. It is based on the idea that each document can be represented as a distribution of different topics, while each topic can be represented as a distribution of key words. This means that each document has a set of topics it relates to with certain weight. We can say that the dominant topic of the document is the one with the largest corresponding weight. In order for LDA to model the topics for a given set of documents, it should be provided with the certain number of topics we expect it to model. For determining the optimal number of topics for any set of documents, some experiments should be conducted and evaluated based on the topic coherence score. Topic coherence can be defined as a degree of semantic similarity between frequent and relevant words and/or phrases for each topic. In other words, topic coherence is an indicator of how interpretable the output topics are to humans.

Several metrics for topic coherence exist – C_v, C_p, C_uci, C_umass, etc. For the purpose of the study the C_umass metric was selected for evaluation. It measures the co-occurrence of words in the topic and is recommended over the C_v metric[4].

The multithreaded Gensim[5] implementation of LDA was used in the study. Before applying the method, the documents were preprocessed – only nouns, adjectives, adverbs, and verbs were left, in order to pay attention to the more meaningful words. Also, phrases (chunks) were built, as this allows for words that often occur next to each other to be processed as a single entity, and lemmatization was done.

[1] https://www.crowdtangle.com/
[2] https://help.crowdtangle.com/en/articles/1140930-what-data-is-crowdtangle-tracking
[3] https://spacy.io/
[4] https://www.baeldung.com/cs/topic-modeling-coherence-score
[5] https://github.com/RaRe-Technologies/gensim

## 4.2 BERTopic

BERTopic is a relatively new model, discussed in [10], which uses transformer-based language models and operates in three distinct steps. First, embeddings are created for each document. Then the embedded documents are being semantically clustered, which includes two tasks –reduce the dimensions of the document embeddings, using UMAP [11], and then create the clusters using HDBSCAN. As a third and final step of BERTopic a tf-idf-based approach is applied at cluster level – c-tf-idf. This allows to find the most relevant words and/or phrases for each cluster of documents and in this way to determine the topic of the cluster. We should note that topic modeling using BERTopic results in each document having a single topic, in contrast to using LDA, where each document is a distribution of topics. Another difference is that BERTopic determines the number of topics itself instead of receiving it as a parameter. The official BERTopic python implementation was used [10]. Some preprocessing of the documents was made – n-grams where n=1 and n=2 are used in the analysis, all stopwords and words occurring in less than 10 documents were discarded.

# 5 Experimental Results

As two different approaches to topic modeling were applied, the results will be presented separately.

## 5.1 Results with LDA

As was mentioned in the previous section, LDA needs the number of topics to model specified in advance. In order to determine the optimal number of topics for the given set of documents some experiments were conducted. These experiments showed that for the current dataset, consisting of around 280 000 documents, 150 topics is optimal (for LDA), given the C_umass metric for topic coherence. The results from the experiments are summarized in Figure 1.

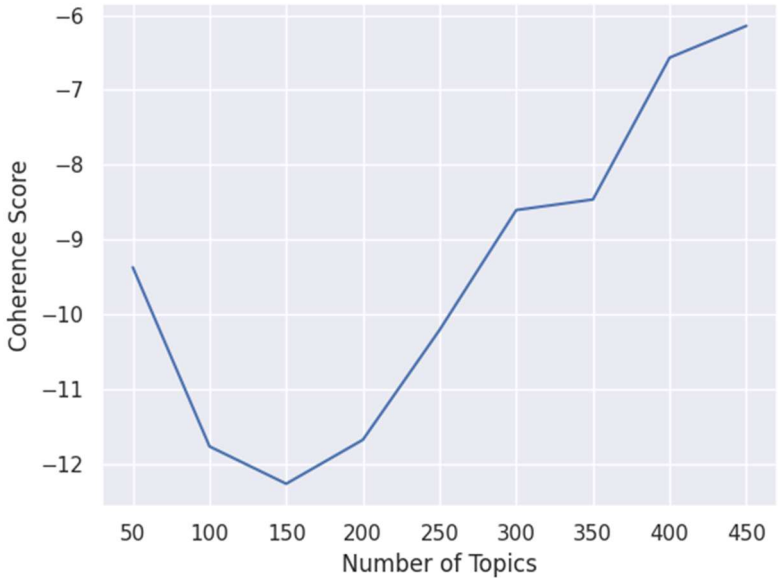


**Figure 1** A graph showing how the C_umass metrics changes with the number of topics

Some of the topics produced by LDA on the given dataset are presented in Figure 2. We can say that (although subjectively) the topics look generally good, but are not easily interpretable by humans.

As explained in section 4.1, with LDA each document can be represented as a distribution of different topics and each topic can be represented as a distribution of key words. **Figure 3** shows the distribution of key words for Topic 1 from Figure 2.

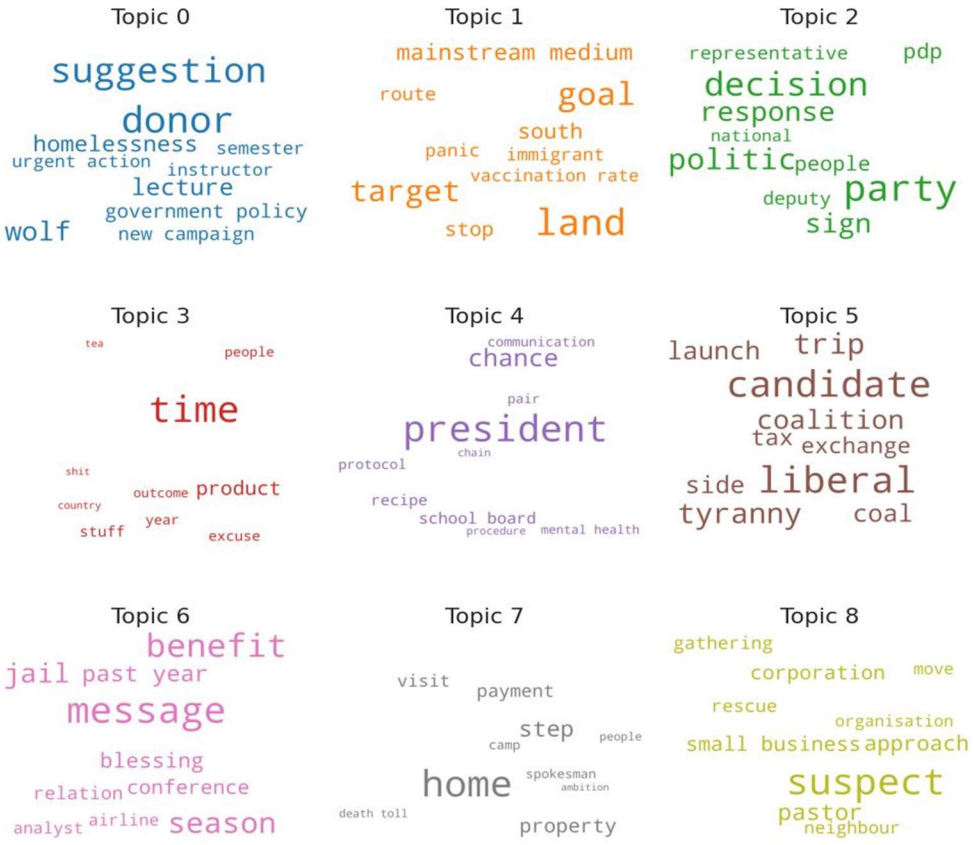


Figure 2 Sample topics produced by LDA

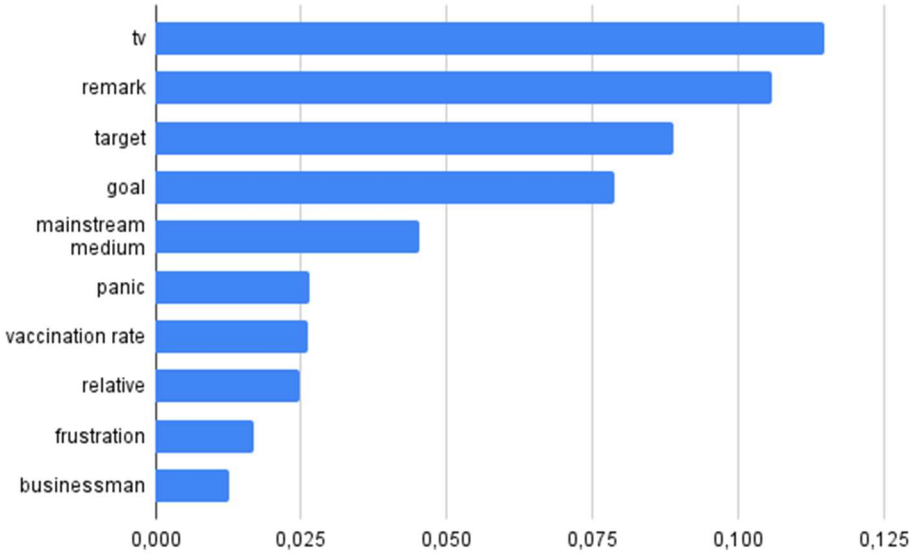


**Figure 3** Distribution of key words for LDA Topic 1

## 5.2 Results with BERTopic

For the given dataset BERTopic outputs 278 topics. The top 25 topics (with most relevant documents) are presented in Figure 4. The first column contains the number of the topic, the second – the number of documents in the corresponding cluster, and the third – the most relevant to the topic n-grams. This data is visualized in the chart on **Figure 5**.

| | | |
|---|---|---|
| -1 | 110886 | -1_covid_people_covid19_new |
| 0 | 23838 | 0_buhari_nigeria_state_president |
| 1 | 8554 | 1_khmer_khmer times_cambodia_times |
| 2 | 5369 | 2_cancer_body_foods_benefits |
| 3 | 4701 | 3_chelsea_barcelona_cup_liverpool |
| 4 | 4463 | 4_big tech_tech_facebook_censorship |
| 5 | 3301 | 5_reshare post_reshare_post_post lets |
| 6 | 2928 | 6_convoy_trudeau_canada_freedom convoy |
| 7 | 2883 | 7_ukraine_russian_russia_putin |
| 8 | 2832 | 8_nfl_nba_brandon_bowl |
| 9 | 2738 | 9_god_lord_shall_jesus |
| 10 | 2610 | 10_biden_joe_joe biden_bidens |
| 11 | 2569 | 11_bbnaija_fan journal_sports fan_fan |
| 12 | 2425 | 12_crypto_stocks_bitcoin_investors |
| 13 | 1926 | 13_true_lol_know_right |
| 14 | 1849 | 14_school_school board_parents_race theory |
| 15 | 1786 | 15_shot_police_suspect_shooting |
| 16 | 1531 | 16_follow occupy_occupy democrats_occupy_democrats |
| 17 | 1501 | 17_border_migrants_border patrol_patrol |
| 18 | 1500 | 18_afghanistan_taliban_afghan_withdrawal |
| 19 | 1359 | 19_mandate_vaccine mandate_vaccine_mandates |
| 20 | 1329 | 20_modi_india_singh_delhi |
| 21 | 1285 | 21_omicron_variant_omicron variant_cases |
| 22 | 1240 | 22_abortion_texas_law_abortions |
| 23 | 1183 | 23_rittenhouse_kyle_kyle rittenhouse_trial |

**Figure 4** Top 25 topics according to BERTopic

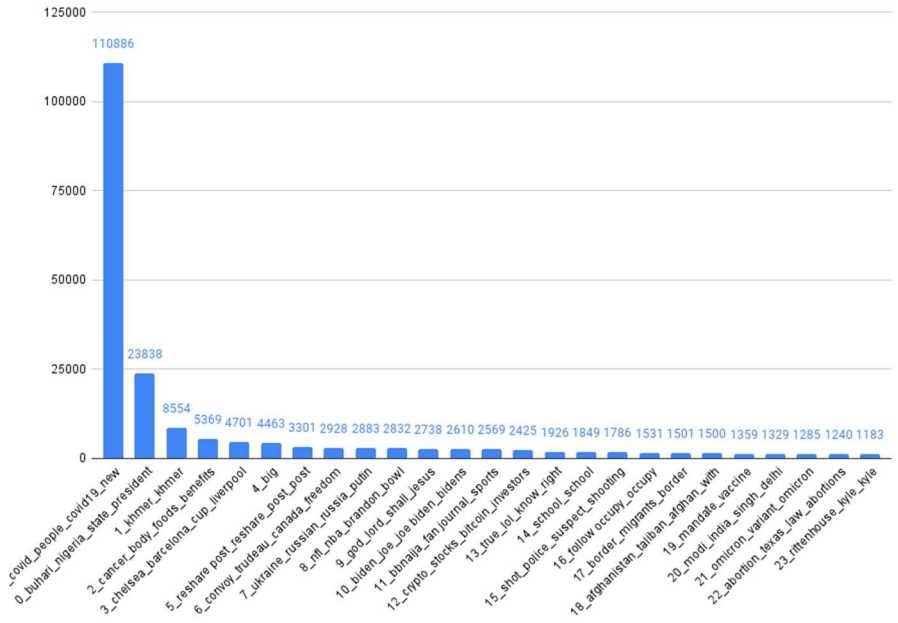


**Figure 5** A chart showing the top 25 topics from BERTopic and the number of their corresponding documents

Clearly the most documents (ca. 40%) fall into the cluster for topic -1. This cluster contains the outliers, which could not be clustered into any other topic. Regardless, this topic could easily Figure 2: Sample topics produced by LDA be interpreted by humans, as could all the others. In general, the presented topics correspond to easily identifiable events or popular news discussions that occurred in the time frame over which the data was collected. This includes events such as the Chelsea vs Liverpool cup final, the Canadian truck convoy, the discussion on teaching critical race theory in American schools, the Khmer times plagiarism scandal and so on.

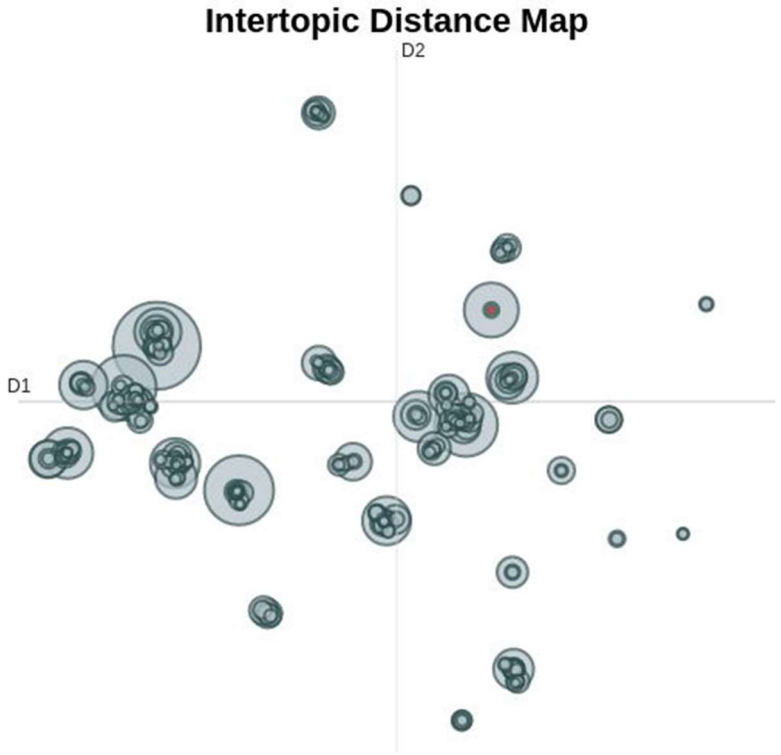


**Figure 6** Distance map of the topic clusters produced by BERTopic

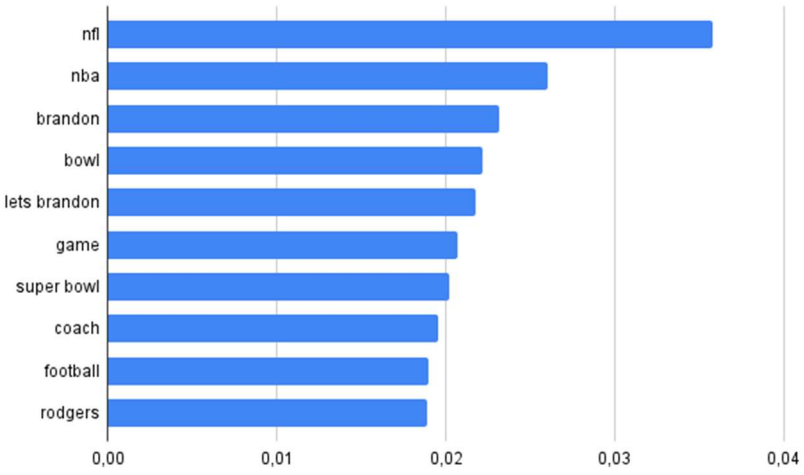


**Figure 7** Distribution of key words for a BERTopic topic (topic 8)

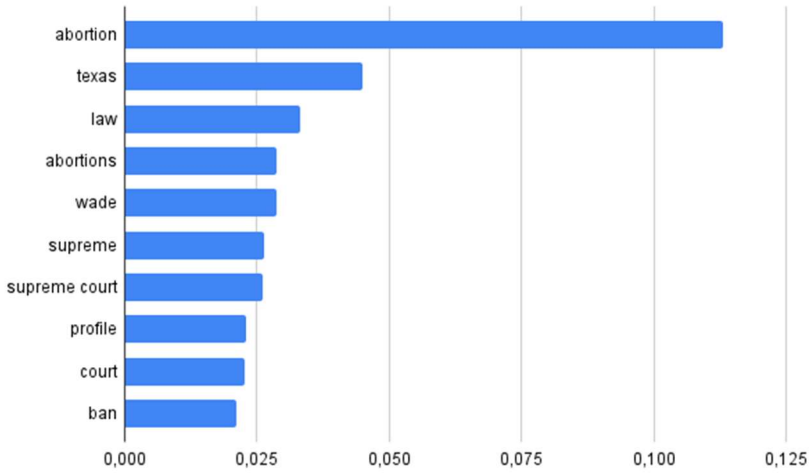


**Figure 8** Distribution of key words for a BERTopic topic (topic 22)

It is also easy to note that some of the topics, such as the connections between foods and cancer or Joe Biden, are quite generic. Among them, however, we can still discover discussions about omicron or border and migrants which are quite fertile and persistent ground for disinformation so still valuable to track in our context.

**Figure 6** is showing the distance map of the produced clusters projected into a two-dimensional space.

**Figure 7** and **Figure 8** show the distribution of keywords for two of the top 25 topics – topic 8 and topic 22. In contrast to the topics modelled by LDA, here we see a very clear and coherent keyword distribution, which can easily be interpreted.

## 6 Discussion

The obtained results could be discussed from several points of view. Firstly, the task at hand is an unsupervised machine learning task – the documents we are handling have not been previously annotated, so we cannot evaluate precisely how accurately each of the applied methods predicts the topics. Moreover, the dataset is too large for a manual validation of the whole of it to be feasible. Nevertheless, by looking at some of the topics, produced by both methods, we can conclude that the topics that come as output from BERTopic seem more coherent and easier to interpret by humans (although somehow subjectively).

We could say that this was expected, as BERTopic is a model which uses more elaborate methods for word and document representation, which take the context into consideration as well, namely transformers. On the other hand, LDA is considering each document as a Bag of Words, meaning the order in which the words occur has no significance.

Another advantage of BERTopic is the possibility to adapt the model for the multilanguage dimension of the task by simply changing the transformer, which is used for the document embeddings.

## 7 Conclusion and Further Work

The results of our experiments have shown that while a strict evaluation is difficult, topic clustering methods are capable of identifying and grouping together discussions on disinformation spaces into clusters that are useful for analysis. The output of the algorithms is of sufficient quality to be presented to non-technical partners for evaluation and utilization in their efforts.

Of the two algorithms we experimented with, BERTopic showed more promise. This is due both to its automatic ability to select a number of topics appropriate to the available data and the more sharply defined and identifiable nature of the topics produced. It has other benefits as well, such as the ability

to utilize a language model in the first step for truly multilingual clustering. Since the dataset we are working with is inherently multilingual, the next logical step in our clustering experiments is, in fact, to experiment with BERTopic's ability to produce truly multilingual clusters.

Another direction of further research is exploring the opportunity for multi-modal clustering. This is a much less explored field, but disinformation is a field that uses text, images, videos and even audio files, not to mention more nebulous formats such as memes which blur the distinction between image and text.

Ultimately, the goal of these experiments is to produce a system that can sift through the massive amounts of data being generated and allow journalists and debunkers to focus their attention on the most important, interesting, and faster growing topics. Creating these clusters is the first crucial step towards that goal and enables us to continue with the next steps of analysis such as examining the evolving activity of users sharing disinformation, identify coherent groups of these users that act in sync with each other and possibly even identify larger scale disinformation campaigns aimed at affecting the general perception and opinion in society on specific topics.

# 8 References


[1] S. Boberg, T. Quandt, T. Schatto-Eckrodt, L. Frischlich, Pandemic populism: Facebook pages of alternative news media and the corona crisis–a computational content analysis, arXiv preprint arXiv:2004.02566 (2020).

[2] K. Chen, M. Babaeianjelodar, Y. Shi, R. Aanegola, L. Y. Cheung, P. I. Nakov, S. Yadav, A. Bancroft, A. Khudabukhsh, M. De Choudhury, et al., Us news and social media framing around vaping, arXiv preprint arXiv:2206.07765 (2022).

[3] R. Debnath, R. Bardhan, D. M. Reiner, J. Miller, Political, economic, social, technological, legal and environmental dimensions of electric vehicle adoption in the united states: A social-media interaction analysis, Renewable and Sustainable Energy Reviews 152 (2021) 111707.

[4] C. M. Valensise, M. Cinelli, M. Nadini, A. Galeazzi, A. Peruzzi, G. Etta, F. Zollo, A. Baronchelli, W. Quattrociocchi, Lack of evidence for correlation between covid-19 infodemic and vaccine acceptance, arXiv preprint arXiv:2107.07946 (2021).

[5] P. Chatterje-Doody, R. Crilley, Three lessons for the future of public service broadcasting: Information, confrontation and russia's war on ukraine, IPPR Progressive Review (2022) Early–Access.

[6] D. A. Broniatowski, D. Kerchner, F. Farooq, X. Huang, A. M. Jamison, M. Dredze, S. C. Quinn, J. W. Ayers, Twitter and facebook posts about covid-19 are less likely to spread misinformation compared to other health topics, PloS one 17 (2022) e0261768.

[7] W. Zhang, S. Mukerjee, H. Qin, Topics and sentiments influence likes: A study of facebook public pages' posts about covid-19 vaccination, Cyberpsychology, Behavior, and Social Networking (2022).

[8] A. Tagarev, K. Bozhanova, I. Nikolova-Koleva, I. Ivanov, Tackling multilinguality and internationality in fake news, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), INCOMA Ltd., Held Online, 2021, pp. 1380–1386. URL: https://aclanthology.org/2021.ranlp-1.154.

[9] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 (2003) 993–1022.

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

[11] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426 (2018)