The Impact of Stopwords Removal on Disinformation Detection in Ukrainian language during Russian-Ukrainian war

Halyna Padalko[1,2,3], Vasyl Chomko[2] and Dmytro Chumachenko[1,2,4]

[1] National Aerospace University "Kharkiv Aviation Institute", Vadym Manko str., 17, Kharkiv, 61070, Ukraine
[2] University of Waterloo, 200 University Ave W, Waterloo, N2L 3G5, Canada
[3] Centre for International Governance Innovation, 67Erb Str W, Waterloo, N2L 6C2, Canada
[4] Balsillie School of International Affairs, 67Erb Str W, Waterloo, N2L 6C2, Canada

Abstract

Political disinformation is a growing threat to democracy, particularly in the context of warfare like the full-scale Russian invasion of Ukraine. Analyzing Ukrainian-language disinformation, especially on platforms like Telegram, is essential for understanding the narratives used by hostile actors. This study addresses the gap in Ukrainian-language research by applying advanced topic modelling techniques to improve disinformation analysis. Using a dataset of Ukrainian news articles and titles, we employed the BERTopic model, leveraging BERT-based embeddings and hierarchical clustering. The results showed that topic modelling performs better on full news bodies than titles, and removing stopwords significantly enhances topic clarity. Hierarchical clustering and topic modelling revealed consistent patterns, highlighting the importance of using both methods for comprehensive analysis. This study offers valuable insights into Ukrainian disinformation tactics and methodological improvements for more accurate topic modelling, aiding efforts to counter disinformation in politically sensitive contexts.

Keywords

disinformation, Telegram, topic modelling, BERT, stopwords

1. Introduction

The problem of political disinformation has become a critical issue in preserving democracy, especially in the digital era, where information can spread rapidly across multiple platforms [1].

Disinformation undermines democratic processes by influencing public opinion, spreading confusion, and eroding trust in institutions [2].

With the development of strategic disruptive technologies, including artificial intelligence (AI), the potential for manipulating information has expanded significantly, making it harder to distinguish between truth and falsehood [3].

Russia's war against Ukraine serves as a prime example of how hybrid warfare, which combines military action with psychological operations or cognitive warfare, is used to destabilize the situation [4]. The conflict highlights the fusion of conventional warfare with disinformation campaigns, creating a complex battlefield of information. The experience gained from analyzing Russia's informational operations against Ukraine presents a valuable case study for researchers aiming to analyze new disinformation techniques and narratives in a rapidly developing AI environment.

One of the most significant characteristics of Russian disinformation is its reliance on emotional appeals and narrative development. By tapping into the public's fears, frustrations, and personal experiences, Russian disinformation seeks to manipulate emotions [5]. In contrast, the Western approach to information dissemination focuses more on presenting facts and reasoned arguments, which may be less effective in emotionally charged environments.

ProfIT AI 2024: 4th International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2024), September 25–27, 2024, Cambridge, MA, USA

galinapadalko95@gmail.com (H. Padalko); vchomko@uwaterloo.ca (V. Chomko); dichumachenko@gmail.com (D. Chumachenko)

0000-0001-6014-1065 (H. Padalko); 0009-0009-4419-6651 (V. Chomko); 0000-0003-2623-3294 (D. Chumachenko)

Understanding the role of narratives is crucial in countering disinformation. Narratives shape how information is received and processed, and studying these narratives can reveal underlying tactics and strategies in disinformation campaigns. Analyzing these narratives helps policymakers develop effective responses, including debunking and prebunking false information and reinforcing strategic communication efforts.

Topic modelling effectively analyses disinformation by identifying patterns, themes, and narratives within large text datasets [6].

It helps researchers detect and categorize the underlying topics disseminated by disinformation actors. Latent Dirichlet Allocation (LDA) and other advanced algorithms are commonly used to parse disinformation content from social media, news, and other text sources [7].

These models uncover the main disinformation themes, providing insights into how false narratives evolve and spread [8]. For instance, topic modelling has been used to analyze disinformation campaigns related to elections and public health, revealing coordinated efforts to manipulate public opinion [9].

While Russian disinformation targets Western societies, Ukraine has been the primary focus of its campaigns. Using various channels and techniques to spread disinformation against Ukraine offers a unique opportunity to study the methods used by authoritarian regimes against democratic adversaries. The lessons learned from Ukraine's experience with Russian disinformation are invaluable to Western countries, showcasing Russia's unconventional warfare tactics. However, there is a notable gap in research focused on analyzing Ukrainian datasets, limiting the full understanding of Russia's approach.

This research is particularly urgent because it fills a gap in analyzing Russian disinformation campaigns in the Ukrainian context. We have used datasets from Telegram, Ukraine's most popular news platform, to carry out this analysis. Telegram has become a crucial tool for Ukrainians, not only for tracking news but also for monitoring air raids and receiving updates on the war. Telegram's user base grew from 20% in 2021 to 72% in 2024, driven by the need for real-time updates and the ability to share information quickly during the full-scale war [10].

Most existing research on disinformation related to the Russia-Ukraine war focuses on English-language datasets, which limits the accuracy of topic modelling for non-English data. The lack of models trained on Ukrainian language corps of data further complicates analysis. This study addressed these limitations by running machine learning models on the Ukrainian dataset. We try to enhance analysis by removing commonly used stopwords from the initial dataset, aiming to improve the interpretability of topic modelling and clarify the narratives being propagated. This methodological improvement aims to clarify the disinformation tactics used against Ukraine and how they can inform more effective responses in democratic nations.

2. Current Research Analysis

The study of misinformation and disinformation in the Ukrainian language, particularly in the context of the Russian-Ukrainian war, has gained increasing attention in recent years. Several researchers have explored how false or misleading information is disseminated and consumed, often focusing on the linguistic and thematic elements unique to the Ukrainian language. Notable works in this field investigate the role of social media platforms, the spread of propaganda, and the strategies hostile actors use to influence public opinion. This section reviews key contributions to the field, focusing on the methodologies used to analyze misinformation and how these approaches can be improved.

The paper by Maathuis and Kerkhof [11] analyses Ukrainian-language discourse on Telegram during the first six months of the Russian-Ukrainian war, applying machine learning techniques to capture the main topics discussed and their sentiments. The study leverages a dataset of nearly 46,000 messages, applying topic modelling and sentiment analysis to understand how Ukrainian users communicated in this period. The research contributes significantly to understanding the public's responses to the war, particularly through social media, which has played a crucial role in

disseminating information during the conflict. However, a notable limitation of the study is its reliance on machine translation via Google Translate, which may introduce sentiment and topic identification inaccuracies due to language nuances that automated translation systems may not fully capture.

The paper [12] explores the role of automated agents (bots) in shaping public discourse during the Russian-Ukrainian war. Utilizing a dataset of over 1.6 million bot-driven tweets, the study employs machine learning techniques like TweetBERT and frameworks like the BEND framework and Moral Foundation Theory to analyze Russian disinformation campaigns and Ukrainian counter-narratives. The research reveals how bots were used to amplify political narratives, creating echo chambers that manipulated public perception. It highlights the tactical deployment of bots during key conflict events, showing distinct patterns of narrative control by pro-Russian and pro-Ukraine forces. One notable limitation of the study is its exclusive focus on bot-generated content, which excludes human-driven interactions and thus might overlook the full complexity of digital propaganda dynamics.

A study [13] analyzes the performance of various machine learning models for identifying disinformation in Ukrainian news headlines. Using a dataset collected during the Russian-Ukrainian war, the authors assess several classifiers, including logistic regression, support vector machines (SVM), random forest, gradient boosting, KNN, decision trees, XGBoost, and AdaBoost. Their evaluation focuses on key metrics like precision, recall, F1-score, and accuracy, with the random forest model achieving the highest accuracy (95.3%), proving to be the most effective at distinguishing true from false news items. This study underscores the critical role machine learning plays in automating the detection of disinformation, particularly in the context of information warfare. However, while high-performing, the models may struggle with more nuanced or contextually complex forms of misinformation, requiring further refinement for broader applications.

The paper [14] presents a hybrid approach for detecting hidden propaganda and sentiment analysis of media in Ukrainian and Russian. Using rule-based methods, dictionary approaches, and machine learning models, the authors developed a system capable of processing large volumes of media content, focusing on identifying manipulative language and emotional tones. The methodology includes named-entity recognition (NER) and morphological tagging to analyze over 630,000 articles from media sources, demonstrating high accuracy in detecting sentiment and propaganda, particularly emphasizing the context of Ukraine. However, a limitation of the study is its reliance on handcrafted dictionaries, which may limit the model's ability to generalize across different domains or evolving language patterns.

Study [15] explores the use of machine learning operations (MLOps) for analyzing online discussions related to the Russian-Ukrainian war. The authors used social media data from VKontakte, covering the period from January 2022 to May 2023, to model topics and identify discussion trends. The LDA algorithm was applied to classify text data into topics, and a set of dashboards was developed using Splunk Enterprise for real-time monitoring and analysis of model performance. The study highlights how social media discussions evolved, especially focusing on the emotional tone, keywords, and misinformation trends, revealing how antiwar hashtags were used to promote pro-war content. One key limitation of the study is its reliance on LDA, which does not account for sentiment or topic evolution over time, potentially overlooking nuances in the sentiment dynamics of the discussions.

The reviewed studies provide valuable insights into analyzing misinformation and social media discussions in the context of the Russian-Ukrainian war, employing various methodologies such as machine learning, sentiment analysis, and topic modelling. These papers demonstrate the significance of language-specific and platform-specific factors in shaping online narratives. Yet, they also reveal limitations, including the need for more advanced tools to handle evolving topics, capture sentiment, and generalize across different datasets. Our study builds on this body of work by focusing specifically on Ukrainian-language disinformation and improving the accuracy of topic modelling by eliminating stopwords. By addressing the linguistic nuances unique to Ukrainian, our research

fills an important gap, enhancing the precision of topic analysis in the ongoing information warfare context.

3. Materials and Methods

For our analysis, we used two Ukrainian language datasets from Telegram: one containing titles only and another with the full bodies of Telegram channel messages [16]. This substantial dataset comprises approximately 50,000 news articles and 11000 news titles, meticulously labelled as “fake” or “true,” thereby providing a balanced foundation for qualitative and quantitative analyses. The dataset spans a significant period from the 24th of February, 2022, to the 11th of December, 2022, a time frame marked by notable political and social events in Ukraine. This temporal coverage ensures that the dataset encapsulates various topics and narratives, reflecting news content’s diverse nature and misinformation strategies’ evolution.

Each entry in the dataset contains four key attributes: a unique identifier (id), the headline of the article (title), the full textual content (text), and a classification label (label) indicating the authenticity of the news piece. The articles labelled as “true” were sourced from reputable Ukrainian news outlets known for their journalistic standards and adherence to factual reporting. Conversely, the “fake” articles were collected from sources identified as propagators of misinformation.

To analyze the existing narratives within these datasets, we employed topic modelling as a key approach for data clustering. This method allowed us to group messages based on common themes and narratives, providing insight into the structure and content of disinformation spread via Telegram.

To analyze the dataset of fake news articles, we began by filtering the data to include only fake news entries. Next, we cleaned the text data by removing punctuation, numbers, and stopwords. We used large set of 1,983 Ukrainian stopwords including numbers from the dataset [17]. We have also removed specific phrases related to Ukrainian Telegram channels and information sharing such as "українські телеграмканали," "телеграмканали повідомляють," "повідомляють українські телеграмканали," and "telegram", “українських телеграмканалах”, “часто згадується”, “інформація розповсюджувалась”, “повідомлення поширювались”, “новинних телеграмканалів”, “така інформація”, “про це пишуть”, “українські джерела”, “посиланням на твітер”, “повідомлення поширюються”, “це пишуть місцеві”. They do not have any particular sence for topic extraction and could rather confuse the model than produce clear outcome of topic modelling.

Two versions of the processed text were prepared: one with stopwords removed and another with stopwords retained, allowing for an assessment of the impact of stopwords on topic modelling outcomes.

We employed the “sentence-transformers/facebook-dpr-ctx_encoder-single-nq-base” model from the SentenceTransformers library to generate dense vector embeddings for our text data. This model, a Context Encoder designed for Dense Passage Retrieval (DPR), encodes sentences and paragraphs into a 768-dimensional vector space. Pre-trained on the Natural Questions dataset, it captures rich contextual semantics from the passages it processes. Utilizing a transformer-based architecture, the “dpr-ctx_encoder-single-nq-base” effectively grasps the relationships between words and their broader textual context, making it ideal for tasks requiring deep semantic retrieval.

To capture the contextual meaning of words in the dataset, we used transformer-based embeddings as input for the BERTopic model. Specifically, we employed a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model to generate embeddings from Ukrainian text data [18]. The embeddings are obtained by passing the input text through the BERT architecture, generating a dense vector for each word, where the vector dimension represents the word’s semantic context. The BERT model utilizes self-attention mechanisms, described by the following equation:

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (1)$$

where Q represents the query matrix, K is the key matrix, V is the value matrix, and d_k is the dimensionality of the key vectors.

The BERT embeddings are produced using this mechanism, where each word's representation is adjusted based on its surrounding words, thus preserving semantic relationships. These dense embeddings provide the basis for the subsequent topic modelling process.

The core topic modelling algorithm we used is BERTopic, which builds on transformer-based embeddings to cluster similar text fragments into topics. BERTopic incorporates two primary steps: dimensionality reduction of the embeddings and topic formation through clustering.

First, we reduced the dimensionality of the high-dimensional BERT embeddings using the Uniform Manifold Approximation and Projection (UMAP) technique. UMAP operates by projecting the high-dimensional embeddings into a lower-dimensional space while preserving the topological structure of the data. The UMAP algorithm minimizes the following loss function:

$$L_{UMAP} = \sum_{(i,j)\in S} \log\left(\frac{p_{ij}}{q_{ij}}\right), \quad (2)$$

where p_{ij} represents the high-dimensional similarity between points i and j, q_{ij} represents the low-dimensional similarity, and S is the set of neighboring points in the high-dimensional space.

The output is a low-dimensional embedding space where similar text fragments are positioned closely together, facilitating the clustering process.

Once the dimensionality of the embeddings was reduced, the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithm was applied to group similar text fragments. HDBSCAN, a density-based clustering algorithm, identifies clusters of various densities by analyzing the density of points in the reduced embedding space. It works by finding clusters where points are densely packed together while treating sparse regions as noise. The key advantage of HDBSCAN is its ability to identify variable-density clusters without requiring a pre-specified number of clusters, making it well-suited for discovering dynamic and complex topics in disinformation campaigns.

HDBSCAN's clustering algorithm is based on calculating the mutual reachability distance, defined as:

$$d_{mreach}(\text{x}, \text{y}) = \max\left(core\_distance(x), core\_distance(y), d(x, y)\right), \quad (3)$$

where d(x,y) is the Euclidean distance between points x and y, core_distance(x) is the minimum distance required to form a cluster around x, d_{mreach} is the modified distance used to determine whether points belong to the same cluster.

HDBSCAN produces clusters corresponding to distinct topics based on these mutual reachability distances.

To interpret and represent the topics generated by the clustering process, we utilized a KeyBERT-inspired representation model. KeyBERT extracts each topic's most representative words or phrases by leveraging the cosine similarity between the topic embeddings and individual word embeddings. Cosine similarity is calculated as:

$$\text{cosine\_similarity}(\text{A}, \text{B}) = \frac{A \cdot B}{\|A\|\|B\|}, \quad (4)$$

where A and B are the embeddings of two words or phrases.

KeyBERT ranks the terms based on their similarity to the topic embeddings, ensuring that the most relevant terms or phrases are selected to represent the topics.

To further enhance the topic representation, we incorporated a CountVectorizer to detect n-grams ranging from unigrams (single words) to trigrams (three-word phrases). The CountVectorizer creates a frequency-based matrix of terms within the text, capturing how often n-grams appear across the dataset. For each text t_i, the CountVectorizer calculates the frequency of each n-gram n_j as:

$$\text{freq}(n_j) = \frac{\sum_{t_i} 1(n_j \in t_i)}{N}, \quad (5)$$

where $1(n_j \in t_i)$ is an indicator function that returns 1 if n_j appears in t_i, N is the total number of documents in the dataset.

The inclusion of n-gram analysis allowed us to capture isolated keywords and multi-word phrases that contributed to the disinformation narratives, further improving the model's ability to detect complex patterns in the data. Both flat and hierarchical clustering methods were used to uncover the main topics and their potential subtopics.

After running the BERTopic model and identifying clusters, we manually reviewed each topic's top keywords and phrases. This qualitative analysis helped refine the understanding of the identified topics, ensuring they were both relevant and interpretable within the context of Ukrainian-language disinformation. The identified topics were categorized based on their dominant themes, such as emotional manipulation, propaganda techniques, or political narratives. By interpreting these topics, we aimed to provide actionable insights into the disinformation tactics used by Russian actors during the full-scale invasion.

Visualizations, such as the intertopic distance map and hierarchical clustering diagrams, were generated to illustrate the relationships between identified topics. This approach allowed for analyzing fake news topics, highlighting key themes and their interconnections.

Intertopic distance map visualization helps users interpret topic modelling results to represent the relationships between topics in a low-dimensional space that shows the proximity of topics, topic size and topic overlap. The proximity of the topics is the distance between the topics on the map, which indicates their similarity. Topics closer together share more words and are thematically related, while those farther apart have fewer words in common and represent more distinct themes. Topic size is the size of each circle on the map is proportional to the frequency of the topic in the dataset, representing how much of the overall text corpus each topic covers. Topic Overlap is the degree of overlap between the circles, showing whether topics share significant words, indicating thematic overlap. A clear separation of circles indicates distinct and well-separated topics.

Hierarchical clustering visualization is a way to display the relationships between different data points in a hierarchical structure. The dendrogram is the most common visualization of hierarchical clustering, which visually represents how clusters are merged or split during the clustering process. The diversity of colours indicates that the data points are distributed across multiple groups, forming smaller and more distinct clusters. The branches merge progressively, meaning the clusters are gradually combined.

4. Results

Analyzing 2 datasets of news titles and the body of the news and dividing them into 4 datasets, including and excluding stopwords, showed us the following results. We compare them within pairs of the original 2 datasets. Topic distribution is presented using 2 visualization approaches: intertopic distance map and hierarchical clustering.

Dataset "News bodies" with excluding stopwords, initially have fewer topics (34) than news bodies including stopwords (43), which shows better topic distribution.

Figure 1 presents news bodies excluding stopwords clustering. The topics seem more widely distributed across the map, with several clusters formed, and topics are more spread out. While some topics appear closer together, many circles are positioned far apart, indicating more distinct topic

separations with less overlap in word distributions. The size of the clusters varies, which may indicate a mix of large and small topics.

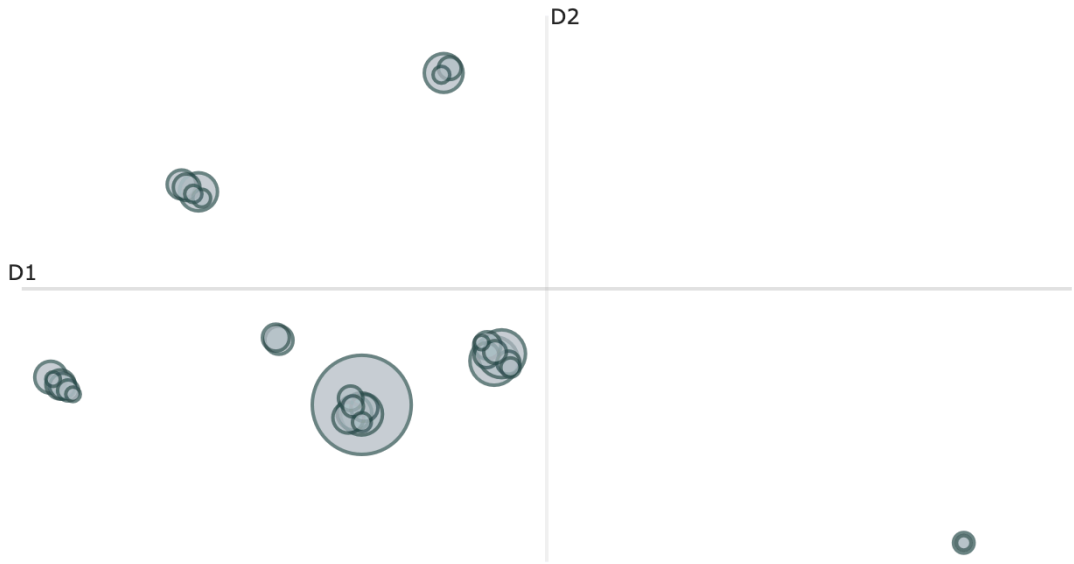

Figure 1: Intertopic Distance Map for "News bodies" dataset excluding stopwords

Figure 2 presents news bodies including stopwords clustering. The graph shows that several topics are grouped closely together. The density of topics in the top-right quadrant suggests that the topic model identifies more related themes that share common words. The circles are more closely packed and frequently overlap, indicating that many topics are more similar or share a greater number of words. However, there is no clear distribution of the topics in this cluster, suggesting that topics are divided and grouped chaotically, without clear distribution which makes interpretation of the graph harder.

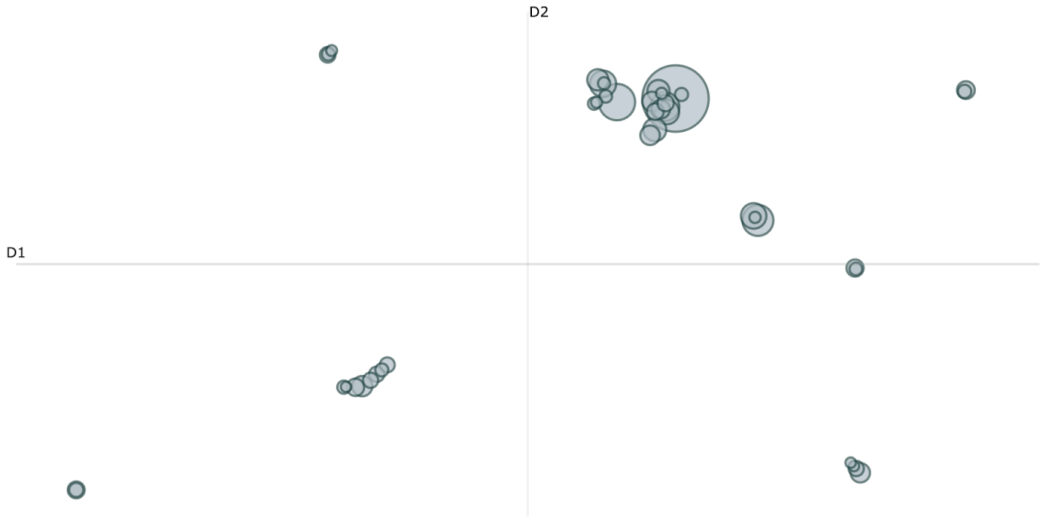

Figure 2: Intertopic Distance Map for "News bodies" dataset including stopwords

Figure 3 presents news bodies excluding stopwords. It shows that the clusters (branches) seem more evenly distributed across different groups, with a clear separation between different sets of branches. There is a wider range of distinct clusters (shown by a variety of colors), indicating the presence of multiple separate groups. The clusters in this graph seem to be smaller and more fine-grained, meaning the data is more evenly divided into different segments. The clusters seem to merge at a more gradual pace, which means each group remains distinct for longer as the hierarchy is built.

Figure 3: Hierarchical clustering for news bodies excluding stopwords

Figure 4 presents news bodies including stopwords clustering. The graph has a more pronounced structure of clusters within the lower part (indicated by the larger purple section). The top portions (green and red) still maintain a relatively even cluster distribution, similar to the first graph, but there is more focus on the larger segments within the bottom half (purple region). We observe that the merging occurs earlier in certain sections (especially in the purple region), indicating that these data points are more similar, leading to larger early mergers.

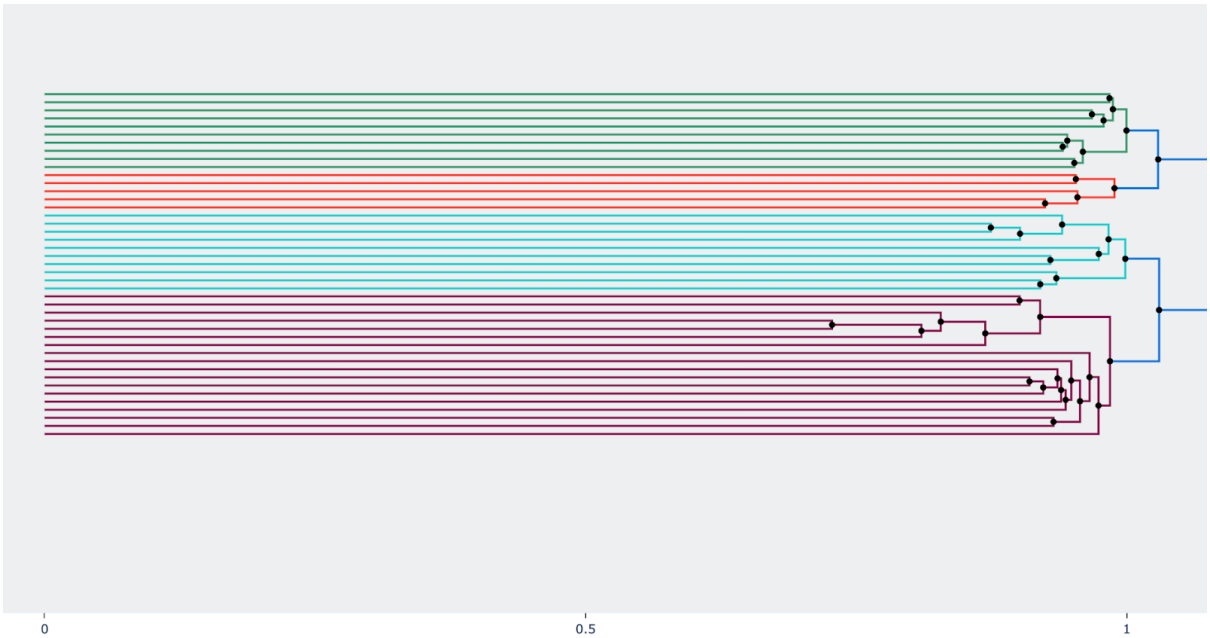

Figure 4: Hierarhical clustering for news bodies including stopwords

Dataset for “News titles” with excluding stopwords, initially shows 49 topics (49) and news names including stopwords 54 topics.

Figure 5 presents news titles excluding stopwords clustering.

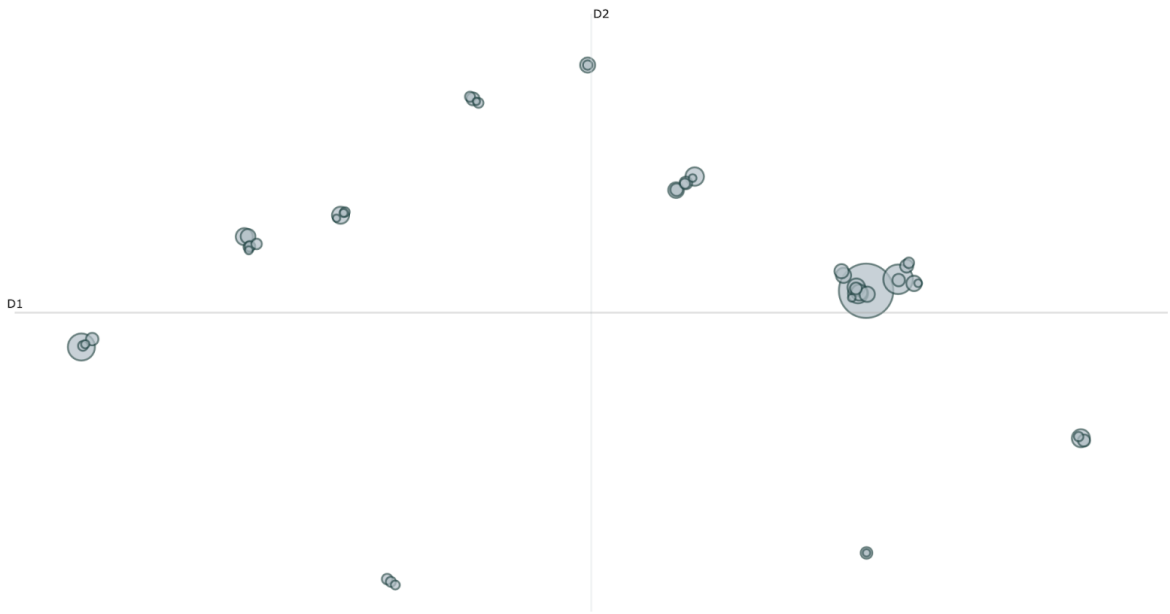

Figure 5: Intertopic Distance Map for “News titles” dataset excluding stopwords

The topics in this map are more spread out across the graph, with several distinct clusters of circles. The distances between the clusters are relatively large, which indicates that the topics are well-separated from each other. Due to the larger separation between topics, this map is likely easier to interpret. The topics appear more distinct, which means the themes they represent may be quite different from one another. This clear separation suggests that the model has successfully differentiated between unique themes in the dataset. There is minimal overlap between the circles, which indicates that the topics are more independent of each other. This structure makes it easier to identify distinct themes without confusion from overlapping or closely related topics. The overall distribution suggests balanced clusters with a range of topic sizes, but none of them dominate excessively. The first intertopic distance map offers better separation and distinct topic identification, making it easier to interpret the dataset’s thematic structure. Topics are well-separated, and the map suggests a clear distinction between different themes.

Figure 6 presents news titles including stopwords clustering.

In this map, the topics are more clustered together compared to the first map. Several topics appear closer to each other, forming denser groups. While there are still some distinct clusters, the overall distribution suggests more overlap and proximity between topics. Due to the closer grouping of topics, this map may be more difficult to interpret compared to the first. The proximity between topics suggests that some themes might be related or share similar terms, leading to less distinct clusters. The clusters are harder to differentiate, which could make it challenging to pinpoint specific themes. More overlap is visible in this map, with several circles touching or being very close to each other, indicating that the topics might share common terms or themes. This overlap can make it difficult to distinguish between the topics, as they may not be as clearly separated as in the first map. There is a more concentrated cluster in one section of the map, where a larger topic dominates with several smaller circles clustered around it.

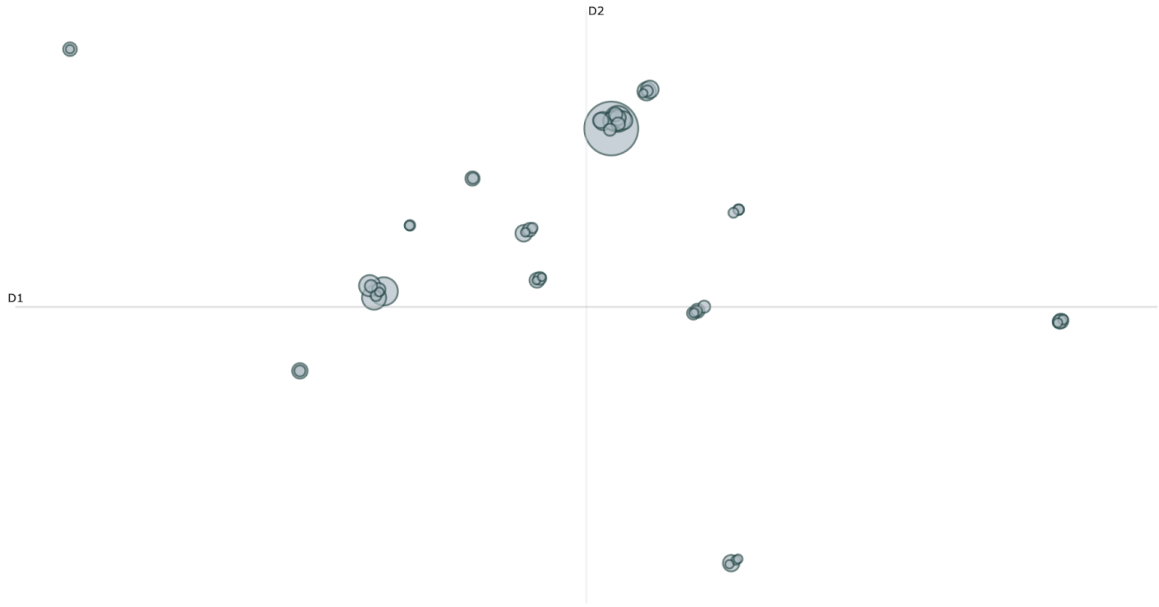

Figure 6: Intertopic Distance Map for "News titles" dataset including stopwords

Figure 7 presents hierarchical clustering for news titles excluding stopwords.

The clusters in this graph are well-distributed with noticeable differences between several distinct clusters. Each color represents a separate group of data points, and the branching suggests a gradual merging process. The smaller branches indicate that the clusters remain distinct for longer periods before merging, suggesting more precise grouping.

This graph seems to maintain clear boundaries between groups (e.g., green, red, yellow, black). The branching structure reflects a more detailed division between clusters, with specific groups (e.g., the red and yellow groups) having distinct sub-groups before they eventually merge into larger clusters. This graph allows for better interpretability, as the clear hierarchical structure offers a more transparent view of how the individual clusters are related and how distinct sub-groups remain for longer before merging.

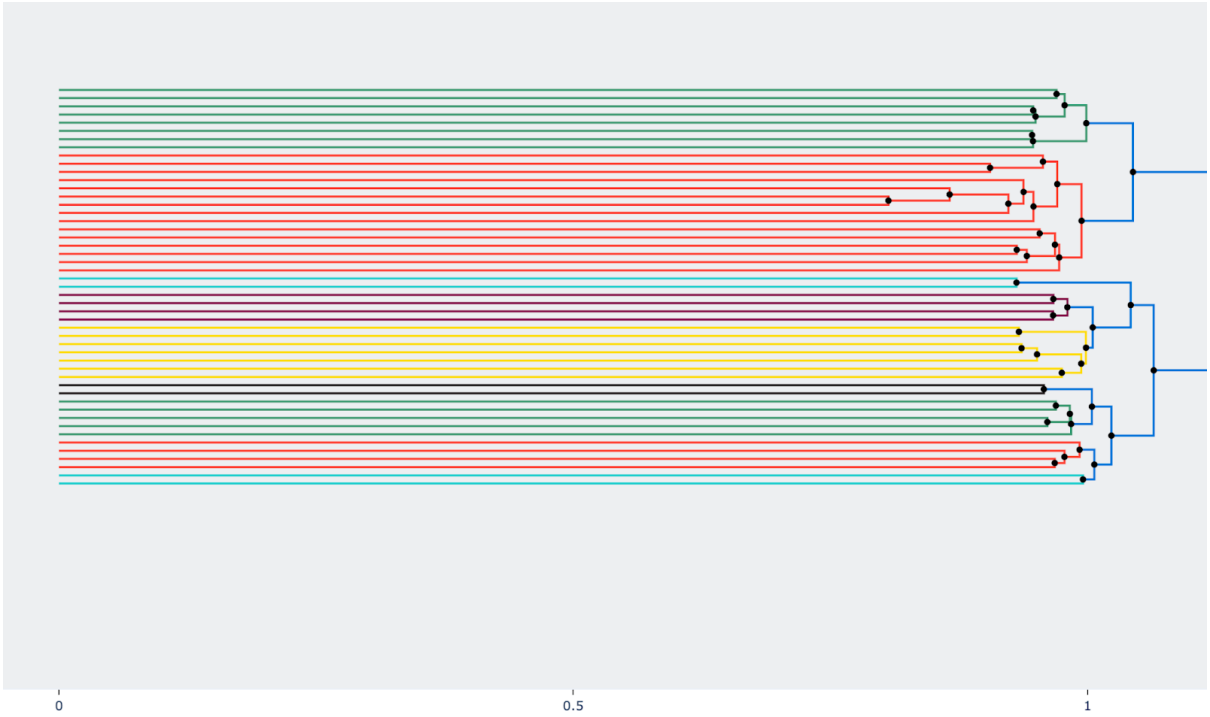

Figure 7: Hierarhical clustering for news titles excluding stopwords

Figure 8 presents hierarchical clustering for news titles excluding stopwords.

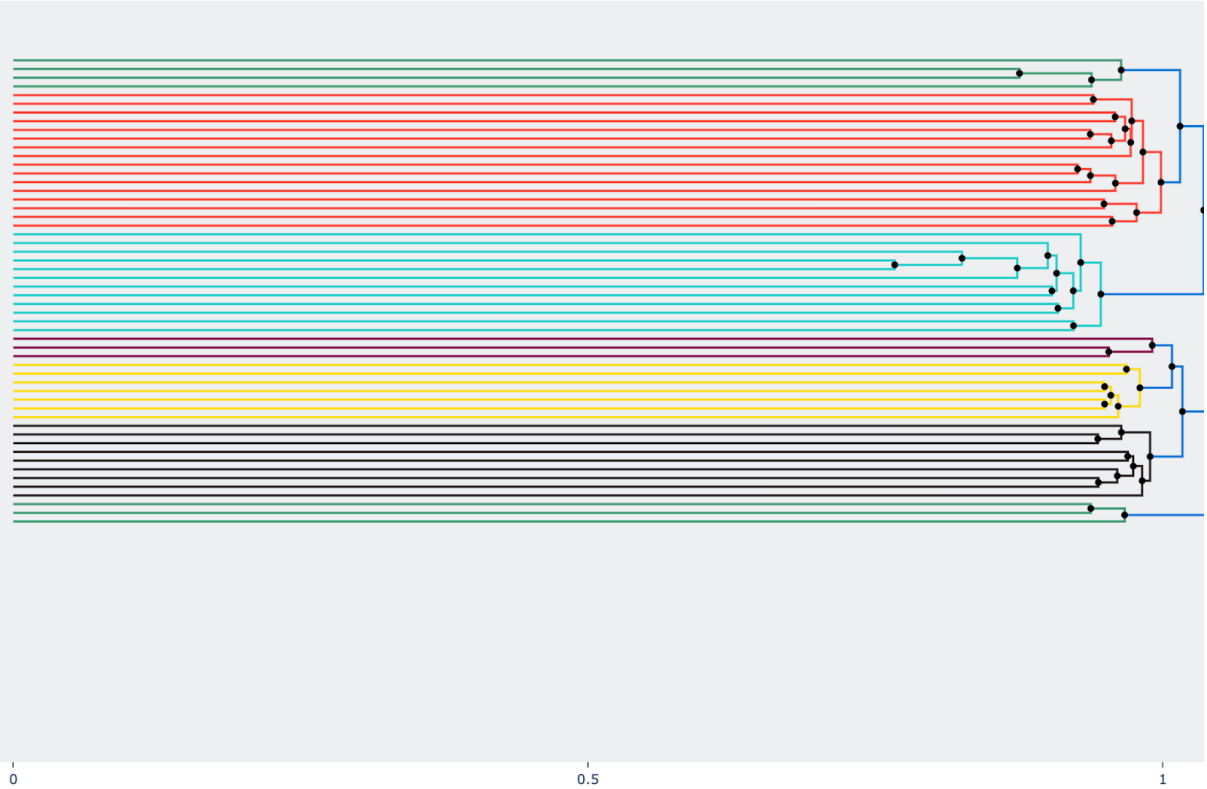

Figure 8: Hierarhical clustering for news titles including stopwords

The clusters in this graph appear more uniform, with the colored sections spanning larger areas horizontally. The branches merge quicker and appear to group more data points into larger clusters earlier. This suggests less granularity in the initial clustering, with groups forming larger, broader clusters early in the hierarchy. Compared to the first graph, the separation between clusters is less defined. Many clusters merge relatively early in the process (for example, the blue, red, and yellow groups), which suggests that the model may be grouping data points that share broader similarities, but with less specificity. This graph shows larger, more general clusters, which may be less interpretable at the individual topic level. The early merging means the topics may be less distinct, and the clustering is based on broader patterns rather than finer details.

5. Discussion

Intertopic distance analysis for two datasets reveals varying behaviours in topic modelling outcomes. For text bodies, topic modelling tends to yield fewer topics, and the identified clusters are more distinct, suggesting clearer separation between topics. This may be due to the richer context provided by the larger text bodies, which allows the model to distinguish between topics with fewer clusters more accurately. Additionally, the inclusion or exclusion of stopwords has a more pronounced impact on the number of topics generated, as the stopwords in full-text bodies tend to influence topic coherence more strongly.

On the other hand, datasets consisting of news titles show a larger number of topics overall, likely because titles are shorter and more varied in content, leading to greater fragmentation in clustering. Despite the larger number of topics, the difference in the number of topics when stopwords are included versus excluded is relatively small, with only about a five-topic difference. This suggests that stopwords in shorter text segments, like titles, have less influence on the overall topic structure than full news bodies, where stopword removal has a greater impact on refining the topic clusters.

Quantitative and qualitative analysis of the datasets indicates that removing stopwords results in a more distinct and interpretable topic distribution when compared to applying topic modelling to an initial dataset that includes stopwords. The influence of stopwords on topic modelling is substantial, as stopwords introduce noise, reducing the clarity of the thematic structure in the data.

An intertopic distance map for a dataset with removed stopwords shows well-separated and distinct topics. Each topic stands out with minimal overlap, allowing for a more precise identification of different themes. This is critical when diverse and unique themes must be recognized within a data corpus.

The distance between clusters in Figure 1 illustrates how removing irrelevant and frequent words helps to highlight the core vocabulary associated with each topic. This results in better-defined clusters, as the algorithm can focus on key terms that genuinely reflect the differences between topics rather than common, high-frequency words like “or”, “and” etc.

In contrast, intertopic distance map, including stopwords, reveals a much less interpretable and chaotic structure. The topics appear to be less distinct, with significant overlap between clusters. This suggests that stopwords hinder the model’s ability to separate topics effectively. The overlapping repetition of topics and chaotic distribution seen in Figure 2 indicate that the stopwords obscure the key themes, leading to difficult-to-interpret clusters. Including common, non-thematic words makes it challenging to distinguish meaningful patterns, ultimately affecting the quality of topic modelling.

Figure 3, which showcased the hierarchical clustering of the dataset where stopwords were removed, supports the same finding. The clusters are more finely distinguished, meaning the data points are grouped into well-defined themes. The more granular separation before clusters merge suggests that removing stopwords improves the initial topic distribution and enhances the hierarchical clustering process.

Figure 4, which retains stopwords, suggests that the data points are grouped into larger clusters. These larger clusters lack the finer distinctions seen in Figure 3. The presence of stopwords here leads to fewer smaller, distinct groups. Instead, data points are aggregated around common words, which may not be meaningful for topic identification.

This results in less interpretable clusters, where topics may appear artificially similar due to the influence of frequent, non-discriminative words.

In Figure 5, the distance between clusters in the hierarchical clustering graph illustrates how removing stopwords allows the model to focus on meaningful vocabulary. The distinct separation between clusters suggests that the data points are more clearly divided into well-defined groups. This reflects the core vocabulary associated with each topic, making it easier to distinguish between themes. The absence of frequent, irrelevant words like “and,” “or,” and similar stopwords allows the algorithm to identify true thematic differences.

In contrast, the intertopic distance map from Figure 6 includes stopwords, which results in a more chaotic and less interpretable structure. The topics appear less distinct, with significant overlap between clusters. This indicates that stopwords hinder the model’s ability to separate topics effectively, leading to overlapping repetition of topics and a lack of clear thematic separation. Including these common, non-discriminative words makes identifying meaningful patterns in the data harder, leading to poorly defined clusters.

Similarly, Figure 7, which represents the hierarchical clustering after removing stopwords, supports these findings. The clusters are more finely distinguished, and the data points are grouped into distinct themes. The more granular separation between clusters before they merge reflects how removing stopwords improves the initial topic distribution and enhances the overall quality of the clustering.

In Figure 8, where stopwords are retained, the data points are grouped into larger, less distinct clusters. The presence of stopwords leads to fewer smaller groups, and the clustering appears less refined. This results in clusters that may seem artificially similar due to the inclusion of frequent, non-thematic words, making the overall structure less interpretable and reducing the quality of the topic modelling process.

Without stopwords, the model produces clearer clusters that are more coherent and easier for humans to interpret. The themes become more transparent, and the clustering process more accurately reflects actual thematic differences in the data. With stopwords, the model struggles to effectively separate topics. This results in confusion and overlap, where topics that should be distinct merge due to the influence of irrelevant words.

Further supporting the quantitative results, a qualitative analysis of the topic clusters reveals that the clusters formed after stopword removal are more interpretable and coherent for human analysis. The topics make more sense within their respective clusters, leading to insights that are not clouded by the presence of stopwords. For example, the analysis of Figure 1 shows that topics 13 and 11 are in the bottom-left quadrant and are united into one cluster. They discuss the military achievements of the Russian and Ukrainian armies on the battlefield. Although, at first glance, they seem to represent adversaries and, therefore, opposing sides of the war and can not be grouped, they are connected by the shared theme of the outcomes of specific military operations. The other example shows a cluster of topics 5, 10, 15, 27, and 28 grouped in the top left quadrant. All topics of this cluster are united by the overarching topic of military activities in different regions, particularly Ukraine and certain Russian regions, mobilization and how it is communicated via social media (especially Telegram). There is a strong emphasis on different regional areas like Ukrainian Kherson and Kharkiv alongside Russian Kaliningrad and Belgorod data, which also indicate the ongoing drafting and participation in the war effort. Another cluster in this quadrant that elaborates on topics 4, 21 and 29 revolves around international involvement regarding Russia's war against Ukraine, particularly focusing on the role of the U.S. and Finland in military support.

Hierarchical clustering and topic modelling often reveal similar data patterns because both methods aim to group similar items based on shared characteristics within the dataset. In text analysis, hierarchical clustering organizes documents or words into clusters based on their similarity, forming a tree-like structure that shows how these groups are related at different levels of granularity. Similarly, topic modelling groups words into topics by analyzing word co-occurrence patterns and identifying latent themes across the dataset. Since both methods rely on the distribution of words and their associations, they frequently uncover comparable clusters or topics, reflecting the same underlying thematic structure in the data.

Both approaches are important because they provide complementary insights that enhance text analysis. Hierarchical clustering visually represents relationships between clusters, helping to explore how broader themes are subdivided into more specific groups. Topic modelling, on the other hand, focuses on the distribution of topics within documents, showing how different themes overlap and interact. By combining these methods, you gain a fuller understanding of both the global structure and the nuanced, probabilistic associations within the dataset. This dual approach also strengthens the validation of findings, as observing consistent patterns across both models increases confidence in the robustness of the analysis.

6. Conclusions

This study highlights the efficacy of topic modelling and hierarchical clustering in analyzing disinformation within the Ukrainian-language dataset, particularly during the full-scale Russian invasion of Ukraine. Several key findings have emerged from the analysis.

Firstly, topic modelling performed better on datasets containing the full bodies of news articles rather than just titles. The richer context provided by full text allowed the models to generate more distinct and coherent topics, enhancing the clarity and precision of disinformation themes. Titles alone, being brief and diverse, led to more fragmented and less interpretable clusters, underscoring the importance of using comprehensive text data for accurate topic modelling.

Secondly, removing stopwords significantly improved topic distribution. By eliminating frequent and irrelevant words, the model could focus on key terms and phrases that genuinely reflect the underlying themes in the dataset. This improvement was evident in both the intertopic distance maps

and hierarchical clustering visualizations, where clearer topic separation was achieved, leading to more interpretable and coherent clusters.

Thirdly, the combination of hierarchical clustering and topic modelling revealed similar data patterns, reinforcing the need to use both methods for a comprehensive analysis. While topic modelling provided insights into the distribution of themes within the data, hierarchical clustering helped understand how these themes were related at different levels of granularity. This dual approach confirmed the findings' consistency and enhanced the analysis's robustness.

This research contributes scientific and practical insights into disinformation analysis by focusing on Ukrainian-language data, an underexplored area in the context of the Russian full-scale invasion of Ukraine. The novelty of this study lies in its application of advanced machine learning techniques like BERTopic and hierarchical clustering on Ukrainian datasets, coupled with the impact of stopword removal to improve topic modelling outcomes. The research also bridges a critical gap by providing more accurate topic modelling for non-English data, addressing the limitations of previous studies that relied on English-language datasets for disinformation analysis in Ukraine.

Future studies can expand on this work by exploring additional methods to enhance the accuracy of topic modelling, such as incorporating sentiment analysis or temporal analysis to observe how disinformation narratives evolve. Additionally, refining models that account for linguistic nuances in the Ukrainian language could further improve the quality of disinformation detection. A broader comparative analysis of different disinformation datasets, including those in other languages or from various regions, could also be pursued to understand global disinformation patterns better and develop more effective countermeasures.

References

[1] U. Ecker *et al.*, "Misinformation poses a bigger threat to democracy than you might think," *Nature*, vol. 630, no. 8015, pp. 29–32, Jun. 2024, doi: 10.1038/d41586-024-01587-3.

[2] S. Lewandowsky, S. van der Linden, and A. Norman, "Disinformation Is the Real Threat to Democracy and Public Health," *Scientific American*, Jan. 30, 2024. https://www.scientificamerican.com/article/disinformation-is-the-real-threat-to-democracy-and-public-health/ (accessed Aug. 11, 2024).

[3] J. Endert, "Generative AI is the ultimate disinformation amplifier | DW | 26.03.2024," *DW Akademie*, Mar. 26, 2024. https://akademie.dw.com/en/generative-ai-is-the-ultimate-disinformation-amplifier/a-68593890 (accessed Aug. 10, 2024).

[4] M. J. Kelley, "Understanding Russian Disinformation and How the Joint Force Can Address It," *US Army War College Publications*, May 29, 2024. https://publications.armywarcollege.edu/News/Display/Article/3789933/understanding-russian-disinformation-and-how-the-joint-force-can-address-it/ (accessed Aug. 11, 2024).

[5] N. Imedashvili, "'Captured emotions' – Russian propaganda," *Rondeli Foundation*, Jul. 11, 2022. https://gfsis.org.ge/blog/view/1512 (accessed Aug. 11, 2024).

[6] M. Hosseini, A. J. Sabet, S. He, and D. Aguiar, "Interpretable fake news detection with topic and deep variational models," *Online social networks and media*, vol. 36, pp. 100249–100249, Jul. 2023, doi: 10.1016/j.osnem.2023.100249.

[7] T. Ahammad, "Identifying hidden patterns of fake COVID-19 news: An in-depth sentiment analysis and topic modeling approach," *Natural Language Processing Journal*, vol. 6, pp. 100053–100053, Jan. 2024, doi: 10.1016/j.nlp.2024.100053.

[8] H. Padalko, V. Chomko, and D. Chumachenko, "Misinformation Detection in Political News using BERT Model," *CEUR Workshop Proceedings*, vol. 3641, pp. 117–127, 2023.

[9] S. A. John and P. Keikhosrokiani, "COVID-19 fake news analytics from social media using topic modeling and clustering," in *Big Data Analytics for Healthcare*, Elsevier BV, 2022, pp. 221–232. doi: 10.1016/b978-0-323-91907-4.00003-0.

[10] USAID-Internews Media Consumption Survey, "Ukrainian media use and trust in 2023," Nov. 2023. Accessed: Aug. 11, 2024. [Online]. Available: https://internews.in.ua/wp-content/uploads/2023/10/USAID-Internews-Media-Survey-2023-EN.pdf

[11] C. Maathuis and I. Kerkhof, "First Six Months of War from Ukrainian topic and sentiment analysis," *European Conference on Social Media*, vol. 10, no. 1, pp. 163–173, May 2023, doi: 10.34190/ecsm.10.1.1147.

[12] R. Marigliano, L. Hui, and K. M. Carley, "Analyzing digital propaganda and conflict rhetoric: a study on Russia's bot-driven campaigns and counter-narratives during the Ukraine crisis," *Social Network Analysis and Mining*, vol. 14, no. 1, p. 170, Aug. 2024, doi: 10.1007/s13278-024-01322-w.

[13] K. Lipianina-Honcharenko, M. Soia, K. Yurkiv, and A. Ivasechko, "Evaluation of the effectiveness of machine learning methods for detecting disinformation in Ukrainian text data," *CEUR Workshop Proceedings*, vol. 3702, pp. 97–109, 2024.

[14] R. Strubytskyi and N. Shakhovska, "Method and models for sentiment analysis and hidden propaganda finding," *Computers in Human Behavior Reports*, vol. 12, pp. 100328–100328, Dec. 2023, doi: 10.1016/j.chbr.2023.100328.

[15] T. Ustyianovych, N. Kasianchuk, H. Falfushynska, and S. Siemens, "Dynamic Topic Modelling of Online Discussions on the Russian War in Ukraine," *Proceedings of International Conference on Applied Innovation in IT*, vol. 11, no. 2, pp. 81–89, Nov. 2023, doi: 10.25673/112997.

[16] V. Petyk, "Ukrainian news," *Kaggle.com*, 2022. https://www.kaggle.com/datasets/zepopo/ukrainian-fake-and-true-news/data?select=news_data.csv (accessed Aug. 01, 2024).

[17] S. Kupriienko, "Ukrainian-Stopwords: The list of Ukrainian stopwords (with numbers) for Data Cleaning and NLP tasks," *GitHub*, 2020. https://github.com/skupriienko/Ukrainian-Stopwords (accessed Aug. 01, 2024).

[18] J. Devlin, M.-W. Chang, K. Lee, K. Google, and A. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT 2019*, 2019, pp. 4171–4186.