UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Ivan Hladkyi

RUSSIAN INVASION OF UKRAINE – TOPICAL EVALUATION OF WORLD NEWS SOURCES WITH MACHINE LEARNING

Master's Thesis (30 ECTS)

Supervisor: Kairit Sirts, PhD

Tartu 2022

Title: Russian invasion of Ukraine – topical evaluation of world news sources with machine learning

Abstract: On the morning of the 24th of February 2022, Russia launched a full-scale invasion of Ukrainian territory. The war erupted in many different places in Ukraine, the Russian armies bombed almost every major city's infrastructure, and as of August 2022, the conflict is still ongoing.
The attention of the whole world is focused on the events unfolding in Ukraine through numerous international news media sources. Different information resources can spotlight the same event from different perspectives depending on factors like audience type, political agenda, degree of speech freedom, etc.
The goal of this thesis was to collect a dataset of news from such resources and then build the pipeline for topic modelling and sentiment classification to analyze the differences and similarities between the news sources. Firstly, we selected several of the most considerable world information resources in our work and collected a dataset of news. Secondly, we created a topic modelling and sentiment analysis pipeline supported by visualization tools. Finally, we analyzed the outcomes of the pipeline and discovered distinctions in the most frequently discussed topics, the sentiment and changes in the popularity of these topics through the timeline. The practical contribution of the thesis consists of several aspects: the novel dataset of news from various sources that spotlight the war, which can be used for further study and the created topical analysis pipeline that consists of the topic modelling and sentiment analysis parts.

Keywords: Russia, Ukraine, war, topic modelling, sentiment analysis, text analysis, dataset collection

CERCS: P170 Computer science, numerical analysis, systems, control

Pealkiri eesti keeles: Venemaa invasioon Ukrainasse – maailma uudisteallikate teema hindamine masinõppega

Sisukokkuvõte: 24. veebruari 2022 hommikul alustas Venemaa täiemahulist sissetungi Ukraina territooriumile. Sõda puhkes Ukrainas paljudes eri paikades, Vene armee pommitas peaaegu iga suurema linna taristut ning 2022. aasta augusti seisuga konflikt veel kestab. Arvukate rahvusvaheliste uudismeedia allikate kaudu on kogu maailma tähelepanu koondunud Ukrainas arenevatele sündmustele. Erinevad infoallikad võivad valgustada sama sündmust erinevatest vaatepunktidest sõltuvalt sihtgrupist, poliitilisest olukorrast, sõnavabaduse astmest jne.
Magistritöö eesmärgiks on koguda uudiste andmestik erinevatest meediaallikatest ning luua tehniline raamistik teemade modelleerimiseks ja meelsuse klassifitseerimiseks, et analüüsida nende meediaallikate sarnasusi ja erinevusi. Kõigepealt valiti töös välja hulk olulisi uudisteallikaid ning koguti nende baasil uudiste andmestik. Seejärel loodi visualiseerimisvahenditega toetatud teemade modelleerimise ja meelsuse analüüsi tehniline raamistik. Lõpuks rakendati loodud raamistikku kogutud andmestikule ning analüüsiti erinevusi enim arutatud teemades ja meelsuses ning samuti muutusi nende teemade populaarsuses ajateljel. Magistritöö praktiline panus koosneb mitmest aspektist: uudne sõda kajastavatest allikatest pärit uudiste andmestik, mida saab ka edasistes uurimustöödes kasutada ning tehniline raamistik teemade ning meelsuse analüüsiks.

Võtmesõnad: Venemaa, Ukraina, sõda, teema modelleerimine, meelsusanalüüs, tekstianalüüs, andmestiku loomine

CERCS: P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine

Content

1. Introduction **4**

2. Related work **6**

3. Technical background **9**

3.1. Topic modelling *9*
3.1.1. Latent Dirichlet Allocation 9
3.1.2. BERTopic 10

3.2. Sentiment analysis *12*

4. Method **14**

5. Dataset creation **15**

5.1. Data collection *15*

5.2. Dataset description *15*

5.3. Preprocessing *17*

6. Pipeline **19**

6.1. Topic modelling *19*
6.1.1. Latent Dirichlet Allocation 19
6.1.2. BERTopic 21
6.1.3. Conclusions 24

6.2. Sentiment analysis *25*
6.2.1. Performance evaluation 25
6.2.2. News sources comparison 26

6.3. Insights and visualizations *27*

7. Discussion **32**

7.1. Implication *32*

7.2. Limitations *32*

7.3. Future work *33*

8. Conclusion **34**

Reference List **35**

Appendix I. **40**

Appendix II. License **41**

1. Introduction

On the morning of February 24th, 2022 president of Russia, Vladimir Putin, announced the beginning of the "special military operation" in Ukraine, and after several moments airstrikes and bombs hit Ukraine in many different cities, including Kharkiv, Kyiv, Dnipro and other big and small places. Simultaneously with the missile shelling, a full-scale inland invasion started in eastern, northern and south parts of Ukraine (Olivia B. Waxman, 2022).

While the war was unfolding, many news sources across the globe drove the readers' attention to this event, and each source was doing it in its own unique approach. The informational resources differ in various ways, starting from obvious geolocation, writing language, audience type and continuing with the presence of hidden political messages, degree of freedom of speech and sensitivity or ignorance of specific topics.

Our study's objective is to gather a dataset of news from these sources and develop a topic modelling and sentiment classification pipeline that would serve as a proof-of-concept solution for finding relationships and dependencies between the extracted topics. In our expectations, we aim to find among the topics the significant events of the war like the Bucha Massacre (UN News, 2022a), Azovstal Siege (Michelle Bachelet, 2022) and Grain Transporting Agreement (UN News, 2022b). We also need to create visualization tools to supply the topic analysis pipeline with instruments for finding insights and conducting future analyses. We want to make the sentiment classification part of the topic analysis pipeline a helpful tool for finding bias in the news sources by comparing the ratios of sentiment classes inside a single news source.

In order to reach the stated goal, we studied the most popular and publicly available worldwide informational resources that have been observing the Russian invasion of Ukraine 2022 and reporting on war news in real-time. We researched commonly used for topic modelling: Latent Dirichlet Allocation (LDA) and BERTopic, applied them both to the problem and compared the results. For sentiment analysis part of the pipeline, we chose BERT based pretrained neural networks. Then we explored the textual information extracted from the selected sources, searching for differences in the news sources' wordings, sentiments of the messages, and topic popularity changes through the fixed timeline of our research - from the beginning of the invasion 24th of February till the 30th of June. Additionally, we searched for insights in the outcomes of the topic analysis pipeline to provide real-world examples of our findings.

Section 2 covers the review of the papers connected thematically and by the nature of research to our work. We compare methodologies and results, then we discuss the differences of the related works. In the end, we give the basis to the novelty of our work by describing several unique aspects that were not covered in any previous research.

We give a brief technical background to the components of our work in Section 3. There we describe the topic extraction techniques used in our words and give an overview of the selected sentiment analysis tools. We explain our choice of approaches and provide an overview of their work alongside visualizations and diagrams for a better understanding.

In Section 4, we provide a high-level description of the developed topic analysis pipeline explaining the connections between its different parts.

The news sources are described in Section 5 along with an explanation of their selection. Following that, we provide a summary of our data collection procedure, a description of the data format and full tabular specification of the dataset broken down by news sources. At the end of this section, we present data preprocessing steps.

Section 6 consists of the results review. It begins with the comparison of the selected topic extraction techniques performance with some examples of their work on the gathered news dataset. It continues with the description of the sentiment analysis results alongside evaluation procedure outcomes and statistic measures per selected news source for each sentiment class. Additionally, in this part of Section 6, we also show the differences in sentiments between the chosen information resources presenting several of our findings. Finally, Section 6 concludes with the analysis of the extracted topics and the description of the visualization tools created for a more detailed and more convenient exploration of the topics.

In Section 7 we give the foundation to novelty of our studies by describing the outcomes and comparing them to the results of the previous studies. This section also covers the unexpected complications encountered during the research, provides our vision of further research and shares the ideas of how we will deliver the currently found insights to the publicity. Finally, in Section 8, we give a conclusion to our research.

2. Related work

In this section, the work will be considered in the context of existing research on the Russian invasion of Ukraine in 2022 and other studies that explore textual information about past conflicts between the two countries. Since the war erupted not far from the time of the beginning of this research and, unfortunately, is still ongoing, not many publications that investigate the current news information field with modern Natural Language Processing (NLP) techniques have been released yet. Still, we will review and compare to our research the methodologies and results of already existing works.

The most recent research that was released in June 2022 (López Ramírez & Méndez Vargas, 2022) provides an overview of the sentiment analysis conducted on the Ukrainian Conflict Twitter Messages dataset as well as the description of the creation of their sentiment classification model based on the Recurrent Neural Network (RNN) architecture. This paper concentrates mainly on the aspect of the design of the RNN-based sentiment extraction model, presenting a solution for the creation of the target sentiment labels for training and evaluation purposes with VADER (Valence Aware Dictionary and Sentiment Reasoner) (Hutto & Gilbert, 2014). This research describes tweets' geographical statistics alongside the most used hashtags and a cloud of words visualizations. The authors used a dataset of messages gathered from all over the world but the part of the main participants of the war (Ukraine and Russia) is relatively small as the number of tweets from the accounts of the corresponding countries represents lesser than 5% fraction of the dataset.

Another similar new paper (Ghosh & Roy, 2022) presents the analysis of the extracted sentiments from the dataset of tweets gathered manually by authors using a set of selected hashtags. The time interval of the research starts before the invasion, at the beginning of the year, when the initial preparations of the Russian army were taking place, and until the initial phase of the conflict – the end of March. The work does not include any implementation details in the methodology section, so it is hard to evaluate the results introduced by authors and use them as a trusted base ground for future studies. Nevertheless, in the results section the work describes the evolution of three sentiment classes (negative, neutral and positive) through the selected timeline, as well as some visualizations of the most popular words in the dataset using 1, 2 and 3-grams (Kapadia, 2019).

The most extensive and close by nature of research to our work is the paper (Tkachenko & Guo, 2019) that describes the examination of how online social network debates unfold in the period from the first Russian intervention in Ukraine in 2014 and the 2019 Ukraine election time interval. The authors of the paper gathered data from Reddit via publicly available python API. They used NLP techniques to comprehend the relevant topics extracted with the LDA model and the evolving sentiment inside the online social network. Moreover, the paper discovers the interactions of individual users' tweets sentiment properties as well as the diversity of the linguistic landscape of Ukraine. Additional attention in the context of Russian and Ukrainian languages is given to the lingua franca, often known as a bridge language, which is a common language spoken by speakers of many native primary languages, many of which

are related linguistically. This work provides a great breakdown of the individual's sentiment evolution alongside the visualization of clusters of LDA extracted topics and clouds of words for each of them.

The following reviewed articles are not as closely related to our research as the previous ones, but they describe solutions to the familiar problems and provide insightful results. The paper about the Ukrainian Maidan Crisis 2013-2014 (Potash et al., 2017) describes the creation of a bias level classification models trained on different articles about one particular event – Ukrainian Maidan. The authors propose an interesting approach of creating target labels of bias for articles based on the discussion types of groups in social media.

The alternative approach to data collection of war news is described in another novel study (Park et al., 2022). The paper describes the authors' way of collecting a dataset of textual media information from the Russian news sources on the VKontakte social network, commonly referred to as "Russian Facebook". The researchers had been working on a problem of dividing the Russian media channels into independent and state-affiliated groups, which they solved using the identification of untrustworthy news sources by Twitter. The authors created the extensive and publicly available dataset of news and described the creation process in detail. In our work we try to extend the idea of news dataset presented by the authors by covering not only the independent and state-affiliated Russian media sources but also Ukrainian, European and US media.

The study of the attempts of the Russian government to shift the public opinion by framing information to support the annex of Crimea by Russia in 2014 (Alzahrani et al., 2018) shows the importance of the time series and frame coding analysis. The results provided in the article describe the approach of early identification of public opinion shifting attempts which in this case led to significant findings of changes in propaganda agenda before the active phase of the conflict highlighting such themes as "Ukrainian fascists in the government" and "brotherly nation calling for help". These outcomes showed the efficiency of the designed solution.

Another similar article (Ngo et al., 2022) uses machine learning to extract the information on public sentiment on economic sanctions from roughly one million Facebook posts made in 109 countries during the Russia-Ukraine conflict. The researchers illustrate the geographic diversity of the majority sentiment and governmental policies.

Finally, we reviewed the article (Justina Mandravickaitė & Tomas Krilavičius, 2014) that analyzes the reflection of dynamics of Crimean conflict 2014 in BBC, RussiaToday, DayKiev and delfi.lt news sources. The authors divide the timeline into three sections (the beginning, the escalation, and the annexation), and they use two different approaches for analysis: co-occurrence networks analysis to reflect rhetorical changes in four different media channels during conflict, and sentiment-based storyline analysis to track sentiment shifts from 2013 to 2014. Results of the research indicate that with utilizing NLP tools, it is relatively simple to monitor the changes in rhetoric and opinion. The authors also present media portrayals of the conflict's events that are proven to help estimate public opinion.

After we finished the literature review for our research, we figured out several unique aspects of our work that had not been done before or were covered only partially:

- In our work, we present the creation of the first dataset of international news spotlighting the Russian invasion of Ukraine gathered from the most significant worldwide informational resources.
- We create a unique topic analysis pipeline which consists of the topic modelling and sentiment classification parts. We also provide an overview of results for both components of the pipeline in the context of the created dataset of war news giving real-world examples of the outcomes.
- In our study, we also describe tools that help to understand the connections between the extracted topics: visualization of the topics' frequencies over timeline and topics hierarchy chart.

3. Technical background

This section includes technical details of the data preprocessing steps and methodologies used to build the topic analysis pipeline: topic modelling and sentiment classification.

3.1. Topic modelling

For the central part of our work, we explored two of the most frequently used approaches in the topic modelling domain of machine learning: Latent Dirichlet Allocation (LDA) and solutions based on neural networks - BERTopic. We will cover both in this section of the chapter.

3.1.1. Latent Dirichlet Allocation

Model description. Latent Dirichlet Allocation (LDA) is a generative statistical model that can extract topics from a set of documents (David M. Blei et al., 2003). Each document is represented as a probability distribution across latent topics, which is the core concept of LDA, and every topic is regarded as a distribution across words that are present in the corpus of documents.

According to the model definition, document-to-topic and topic-to-word distributions are sampled from a probabilistic perspective from two Dirichlet distributions with two main hyper-parameters: α and β. A high α makes the combination of topics associated with each document more uniform across documents. A high β generally makes each topic contain a more uniform mix of words. The Dirichlet distribution itself is a kind of a multivariate probability distribution that represents the probabilities x_i of $K > 2$ distinct categories in a way that

$$\sum_{i=1}^{K} x_i = 1, \qquad where\ each\ x_i \in (0, 1)$$

The entire LDA space is represented by the plate diagram (Figure 1) below:

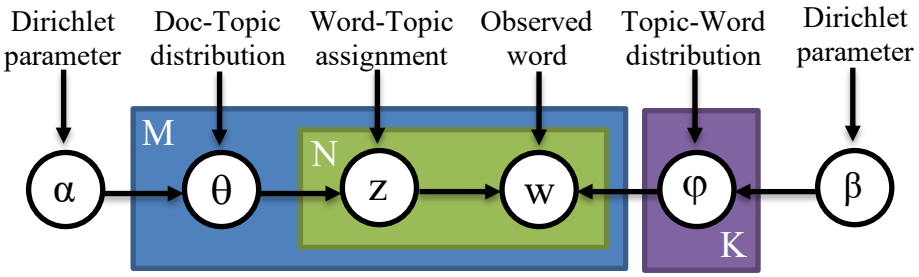

Figure 1. Diagram representation of the LDA model.

Here is a detailed summary of all the components in the diagram:

- The blue box is the number of all documents in a corpus, represented by *M*.
- The following green box is the number of words in a document, given by N.
- The violet box represents the number of topics K.
- w is a word a document.
- z is the latent topic assigned to a word.

- θ is the doc-topic distribution and φ is the topic-word distribution

The training of the LDA models pertains to finding optimal document-topic and topic-word distributions that best explain the data. There are two main inference methods: Gibbs sampling and variational Bayes, which are explained in the original paper (David M. Blei et al., 2003).

Model training. Our LDA model was trained with Python *genism* library (Rehurek, R. & Sojka, P., 2011). In our research we were working with the following input parameters of *gensim LdaModel* class:

- *corpus* – bag of words corpus created in the dataset preparation step,
- *num_topics* – the number of requested latent topics to be extracted from the training corpus,
- *id2word* – dictionary created in the dataset preparation step,
- *passes* – number of passes through the corpus during training,
- *alpha* – α parameter of the document-topic distribution,
- *eta* – β parameter of the topic-word distribution.

Evaluation. We used the coherence score to measure how interpretable the topics are to humans. Our topics are represented as the top N words with the highest probability of belonging to one of the topics. In our research we used CV coherence measure (Röder et al., 2015) which is among the most popular ones. The coherence score measures how similar these words are to each other. Utilizing word co-occurrences, it builds content vectors for the words, and then computes the score using normalized pointwise mutual information (NPMI and) cosine similarity. *Gensim* topic coherence pipeline module uses this metric by default. The CV measure is defined in the interval between 0 and 1, and the more it grows, the more distinguishable and understandable the extracted topics are.

3.1.2. BERTopic

Algorithm description. BERTopic (Grootendorst, 2022) is a topic modelling approach that makes use of transformers and c-TF-IDF to build dense clusters that enable readily understandable topics while preserving key words from the topic descriptions. The algorithm consists of three main parts described in the diagram (Figure 2) bellow:

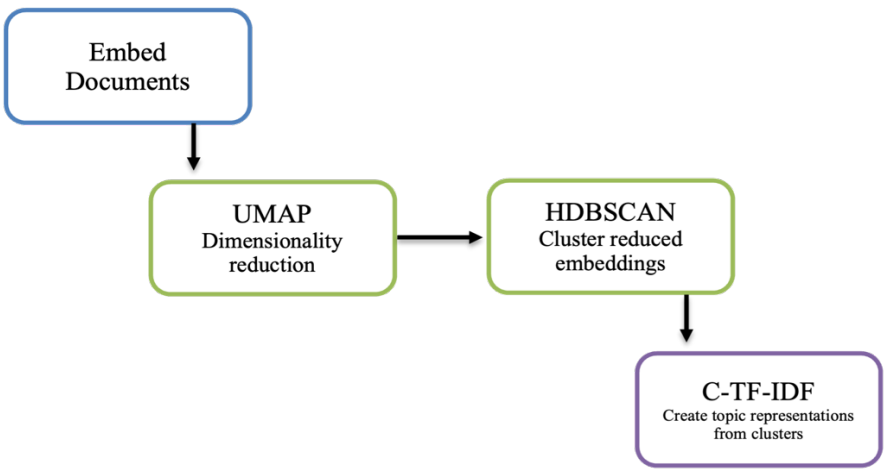

Figure 2. Main components of BERTopic.

Embed Documents. In the first phase, *sentence transformers* (Reimers & Gurevych, 2019), a multilingual framework that creates dense vector representations for each text in our data corpus, is used. Sentence Transformers offers models that are ready to use and have been pre-trained for a variety of languages. These models are excellent for embedding sentences or documents and BERT (Devlin et al., 2019) is typically chosen as an embedding model of preference. Bidirectional Encoder Representations from Transformers, or BERT, is a deep learning model that is based on Transformers (Vaswani et al., 2017). Transformer was the first deep neural architecture for computing text representations that used the attention mechanism as the main building block to consider the connections between words in a sentence. Transformer's attention block generates differential weightings that indicate which other sentence components are more important for understanding the meaning of a problem word. For our case we took *paraphrase-multilingual-MiniLM-L12-v2* model from *sentence transformers* because it has shown the best performance on the multilingual datasets. Any other embedding approaches and libraries like *flair (Akbik et al., 2018)*, *spaCy (Honnibal, M. & Montani, I., 2017), gensim*, *Universal Sentence Encoder (USE) (Cer et al., 2018)*, custom embeddings or even simple TF-IDF can be applied instead of *sentence transformers* models. We configured the following parameters to control this part of the algorithm:

- *language* – either "english" or "multilingual". This parameter selects the pretrained embedding model from *sentence transformers*,
- *nr_topics* – after training the topic model, the number of topics that will be reduced to,
- *top_n_words* – the number of words per topic to extract.

Dimensionality reduction. The dimensionality reduction of the embeddings is the next crucial component of BERTopic. Usually, dense embeddings contain at least several hundreds of components, and many clustering methods struggle to cluster in such a large space (Assent, 2012). The dimensionality of the embeddings can be decreased to a manageable dimensional space (for example 5), in which clustering algorithms can function. Due to its capacity to represent both local and global high-dimensional space in lower dimensions, the most novel dimensionality reduction algorithm UMAP (McInnes et al., 2020) is used inside of the BERTopic by default. BERTopic, however, also functions with other dimensionality reduction techniques, like the well-known PCA. The following set of input parameters was used to optimize dimensionality reduction part:

- *n_neighbors* – it is the number of neighboring sample points that is utilized to approximate the manifold. Smaller numbers produce a more localized perspective of the embedding structure, whereas larger values often produce a more global view. Larger clusters are frequently produced when this value is increased,
- *n_components* – this parameter refers to the dimensionality of the embeddings after reducing them. If this value is increased excessively, clustering method will have trouble clustering the high-dimensional embeddings. If you reduce this amount too much, there won't be enough information in the resultant embeddings to produce accurate clusters.

Clustering. The next step is to group the reduced embeddings into collections of comparable embeddings in order to extract the topics. This clustering procedure is crucial because the more effective the clustering method, the more precise the topic representations will be. As HDBSCAN (Malzer & Baum, 2020) is highly competent of capturing structures with various densities, it is used by default in BERTopic. Because there is no ideal clustering model and the choice of clustering method usually depends on the use case, users of the BERTopic choose some different clustering technique like k-means or Agglomerative Clustering. During the research of the BERTopic clustering step, we have tried to find the most optimal values from the following parameters:

- *min_cluster_size* – it regulates the minimum cluster size and, as a result, the number of clusters that will be produced. By default, it is set to 10. While reducing this number leads to the generation of more micro clusters, increasing it leads to fewer clusters of bigger size,
- *min_samples* – this parameter offers a measure for how conservative you want your clustering to be. More points will be labeled as noise and clusters will be constrained to increasingly denser regions the bigger the number of min samples you specify.

Topic Representation. In order to obtain the representation of topics in this last step, the most relevant words for every cluster are extracted using the class-based TF-IDF technique:

$$\text{for word } x \text{ if class } c, \qquad \mathrm{W}_{x,c} = \ tf_{x,c} \ \times \log\left(1 + \frac{A}{f_x}\right),$$

where $tf_{x,c}$ is frequency of word x in class c, f_x is frequency of word x in all classes and A is the average number of words per class.

3.2. Sentiment analysis

Finding opinions on certain entities inside textual data is the task of sentiment analysis, which is also known as opinion mining. In our case, we want to know how news organisations feel about the conflict between Russia and Ukraine through their writings. The final step of the pipeline research is the conducting of the sentiment analysis on the extracted dataset news and the comparison of the sentiment of different new sources. We wanted to know if these information sources just give the reader the latest news or hide some distinctive political tendency under their messages, and if yes, what is the sentiment of these messages - negative, neutral or positive.

We were trying to find a single multilingual model for this part of the project, but we failed in our search. Instead, we decided to use two pretrained publicly available sentiment classification models from *huggingface transformers* (Wolf et al., 2020) python package, which are both based on BERT architecture:

- *cointegrated/rubert-tiny-sentiment-balanced* – the Russian language model,
- *j-hartmann/sentiment-roberta-large-english-3-classes* – the English language model.

The models take text as input and produce the output - three independent probabilities for the sentiment label – negative, neutral and positive. By default, the predicted sentiment with the highest probability becomes a label for the input news message. We changed the prediction rule to give more importance to negative and positive sentiment predictions. The new sentiment prediction rule works as follows:

1. If predicted negative sentiment probability is higher or equal than 0.5 and predicted positive sentiment probability is lower than 0.5: return negative sentiment label;
2. If predicted positive sentiment probability is higher or equal than 0.5 and predicted negative sentiment probability is lower than 0.5: return positive sentiment label;
3. If none of the above conditions were met, return sentiment with the highest probability.

Now, even if the neutral sentiment has the highest probability, it will not be chosen in the first and second cases.

Because there is no target sentiment labels in our dataset, we need to evaluate the results of the sentiment analysis manually using the following procedure:

For each of the sentiment class (negative, neutral, positive):

1. Take a sample of 100 messages of the current predicted sentiment from the complete news dataset,
2. Give an objective estimate to the predictions and count the number of correctly predicted sentiments in the sample,
3. Return the ratio of correctly predicted messages to the sample size.

4. Method

This section describes the connection between the parts of the developed topic analysis pipeline describing the input data, the order of steps in the pipeline and the outcomes.

The complete pipeline and its parts are described in Figure 3. We will go over each of them giving the overall description of each step of the pipeline together with some of the implementation details.

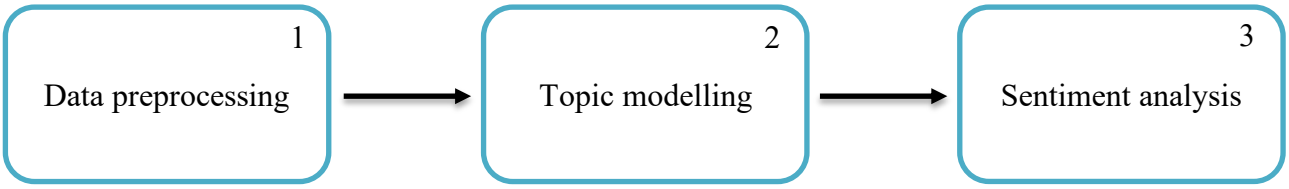

Figure 3. Main components of the topic analysis pipeline.

Firstly, the raw data is preprocessed in the pipeline's first step, which includes cleaning and transforming the information into the input format of the chosen topic modelling technique. The LDA model requires additional preprocessing to compose the filtered dictionary and the bag of words, but this part of the algorithm is described in more detail in Section 6. BERTopic model does not need additional text preprocessing.

Then the data is passed to the topic modelling step of the pipeline. There the algorithm is applied, and the extracted topics can be saved into a .csv file. Additionally, our pipeline provides visualization tools like topic hierarchy and topic frequency through time charts for detailed exploration of the topics and their relationships.

Finally, the extracted topics arrive at the last part of the pipeline - sentiment analysis. In this step, clusters of messages of the extracted topics can be processed through the sentiment classification models to extract the opinion of the selected news sources, which can be used in the comparative analysis of the information media and search for insights.

The codebase of this research and the resulting dataset can be found on the author's GitHub repository page[1].

[1] https://github.com/HladkyiIvan/russian-ukraine-invasion-2022-news-analysis

5. Dataset creation

In this chapter, we describe how our goal of creating the first publicly available dataset of news about the Russo-Ukrainian war 2022 was reached and which methods we used to achieve it. Here, we also extensively describe the selected news sources and give general dataset statistics. In the end of the section, we describe data preprocessing step.

5.1. Data collection

We analysed the possibility of gathering news data directly from the news websites using web scraping techniques such as Beautiful Soup and Selenium Driver but quickly understood that this is not the most efficient way create gather the dataset. There is no common HTML / CSS standard for these websites as they all have unique designs and are implemented in different ways (Figure 4), so creating a separate web scrapping application would require a massive amount of time which we did not possess.

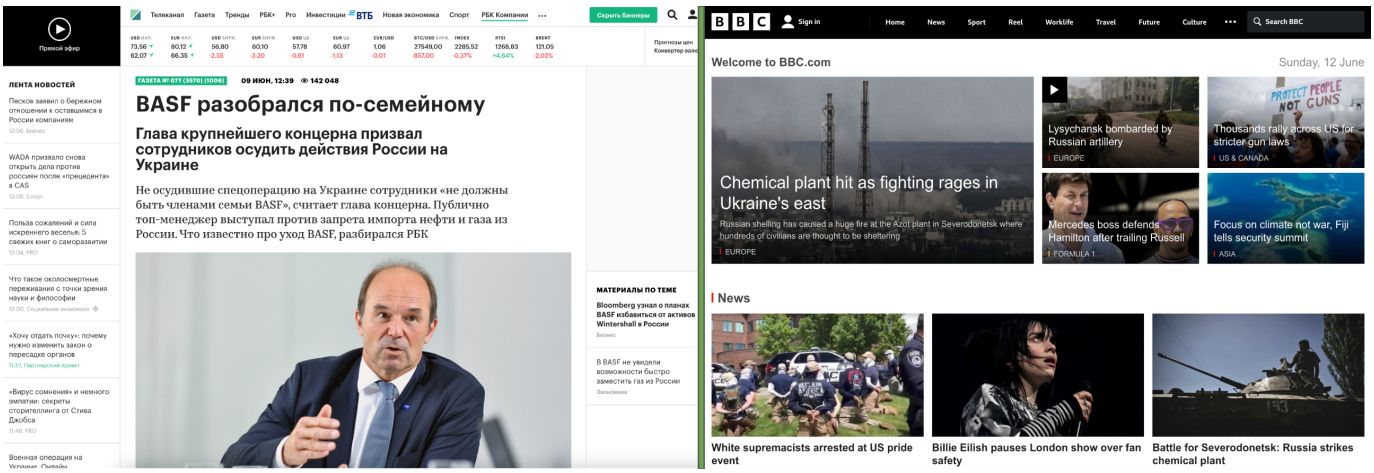

Figure 4. News websites layout examples (RBK – *left*, BBC – *right*).

We found the solution to this problem in Telegram – it is a freeware, cross-platform, cloud-based instant messaging service. The service also provides end-to-end encrypted video calling, VoIP, file sharing and several other features. This messaging platform became popular a couple of years ago, and now more and more news sources create additions to their websites – Telegram channels, to support their spread of information. In these channels administrators are posting short summaries of the news (on average ~150 words) and some additional information such as links to the main webpage of a news company, emails, pictures and videos. Desktop version of Telegram has a built-in channel information export functionality, which allows to download messages, pictures, videos and other content in practical data format – a zip file with JSON messages, png pictures and mp4 videos.

5.2. Dataset description

After a thoughtful consideration and preliminary analysis of available in Telegram world news sources that spotlight the situation in Ukraine, we decided to take the following news sources for our analysis (Table 1):

Table 1. Selected news sources for the analysis.

Id	News source	News source short	Language	Region	Number of extracted messages	Number of subscribers	Telegram link
1	BBC	BBC	RU	Europe	7,282	368K	https://t.me/bbcrussian
2	The New York Times	NYT	ENG	US	532	74K	https://t.me/washingtonpost
3	Washington Post	WP	ENG	US	742	29K	https://t.me/washingtonpost
4	Ukraine24	U24	RU	Ukraine	2,754	485K	https://t.me/ukraina24tv
5	UkraineNow	UNOW	RU	Ukraine	22,209	1.52M	https://t.me/u_now
6	Shocked Ukraine	SU	RU	Ukraine	7,179	220K	https://t.me/voyna_ ukrainavshoke
7	UNIAN	UNIAN	RU	Ukraine	23,574	756K	https://t.me/uniannet
8	RBK	RBK	RU	Russia	9,160	307K	https://t.me/rbc_news
9	TACC	TACC	RU	Russia	31,744	299K	https://t.me/tass_agency
10	RIANews	RIA	RU	Russia	19,947	2M	https://t.me/rian_ru
11	Meduza	MED	RU	Russia	10,640	1.16M	https://t.me/meduzalive
12	Novaya Gazeta	NG	RU	Russia	1,374	443K	https://t.me/novaya_pishet

The total number of messages in the dataset is 137,137. We must mention that created dataset of news also includes topics that have no connection to war, such as “Winter Olympics 2022”, “International Films Awards”, “Weather”, etc. Still, they represent only a minor part of less than 7% of the dataset.

We did not export any graphic information like photos or videos as we were focusing only on textual data in our analysis. The time interval from which the information was extracted is starting from the first day of the invasion, 24th of February and the fixed end date of the research - 30th of June. The following bar chart (Figure 5) describes the proportion of gathered messages across different countries:

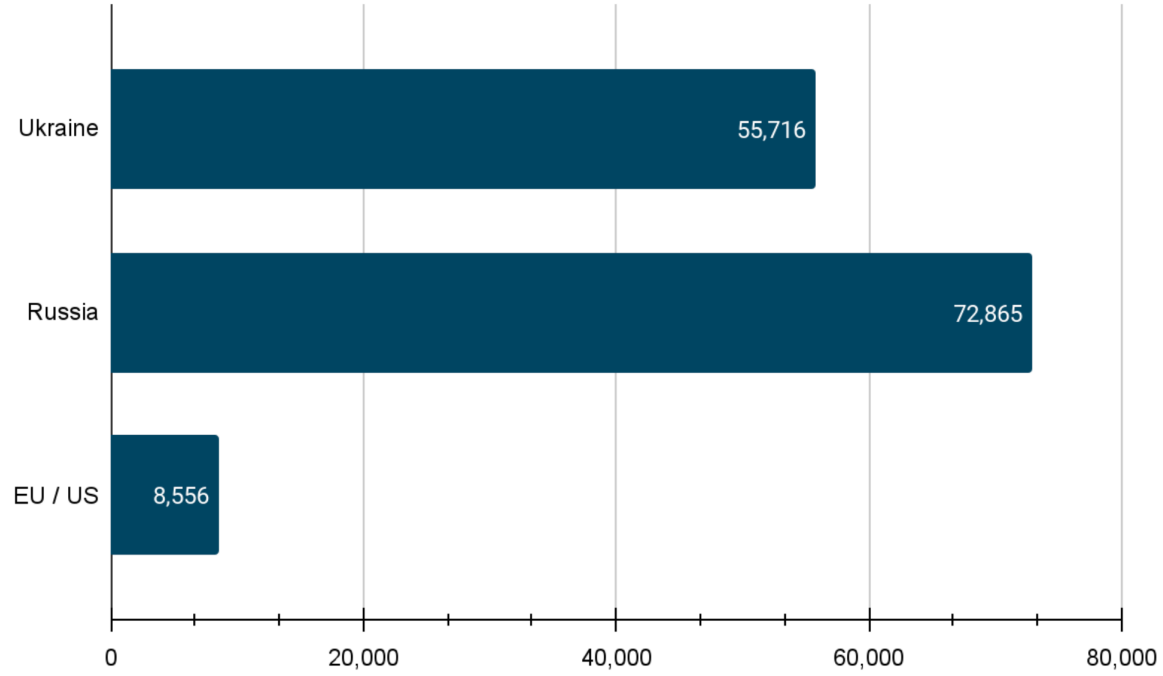

Figure 5. Bar chart: gathered messages per country/region.

The resulting array of datasets is diverse in several different aspects:

- Language,
- Political conviction,
- Target audience, etc.

Another important point is that there are two different types of Russian news sources: those who support the current Russian government (RBK, TACC, RIA News) and oppositional news (Meduza, Novaya Gazeta). The last of the mentioned sources, Novaya Gazeta, had been posting news from the beginning of the war until they were forced to stop working on 27th of March not being able to operate under the pressure of the Russian government. (Roth, 2022)

There are two languages in the composed dataset: Russian and English, and the first one is the leader by the number of messages. One additional aspect to mention is that all the selected Ukrainian news sources, the channels with the biggest audiences, write their posts in the Russian language due to many geopolitical reasons.

5.3. Preprocessing

The next step after gathering raw data is cleaning. Data cleaning's primary goal is finding and eliminating mistakes and duplicating data to provide a reliable dataset. As a result, decision-making is more accurate, and training data for analytics are of higher quality.

In our case, we need to fix several problems caused by the structure of the gathered data:

- Some Russian news channels have constantly repeating messages that appear in some parts of the posts. For example, the proposal to subscribe to the channel or a “foreign agent” notice that all of the oppositional news sources were obliged to add by the

Russian government in the header of each news to inform the reader of their possible untrustworthy (The Moscow Times, 2022);

- We needed to clean messages of emojis and other characters that bear no valuable information for the topic modelling techniques;
- Links, phone numbers, emails and hashtags were also removed as information that is unnecessary for the analysis.

The cleaned dataset was saved in separate .csv files one for each of the news sources. The resulting .csv files contain two columns: the text of the message and its date.

6. Pipeline

In this section of our research, we want to share the application results of the created topic analysis pipeline on the given news dataset and show found insights supported with different visualizations.

6.1. Topic modelling

To understand to what extent we have achieved the defined goal of creating the topic analysis pipeline, we will describe and compare the performance results of each selected topic modelling technique.

6.1.1. Latent Dirichlet Allocation

In the beginning, we must mention that the LDA model requires additional input data preprocessing, which can be described in the following steps:

- Stop words exclusion,
- Text tokens stemming with *SnowballStemmer* from *nltk* python package (Loper & Bird, 2002),
- Dictionary of words creation (the model requires it as an input),
- Extremely rare / common words filtering from the dictionary with the *filter_extremes* function of *gensim.corpora.Dictionary* module,
- Creation of the bag of words – a set of pairs of terms from the dictionary and their appearance frequencies.

After the initial iterations of LDA model creation, this approach has shown a weak capability of extracting topics from the processed dataset – Coherence V measure was on the approximate level of 0.371 ± 0.064 (after 15 runs with a different number of topics between 3 and 15). The extracted topics either seemed not specific enough, too similar or covered to few themes. Therefore, an extension of the training pipeline was created to find the set of hyperparameters resulting in the highest coherence evaluation score. We did a grid search for three main parameters, *num_topics*, *alpha* and *eta*. The hyperparameters tunning step was parallelized using Python standard *multiprocess* library to speed up the computations by separating them into different processes. We obtained the following top 5 sets of hyperparameters sorted by the Coherence V (CV) evaluation metric mentioned in the LDA part of the Section 3 and grouped by the number of topics (Table 2):

Table 2. The results of the grid search.

Number of topics	alpha	eta	Coherence V (CV) evaluation score
5	asymmetric	symmetric	0.514
3	asymmetric	symmetric	0.460
7	asymmetric	0.1	0.438
9	symmetric	0.03	0.429
10	0.1	0.01	0.407

The model fitted with the best set of parameters found with grid search has produced the topics described in Table 3 by the top 7 most probable words of each topic. We have also included the translations for convenience because most of the news dataset was in Russian, so generated topics also mostly consist of Russian words.

Table 3. Topics extracted using the LDA model with the highest performance score and their translations.

Id	Topics	Translated topics
1	обстрел, удар, район, силы, ракеты, армия, продолжает	shelling, impact, area, force, missiles, army continues
2	зеленский, мариуполь, переговоры, министр, путин, нато, оон	zelensky, mariupol, negotiations, minister, putin, nato, un
3	санкции, компании, рубль, сша, банк, доллар, газ	sanctions, companies, ruble, usa, bank, dollar, gas
4	путин, наш, русский, сша, подкаст, журналист, сми	putin, our, russian, usa, podcast, journalist, media
5	суд, территория, дело, человек, плен, задержан, право	court, territory, case, person, captivity, detained, law

We wanted to understand the relationships between the topics extracted with LDA and the internal structure of each separate topic. For this task, we used *pyLDAvis* python package (Sievert, Carson & Shirley, Kenneth, 2014). The package collects data from a fitted LDA topic model to create an interactive web-based visualization which may be saved as a standalone HTML file for sharing purposes, although by default it is meant to be used within an IPython notebook. The next figure (Figure 6) represents the topics extracted by our best LDA model (Table 3):

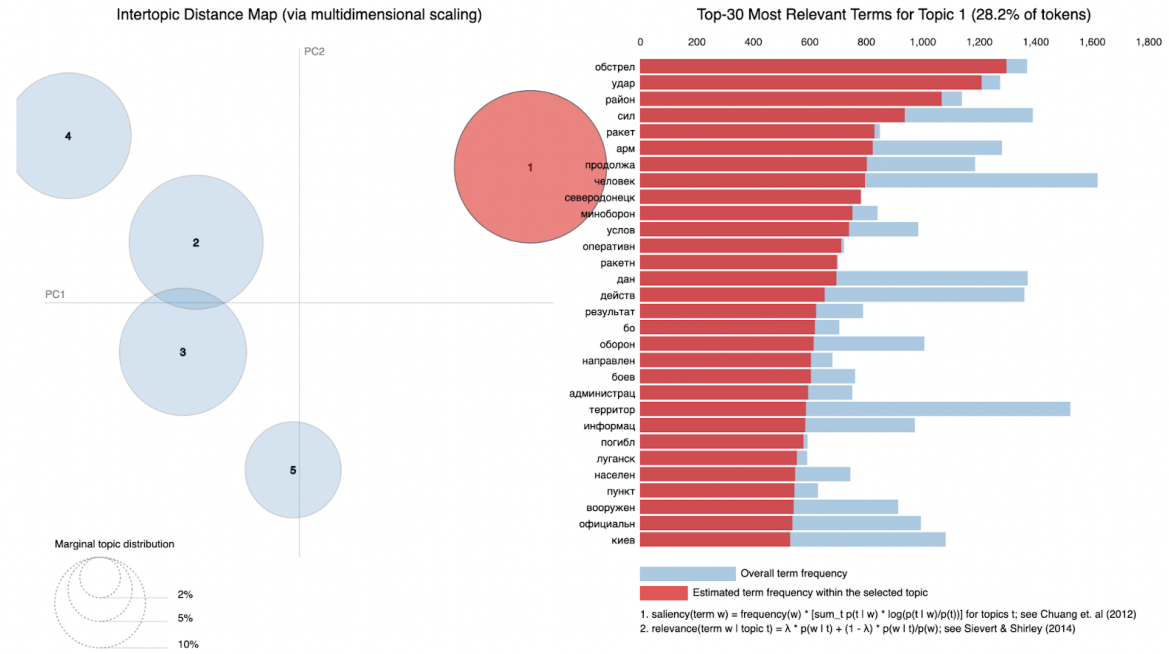

Figure 6. *pyLDAvis* visualization of the topics extracted with the highest coherence LDA model. The chart is cropped by the original design, and terms overlapping is the expected behaviour of the visualization tool.

The visualization is controlled by the following rules:

- Each bubble represents a topic. The bubble size increases as the proportion of messages in the corpus about a particular topic grows.
- The total frequency of each term in the corpus is shown by blue bars.
- The expected frequency of a word for a specific topic is shown by red bars. The term that is most frequently used in messages related to that topic is the one with the longest red bar.
- The distance between the bubbles characterizes how distinct they are from one another.

6.1.2. BERTopic

After tuning the hyper-parameters of the BERTopic model by manual examination of the results and inspection of clusters t-SNE visualization, this topic modelling technique extracted more than a hundred fifty unique topics that cover almost every major event during the war until the 30th of June. The model extracted the topics we used as indicators of meeting the expectations: the Bucha Massacre (UN News, 2022a) and Russian Nuclear Threats (Gordon Corera, 2022). Here is the list of the top 15 topic clusters by the number of messages with the translation column (Table 4):

Table 4. Top 15 by the number of messages topics extracted using the BERTopic model with the highest performance score and their translations.

Id	Number of messages	Topics	Translated topics	Summary title
1	1691	эвакуировать, эвакуацию, эвакуированы, коридоров, мирных, эвакуировали, коридоры	evacuate, evacuation, corridors, peaceful	Civilian evacuation
2	1585	аэрофлота, воздушного, самолет, авиакомпания, аэрофлот, полетов, авиакомпаний	aeroflot, air, aircraft, airline, flights	Airlines
3	1376	крымского, дамбу, водоснабжение, городе, водоснабжения, водой, электроснабжение	crimean, dam, water supply, city, water, electricity	Water supplies
4	1222	российского, дипломаты, болгарии, дипломатических, дипломата, российских, посольство	russian, diplomats, bulgaria, diplomatic, diplomat, embassy	Diplomatic relationships
5	937	россияне, пострадавших, раненых, донецкой, пострадали, погибших, обстреляли	russians, injured, wounded, donetsk, dead, fired	Injuries reports
6	936	возгорания, горят, огонь, пожары, пожарные, пожаров, нефтебазе	burning, fire, fires, firefighters, tank farm	Infrastructure destruction

7	929	вертолет, боеприпасов, вертолетов, авиация, уничтожили, украинских, ракетами	helicopter, ammunition, helicopters, aviation, destroyed, ukrainian, missiles	Aviation vehicles
8	864	балтии, украины, украину, европе, россией, стран, украине	baltic states, ukraine, europe, russia, countries	Eastern Europe diplomacy
9	821	белорусская, белорусы, белорусов, белоруссию, белорусского, белорусских, белорусской	belarusian, belarusians	Belarus
10	764	лисичанск, северодонецк, россия, наступления, войск, армия, украинские	lisichansk, severodonetsk, russia, offensive, troops, army, ukrainian	Donbas fighting
11	756	пользователей, блокировки, доступа, сайты, интернет, вконтакте, сайтов	users, blocking, access, sites, internet, vkontakte	Sites blocking
12	741	харьковская, воздушных, воздушной, днепропетровская, карта, тревоги, областях	kharkiv, air, aerial, dnepropetrovsk, map, alerts, areas	Areal attacks notifications
13	729	китайских, китайская, безопасности, китайские, китаю, тайвань, цзиньпин	chinese, security, china, taiwan, jinping	China
14	593	заболевших, возросло, подмосковье, коронавируса, зараженных, умерли, госпитализированы	sick, increased, suburbs, coronavirus, infected, died, hospitalized	Diseases
15	568	ядерный, ядерного, ядерной, безопасности, российские, чернобыль, электростанции	nuclear, safety, russian, chernobyl, power plants	Nuclear threats

Due to the nature of the gathered news dataset, which consists mainly of similar war-related messages, it isn't easy even for a human to understand to which topic one message should belong. But the HDBSAN clustering part of BERTopic can identify this ambiguous message type as the outliers are considered too generic to belong to any cluster. After our experiments with BERTopic hyperparameters, we managed to tune the model to reduce the number of outliers and increase the number of meaningful smaller clusters of topics. In the end, the number of outliers dropped from 101,123 to 59,995, and the number of topics rose from 112 to 167. After manually cleaning the topics that were subtopics of other topic, which included going one by one and examining the samples of messages from newly formed clusters, we ended up with 150 topics.

We used the following hyperparameters to fit the final BERTopic model:

- BERTopic:
  - top_n_words = 10
  - language = 'multilingual'
- UMAP:
  - n_neighbors = 12
  - n_components = 5
  - metric = “cosine”
- HDBSCAN:
  - min_cluster_size = 35
  - min_samples = 20
  - metric = 'euclidean'

The creators of BERTopic created an interactive visualization of topics reduced into two-dimensional space with *plotly* package. The chart visually resembles the one created for LDA topics with *pyLDAvis*, but, unfortunately, it lacks some functionality of the original visualization like overall and estimated term frequency. On the other hand, *plotly* provides additional features like zoom in and out, download as png, etc. On the next figure (Figure 7) you can see the visualization of the topics extracted with the final variant of the BERTopic with 150 topics:

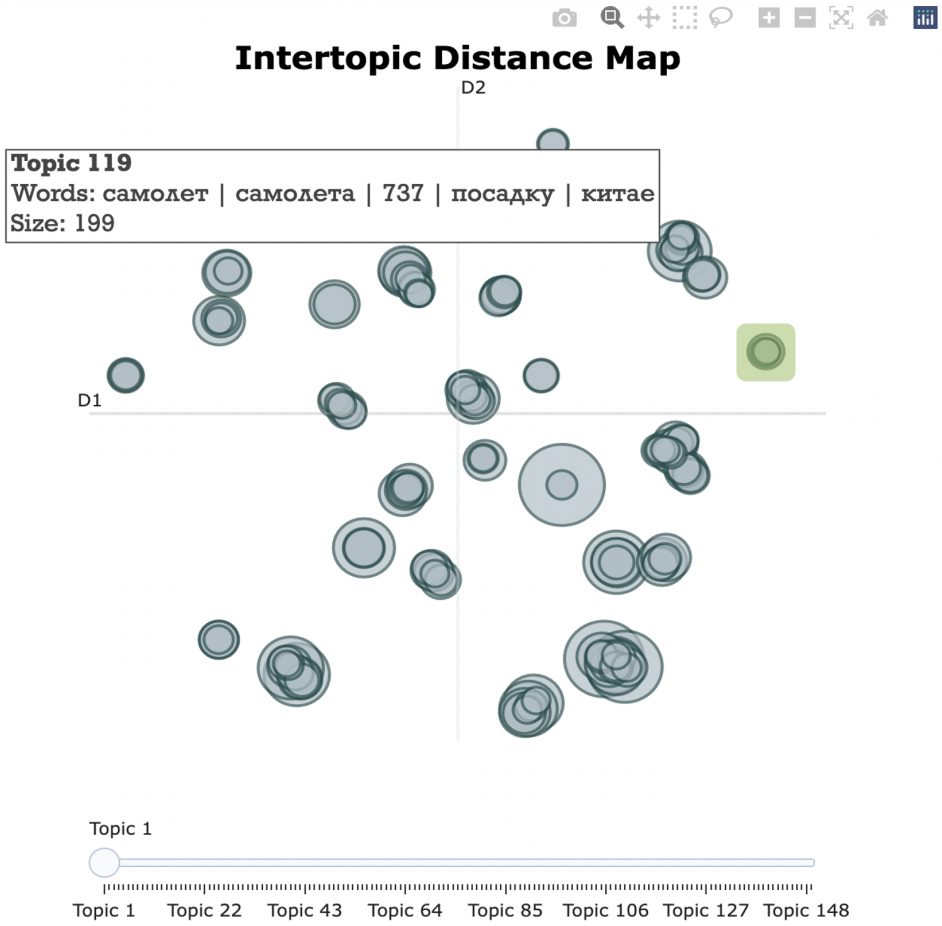

Figure 7. BERTopic visualization of the topics extracted with the final BERTopic model. Banking subcluster is inside the green square.

Many of the two-dimensional projections of the extracted topics seem to overlap, but after zooming in, it turns out that these topics form a subcluster of news themes. For example, the rightmost group of topics inside the green square in Figure 8 in proximity appears to be a "banking" subcluster where you can find such topics as:

- Ban of the Russian banks from the SWIFT system,
- The default of Russian economic,
- Visa and Mastercard suspend Russian operations,
- News about management changes in Russian banks.

We tried to evaluate the model fitted to our news dataset using the coherence metric, as we did with LDA, but we ended up having more than a hundred fifty unique topics, some of which were not as meaningful as the others. In addition to the fact that the coherence metric can sometimes be misleading due to the natural difficulty of the topic meaningfulness evaluation, a new problem appeared – because of the massive number of generated topics, the overall metric could not represent the actual performance of the model correctly. As a result, we decided to manually assess the resulting topic model using the terms used most often in each topic. Additionally, we used the t-SNE technique (van der Maaten & Laurens & Hinton, Geoffrey, 2008) to display the transformed documents where different colours represented clusters and to tune the parameters of the HDBSCAN algorithm by manual investigation of the clusters visualization (Figure 8):

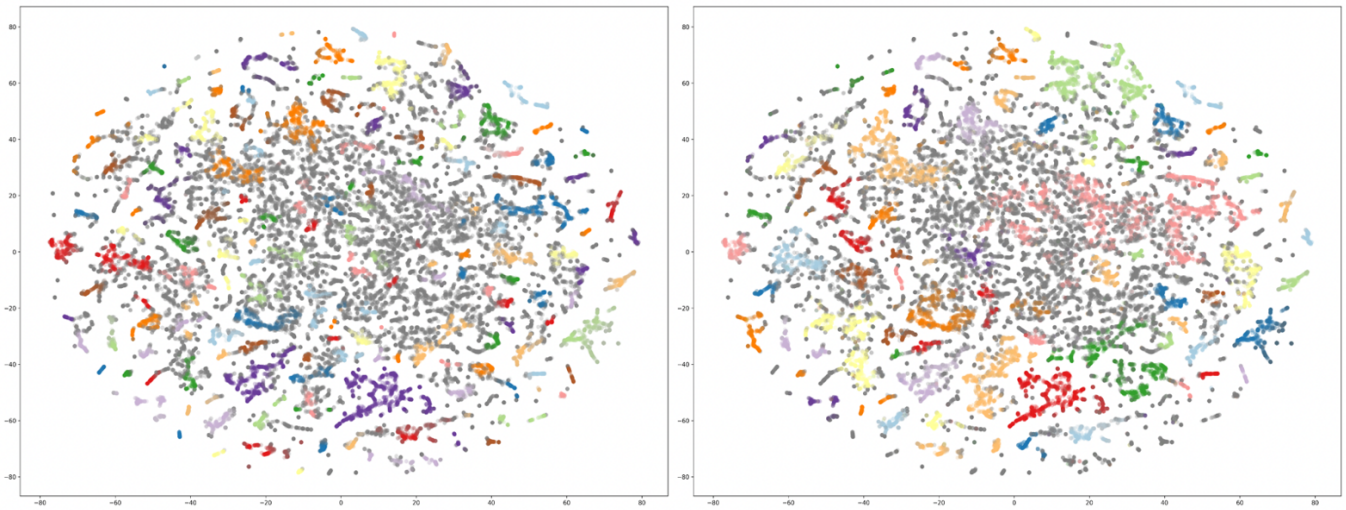

Figure 8. t-SNE visualization of the news documents embeddings. *Left* – HDBSCAN was targeting the precise clusters separation, *right* – HDBSCAN was set to create bigger clusters. Gray points are considered outliers by HDBSCAN.

6.1.3. Conclusions

The results of the LDA model did not follow up with our expectations. The initial idea that the research group kept in mind was to find the topic modelling technique that can extract the subgroups of messages that represent particular events or themes the like the Bucha Massacre (UN News, 2022a) or Russian nuclear threats (Gordon Corera, 2022), but our best LDA model

by CV score is not able to do that having too few number of topics. The model shows its capability of extracting thematically broad topics like sanctions bestowed upon Russia (3rd topic, Table 3) or negotiations between the parties of the conflict and the third-party observers (2nd topic, Table 3), but, unfortunately, neither LDA trained with the most optimal parameters according to the conducted grid search nor other variations of this model with a higher number of topics and lower CV score were unable to achieve the desired results.

BERTopic model, on the other hand, was able to extract a more considerable number of unique topics that touch on various events of the war, giving us the desired outcomes in the end. Our model extracted most of the topics used as indicators of successful topic modelling. We decided to proceed with BERTopic as the preferred model in the topic analysis pipeline, leaving LDA as an alternative.

6.2. Sentiment analysis

6.2.1. Performance evaluation

To prove the statement that selected approach of sentiment analysis is able to extract the correct sentiment, we conducted manual performance evaluation. Firstly, we had been randomly sampling messages from the created dataset of news and giving them target sentiments until we had 100 labelled messages in each of the sentiment class: negative, neutral and positive. Then we ran the sentiment classification part of the pipeline to obtain the predictions. Finally, we create a confusion matrix (Table 6) and calculated the accuracy for each of separate sentiment class and total (Table 5).

Table 5. Sentiment analysis manual evaluation: Accuracy.

Sentiment	Accuracy
negative	72%
neutral	81%
positive	64%
total	**72.3%**

Table 6. Sentiment analysis manual evaluation: Confusion matrix.

		True		
		negative	neutral	positive
Predicted	negative	72	9	1
	neutral	21	81	35
	positive	7	0	64

We find this level of accuracy to be acceptable as it is almost reaching the same level of performance as the other similar approaches described in the related works (López Ramírez & Méndez Vargas, 2022).

6.2.2. News sources comparison

For each of the selected news sources we extracted the sentiment of every message of the source and calculated the percentage of negative-neutral-positive messages. The results are presented in the Table 7:

Table 7. Sentiment analysis general results per news source.

News source short	% of negative messages	% of neutral messages	% of positive messages
BBC	63.3	26.8	9.9
NYT	38	62	0
WP	24	75.7	0.3
U24	55	36.9	8.1
UNOW	55.1	36.9	8
SU	51.2	43.9	4.9
UNIAN	59.7	31.3	9
RBK	45.2	39.8	15
TACC	37.2	52.6	10.3
RIA	39.9	51.8	8.3
MED	49.8	43.2	7
NG	55.4	36.5	8.1

In all sources, most messages are neutral, followed by negative with only a few positive sentiment messages. During the conducted sentiment analysis, we found several components on which the "neutrality" (% of neutral messages) as a measure of objectiveness for each of the news sources depends on:

- The higher the number of topics unrelated to war, the lower the negativity of the news source. For example, news sources controlled by the Russian government (RBK, TACC, RIA) and US informational resources (NYT, WP) both contain such topics as "Winter Olympics 2022" and "International Films Awards" that are not present in any other news sources consist primarily of neutral sentiment messages.
- Less formal news sources with smaller audiences like ShockedUkraine (SU) or NovayaGazeta (NG) appear to have a lower percentage of neutral messages. One possible reason for this could be that these news media present their political opinion on the events unfolding in Ukraine without trying to stay neutral.

Additionally, we divided sentiment predictions by topic and searched for interesting opinion differences between the selected news sources. Here are the most important of them:

- Topics such as "Kaliningrad blockade" (Sytas & O'Donnell, 2022) and "US & Canadian sanctions" that directly affects the affairs of the Russian government have two times higher percentage of negative sentiment messages for news sources controlled by the Russian government (RBK, TACC, RIA) than other information resources.
- On the other hand, the opposite effect is observed when the other topics like "Belorussian involvement" and "Putin's meetings and press releases" – Ukrainian

news sources (U24, UNOW, UNIAN, SU) have increased percentage of negative sentiment messages than the others.

- European, US and Ukrainian news sources (BBC, NYT, WP, U24, UNOW, UNIAN, SU) show positive sentiment on the "Finland and Sweden join NATO" topic (NATO Newsroom, 2022) while all Russian news sources (RBK, TACC, RIA, NG, MED) are expressing neutral sentiment.
- The advances of the Russian army in the Donetsk region (Karolina Hird et al., 2022) in May and June launched a negative sentiment percentage increase for Ukrainian news sources (U24, UNOW, UNIAN, SU). Also, they increased the positive sentiment percentage for news government-controlled Russian information resources (RBK, TACC, RIA).

6.3. Insights and visualizations

Using BERTopic methods, we also created a visualization of topics hierarchy (Figure 9 in Appendix I) which shows how neighbor topics can be grouped into larger clusters. This visualization helps to understand the potential hierarchical structure of the extracted topics and might also help to select an appropriate nr_topics parameter value when reducing the number of topics. The hierarchical clustering visualization shows relevant results on our dataset: it manages to group such logically connected topics as "Gas" and "Oil", "Army casualties" and "Civilian deaths", "Unauthorized protest rally" and "May 9th Victory Parade" into more extensive topics. We have decided not to reduce the number of topics with hierarchical clustering to preserve the topical diversity.

One of the most critical aspects of the expected outcomes was to gain insights into the evolution of different topics over time. By computing the topic representation at each timestep without running the complete model several times, BERTopic enables Dynamic Topic Modeling (Blei & Lafferty, 2006). For each topic and timestep, the model calculates the c-TF-IDF representations that result in a specific topic representations at each timestep without the need to create clusters from embeddings as they were already created. The next step is to adjust a topic representation globally at timestep t by averaging its c-TF-IDF representation with the global representation. This enables each topic representation to lean a little closer to the overall representation while retaining some of its own words.

The resulting *plotly* (Plotly Technologies Inc., 2015) visualization (Figure 10) was created with the built-in functionality of BERTopic by selecting the topics of user preference. This interactive chart gives the opportunity to hover over the points to reveal timestamp representations of the selected topics simultaneously displaying their frequency across time:

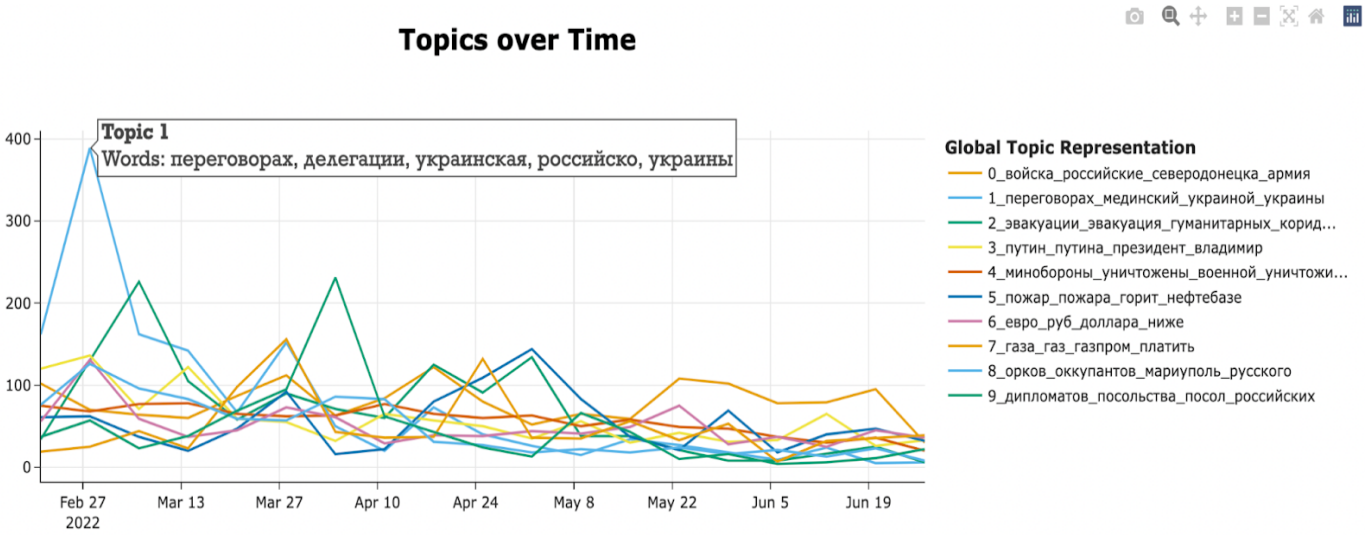

Figure 10. BERTopic visualization of the topics' frequencies over time.

Using "Topics over Time" chart (Figure 10) it is possible to track the number of messages and sudden spikes of public interest to the researched topics. For example, we were able to follow simultaneously the traces of several connected topics "Evacuation", "Negotiations", and "Casualties", which allowed us to discover the simultaneous frequency increase pattern (Figure 11) which works by the following rules:

1. The Negotiations / parley / conference is conducted, and Russian military forces performs artillery shelling / airstrikes in the same day. One of these events becomes the outcome of another but they may happen in different order. Here are two examples:
    a. UN convened urgent meeting due to Russian strike on a shopping center in Kremenchuk (27 June 2022),
    b. During the Ukraine Accountability Conference in Hague (14 July 2022) the topic of claiming Russia the terrorist-state was publicly discussed. As the response Russian army launched artillery attack on Vinnitsa Officers' House.
2. Evacuation calls appears in the news channels as a reaction to the assault.

The same visualization of topic frequencies over the researched timeline also displays some obvious spikes like increased appearance of the "Mariupol" topic when Ukrainian Azovstal defenders were surrounded by the Russian army forces or "Putin" topic's frequency increases which correlated with press releases issued by the Russian President Kremlin Press Secretary.

Figure 11. Negotiations-Casualties-Evacuation frequency increase pattern visualization.

Another important goal of this research was to compare the selected news sources and gain valuable insights into their differences: which topics are most spotlighted in different social channels (Table 8).

Table 8. Top 5 topics (translated) by number of messages per news.

News source short	1st topic (number of messages)	2nd topic	3rd topic	4th topic	5th topic
BBC	Severodonetsk, troops, army, Ukrainian (39)	evacuations, humanitarian corridors (18)	negotiations, Medinsky, Ukraine, (17)	loss of dead, losses, casualties (17)	rally, protest, detained, support (15)
NYT	Severodonetsk, troops, army, Ukrainian (4)	evacuations, humanitarian corridors (3)	dead, burial, grave (3)	billion,Ukraine, million, military (2)	billion,Ukraine, million, military (1)
WP	Severodonetsk, troops, army, Ukrainian (12)	billion,Ukraine, million, military (8)	evacuations, humanitarian corridors (5)	EU, European Commission (4)	Putin, Vladimir (2)
U24	air alert, cover, region (25)	Severodonetsk, troops, army, Ukrainian (11)	Putin, Vladimir (10)	Belarus, Belarusian (8)	fire, burning tank farm (8)
UNOW	air alert, cover, region (109)	evacuations, humanitarian corridors (85)	Putin, Vladimir (79)	explosions, flash, blast (75)	fire, burning tank farm (57)

SU	fire, burning tank farm (122)	invaders, mothers, says, occupiers (72)	rally, protest, detained, support (57)	destroyed, tank, armored, invaders (55)	Putin, Vladimir (24)
UNIAN	Putin, Vladimir (142)	rally, protest, detained, support (102)	Severodonetsk, troops, army, Ukrainian (52)	loss of dead, losses, casualties (49)	destroyed, tank, armored, invaders (48)
RBK	mastercard, SWIFT, banks (43)	directors, post, resigned, chairman (27)	negotiations, Medinsky, Ukraine, Russia (21)	Avtovaz, Volkswagen, cars (20)	Samsung, stores, companies (18)
TACC	aircraft, flights, airlines (151)	negotiations, Medinsky, Ukraine, Russia (96)	fire, burning tank farm (74)	evacuations, humanitarian corridors (70)	mastercard, SWIFT, banks (67)
RIA	aircraft, flights, airlines (68)	negotiations, Medinsky, Ukraine, Russia (54)	coronavirus, Brazil, Indonesia, covid (50)	evacuations, humanitarian corridors (40)	fire, burning tank farm (37)
MED	caution, information, impossible, conflict (106)	evacuations, humanitarian corridors (35)	photographs of the war, photo stories to document (28)	rally, protest, detained, support (27)	aircraft, flights, airlines (24)
NG	aircraft, flights, airlines (9)	Roskomnadzor, vpn, site, access (8)	died, general, death (6)	Instagram, facebook, twitter, extremist (5)	editorial offices, newspapers, news (5)

From the information presented in Table we the following insights:

1. European and US news sources (BBC, NYT, WP) mostly concentrate the attention of their readers of the themes like “Donetsk warfare”, “Evacuation” and “Financial assistance for Ukraine”.
2. Ukrainian news channels (U24, UNOW, UNIAN) share air attack alerts, Ukrainian assaults on Russian ammunition and fuel storage, commentaries on Vladimir Putin speeches and invaders’ elimination related topics.
3. It is visible that informational resources controlled by the Russian government (RBK, TACC, RIA) are trying to mitigate the impact of war by mainly speaking on topics that are not directly connected to the events happening on the battlefront: airlines banned from the EU flight zones, SWIFT system ban and exits of multiple corporations from

the Russian market. But they also post about the evacuation of civilians (mostly deportation of Ukrainians into Russia), rocket airstrikes, and political negotiations.

4. Russian oppositional channels (NED, NG), on the other hand, do post news on such sharp topics as death of the Russian military command, block of the public social media in Russia, video and phono documentation of war events. These news channel cover such common topics as evacuation and problematic airlines communication as well.

To sum up, we found the developed additions to the pipeline to be valuable tools that can help in future studies and lead to multiple findings in the data, which was proven by the presentation of our results.

7. Discussion

In this section, we will think about the obtained results from the point of view of implication of our study, unsolved problems, possible ways of resolving them, and the additions to the conducted work that we did not have time to do or were out of the context of this work.

7.1. Implication

The results of this study have several implications in both research and practice. In our work, we showed the creation process of the novel dataset of news, described the application pipeline for topic modelling and looked at Russo-Ukrainian war events from a fresh perspective discovering patterns in topics' popularity evolution through time, hierarchical structure of the topics and differences in sentiments between the selected group of news sources. Our goal was to create a dataset of Russo-Ukrainian war 2022 news, create topic analysis pipeline and provide an understanding of the informational space in the context of a full-scale invasion.

Before we started working on the analysis part, we needed to create a dataset because there was no publicly available dataset of international news that consisted of the description of the events happening during the Russian invasion of Ukraine in 2022 yet. Also, the studies that provide NLP analyses on this topic mostly use datasets of different social media posts about the war that have other formats and styles from the real news. In this context, our research covers two important gaps in the previous works on this subject:

1. We created a unique, clean dataset of news messages that can be used for future studies on this subject,
2. Our work provided a novel angle of view that discovers the information about the war not from a social media point of view but from the angle of the actual news sources.

Another aspect of our work that has not been covered yet in previous studies about this historical event is the topic extraction and their further analysis. In our research, we provided a scientific overview of the established topic extraction pipeline with a description of alternative methods. Moreover, we present multiple visualizations to describe the insights found during the analysis and give real-world examples to support our findings.

Finally, we also showed in several instances how different sentiments are distributed across the selected news sources and what are the emotional distinctions between the chosen sources on the same topics.

In the context of the defined research goal, we consider our unique dataset and the created topic analysis pipeline a valuable contribution to the growing literature on the understanding of events that are unfolding in Ukraine and the relations between geographically, politically and culturally different information resources.

7.2. Limitations

Several restraints in our work made us decrease the scale of the conducted research that we will cover in sequential order, together with unexpected complications that we have met along the way.

We could not use additional news sources due to the set limit in computational resources. We had a single virtual machine with NVIDIA P100 GPU and 32GB of RAM. We could not increase the number of processed messages through our pipeline without significant losses in inference time, which would suspend the progress of our work.

In the final version of our dataset of news there were less variety of languages and regions as we firstly were hoping to accumulate. The main reason to this is the fact that Telegram, the online communication platform that we used for data collection, is still not so popular in other regions as in Post-Soviet countries. But due to its rapid growth is US and EU we managed to include the opinions of western countries on the current situation in Ukraine.

An unexpected complication was found in the clustering part of BERTopic. The initial results in topic modelling with this technique have shown its weak ability to create consistent clusters of messages, leading to many outliers and unclear topic formations. To find a better tune of the algorithm's parameters, we had to use the t-SNE dimensionality reduction technique to have the possibility to visualize the resulting topics and compare these visualizations during the iterative process of finding the best parameters. This problem was solved by finding better parameters.

7.3. Future work

First and foremost, we need to solve the unresolved problems by spending the additional resources on the following improvements:

- The research of other possibilities of getting news messages from various places of the world different to the locations of the selected news sources should be conducted,
- The potential of finetuning the sentiment extraction pipeline to the needs of our task should be explored, and if this discovery shows positive results, this part of the research should be improved.

Moreover, we are sure that many insights in the dataset are yet to be discovered. To search for novel revelations in data more open, we consider creating a publicly available web service-based solution that would contain all the visualizations and topic exploration tools described in the research. For this, we need to create a backend infrastructural pipeline for updating the dataset with new messages in fixed periods. The described web application would allow individuals to conduct their studies on our dataset comfortably, and it would also serve as a dashboard for tracking changes in the current topics and the appearance of new ones. Currently, we plan to create this solution in a format of dashboards that are commonly used in popular business intelligence analytical tools such as PowerBI or Tableau.

8. Conclusion

In this work, we proposed and implemented the data collection approach to gather news from the worldwide information resources on the example of the Russo-Ukrainian war 2022 dataset of news that we created for our NLP analysis. Secondly, we experimented with the two most widely used topic modelling techniques: LDA and BERTopic. Thirdly, we extracted the set of topics from the gathered dataset and conducted an extensive topical comparison of different news channels supported by visualizations and exploration tools for finding data insights and providing real-world examples. Finally, we explored the involvement of sentiment in the news texts of selected information resources using sentiment classification neural networks. As a result, we have reached our goal of collecting a complete and original dataset of international news messages spotlighting the current situation in Ukraine. Moreover, we also managed to create a proof-of-concept topic analysis pipeline which can provide a possibility to conduct an extended analysis of the behaviour of the extracted topics in different news sources and their intercommunication.

Reference List

Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual String Embeddings for Sequence Labeling. *Proceedings of the 27th International Conference on Computational Linguistics*, 1638–1649. https://aclanthology.org/C18-1139

Alzahrani, S., Kim, N., Ruston, S., Schlachter, J., & Corman, S. (2018, July 1). *Framing Shifts of the Ukraine Conflict in pro-Russian News Media- SBP2018*.

Assent, I. (2012). Clustering high dimensional data. *WIREs Data Mining and Knowledge Discovery*, *2*(4), 340–350. https://doi.org/10.1002/widm.1062

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*, 113–120. https://doi.org/10.1145/1143844.1143859

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., & Kurzweil, R. (2018). *Universal Sentence Encoder* (arXiv:1803.11175). arXiv. https://doi.org/10.48550/arXiv.1803.11175

David M. Blei, Andrew Y. Ng, & Michael I. Jordan. (2003). *Latent Dirichlet Allocation*. 30.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Ghosh, S., & Roy, A. (2022). *Social Media Sentiment Analysis based on Russo-Ukrainian War*. https://doi.org/10.13140/RG.2.2.27245.05605

Gordon Corera. (2022, April 26). Ukraine war: Could Russia use tactical nuclear weapons? *BBC News*. https://www.bbc.com/news/world-60664169

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure* (arXiv:2203.05794). arXiv. https://doi.org/10.48550/arXiv.2203.05794

Honnibal, M. & Montani, I. (2017). *Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.*

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216–225.

Justina Mandravickaitė & Tomas Krilavičius. (2014). *Ukrainian Conflict in Media: Two Approaches to Narrative Analysis*.

Kapadia, S. (2019, August 19). *Language Models: N-Gram*. Medium. https://towardsdatascience.com/introduction-to-language-models-n-gram-e323081503d9

Karolina Hird, George Barros, & Frederick W. Kagan. (2022). *Russian offensive campaign assessment*. Institute for the Study of War. http://dev-isw.bivings.com/

Loper, E., & Bird, S. (2002). *NLTK: The Natural Language Toolkit* (arXiv:cs/0205028). arXiv. https://doi.org/10.48550/arXiv.cs/0205028

López Ramírez, I., & Méndez Vargas, J. (2022, June 14). *A sentiment analysis of the Ukraine-Russia conflict tweets using Recurrent Neural Networks*.

Malzer, C., & Baum, M. (2020). A Hybrid Approach To Hierarchical Density-based Cluster Selection. *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, 223–228. https://doi.org/10.1109/MFI49285.2020.9235263

McInnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* (arXiv:1802.03426). arXiv. https://doi.org/10.48550/arXiv.1802.03426

Michelle Bachelet. (2022). *High Commissioner updates the Human Rights Council on Mariupol, Ukraine*. OHCHR. https://www.ohchr.org/en/statements/2022/06/high-commissioner-updates-human-rights-council-mariupol-ukraine

NATO Newsroom. (2022). *Türkiye, Finland, and Sweden sign agreement paving the way for Finnish and Swedish NATO membership*. NATO. https://www.nato.int/cps/en/natohq/news_197251.htm

Ngo, V. M., Huynh, T. L. D., Nguyen, P. V., & Nguyen, H. H. (2022). Public sentiment towards economic sanctions in the RUSSIA–UKRAINE war. *Scottish Journal of Political Economy*, sjpe.12331. https://doi.org/10.1111/sjpe.12331

Olivia B. Waxman. (2022). *Ukraine rejects Russian demand to surrender port city of Mariupol in exchange for safe passage*. https://www.cbsnews.com/news/ukraine-mariupol-russia-surrender-reject/

Plotly Technologies Inc. (2015). *Collaborative data science*.

Potash, P., Romanov, A., Gronas, M., Rumshisky, A., & Gronas, M. (2017). Tracking Bias in News Sources Using Social Media: The Russia-Ukraine Maidan Crisis of 2013–2014. *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing Meets Journalism*, 13–18. https://doi.org/10.18653/v1/W17-4203

Rehurek, R. & Sojka, P. (2011). *Gensim–python framework for vector space modelling.*

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks* (arXiv:1908.10084). arXiv. https://doi.org/10.48550/arXiv.1908.10084

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408. https://doi.org/10.1145/2684822.2685324

Roth, A. (2022, March 28). Russian news outlet Novaya Gazeta to close until end of Ukraine war. *The Guardian*. https://www.theguardian.com/world/2022/mar/28/russian-news-outlet-novaya-gazeta-to-close-until-end-of-ukraine-war

Sievert, Carson & Shirley, Kenneth. (2014). *LDAvis: A method for visualizing and interpreting topics.*

Sytas, A., & O'Donnell, J. (2022, June 30). Exclusive: EU nears compromise deal to defuse standoff with Russia over Kaliningrad. *Reuters*. https://www.reuters.com/world/europe/exclusive-kaliningrad-row-eu-nears-compromise-deal-defuse-standoff-with-russia-2022-06-29/

The Moscow Times. (2022, July 14). *Putin Signs Expanded 'Foreign Agents' Law*. The Moscow Times. https://www.themoscowtimes.com/2022/07/14/putin-signs-expanded-foreign-agents-law-a78298

Tkachenko, N., & Guo, W. (2019, November 12). *Conflict Detection in Linguistically Diverse On-line Social Networks: A Russia-Ukraine Case Study*. https://doi.org/10.1145/3297662.3365819

UN News. (2022a, April 4). *Bucha killings raise 'serious' questions about possible war crimes: Bachelet*. UN News. https://news.un.org/en/story/2022/04/1115482

UN News. (2022b, July 28). *UN welcomes new centre to put Ukraine grain exports deal into motion*. UN News. https://news.un.org/en/story/2022/07/1123532

van der Maaten & Laurens & Hinton, Geoffrey. (2008). *Viualizing data using t-SNE*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y.,

Plu, J., Xu, C., Scao, T. L., Gugger, S., … Rush, A. M. (2020). *HuggingFace's Transformers: State-of-the-art Natural Language Processing* (arXiv:1910.03771). arXiv. https://doi.org/10.48550/arXiv.1910.03771

Appendix I.

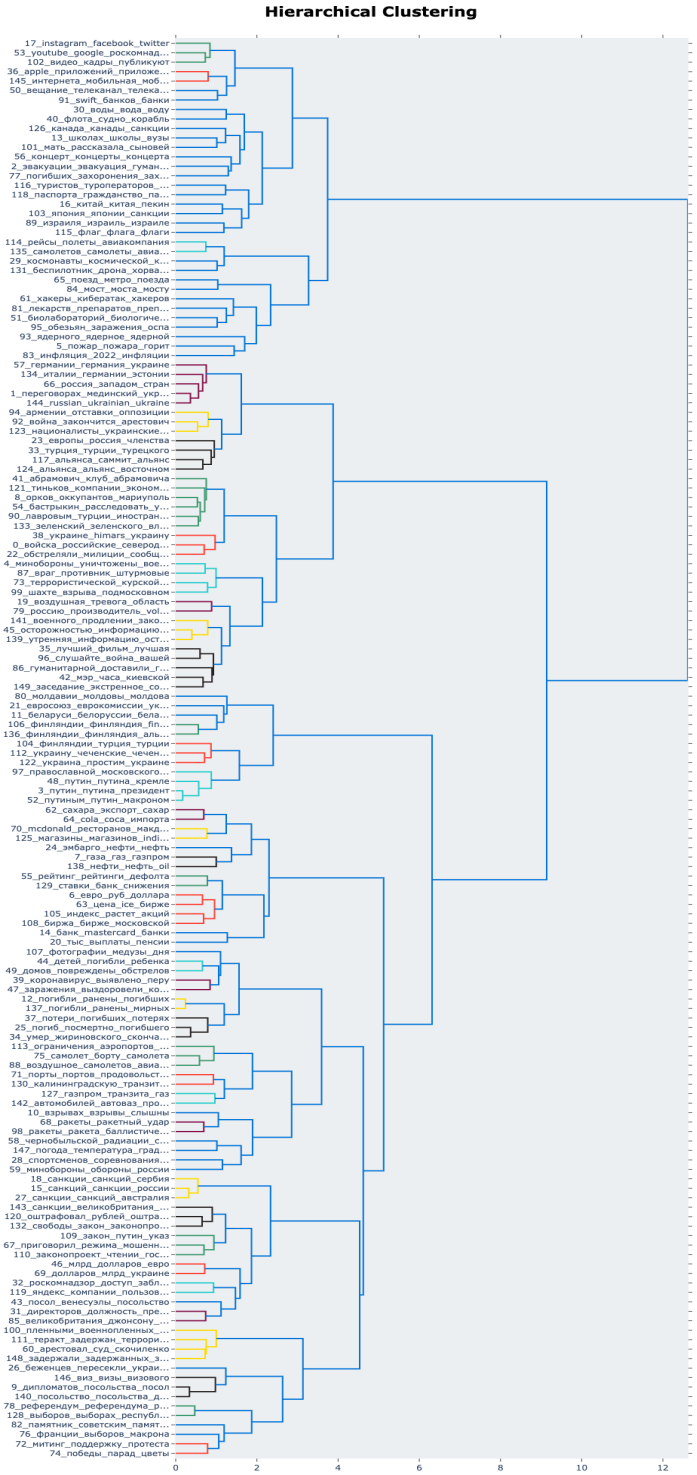

Figure 9. Extracted topics hierarchy tree visualization.

Appendix II. License

Non-exclusive license to reproduce thesis and make thesis public

I, Ivan Hladkyi,
(author's name)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

Russian invasion of Ukraine – topic evaluation of world news sources with machine learning,
(title of thesis)

supervised by Kairit Sirts,
(supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Ivan Hladkyi
Tartu, **06/08/2022**