WAR IN UKRAINE - TWITTER ANALYSIS

A PREPRINT

Jan Skwarek
Faculty of Mathematics and Information Sciences
Warsaw University of Technology
Warsaw, Poland
janskwarek@pm.me

Jakub Kozieł
Faculty of Mathematics and Information Sciences
Warsaw University of Technology
Warsaw, Poland
jakub.koziel@interia.pl

Tomasz Nocoń
Faculty of Mathematics and Information Sciences
Warsaw University of Technology
Warsaw, Poland
tom.nocon20@gmail.com

June 2, 2022

ABSTRACT

In this paper, we examine public's perception on the progression of events in current war in Ukraine, with particular emphasis on its recent eruption to bigger scale on February 24, 2022. On that day, information about Russian aggression spread throughout the world in many different ways, one of them being social media platforms. Identifying huge role of social media in the information flow within countries and among them, we analyze 10 million of tweets from February 21, 2022 to March 03, 2022. We use this data set to determine which topics are most popular and how they correspond with consecutive political and war related events. In order to obtain the best results, we leverage our analysis with usage of unsupervised techniques of Natural Language Processing. Topics acquired with BERTopic provide useful insights about trends and can be used as a measure of popularity of specific events. As a result, we develop application that could facilitate performing fast and easy analysis and could be used in social sciences, for example to investigate aspect of framing.

Keywords Data Science, Natural Language Processing, Twitter, Russia, Ukraine, war, Ukrainian conflict, BERTopic, spaCy, Streamlit, Framing

1 Introduction

It was on August 24, 1991 after the Soviet Union's dissolution when Ukraine declared itself an independent country[Sullivan, 2022]. From then on, the tensions have been high, as many Russians were opposed to the new geopolitical state. The Russo-Ukrainian war began in February 2014 following the Revolution of Dignity[Kariakina, 2019] and was initially focused on annexation of Crimea and war in Donbas[Sullivan, 2022] ongoing since 2014. Russia claimed to have support of Crimean Russian-speaking people, but this was denounced as fraudulent vote[Clinch, 2022] and on the contrary Ukrainian poll shows that some 92% of Ukrainians have negative attitude towards Moscow[Anonymous, 2022]. In 2019 was the election of Volodymyr Zelenskyy, the icon of the ongoing fight for independence. On February 24, 2022 Russian troops concentrated along the borders have begun the invasion[Sullivan, 2022], henceforth, the war erupted to much bigger scale, not seen in this part of the globe for long period of time.

Social media are currently playing a huge role in the information flow within countries and among them. One of the most broadly used platform is Twitter, with 465.1 million active users in April 2022[Kemp, 2022]. Users activity on the platform is reflected in many statistics, one of them being a total number of tweets sent each day fluctuating around 500 million[Sayce, 2019]. Therefore, Twitter appears to be a highly valuable data source for many analyses, providing useful resources not only for studies on past events, but also related to the most recent ones.

Social media activity is also present during the war, and the war itself often becomes the talking point. The importance of information is vivid as both sides of war put on it special emphasis and undertake many actions to propagate occurring events to wider audience as it may influence on behavior not only of regular people, but also country leaders. Furthermore, both sides used disinformation tactics. Russia carried out misinformation campaigns both domestically and abroad. Moreover, there could be observed Ukraine's online offensive with pro-Ukrainian posts flooding platforms thanks to smart hashtags selection[Scott, 2022]. This study takes up on identifying the most common topics coming up in the wide discussions over time. We aim to show the amount of interest among people in specific political events over time, for example: sanctions imposition, military maneuvers and support from different countries. We also examine how well is it possible to recognize war related events through analyzing large amounts of data from social media platforms.

The rest of this paper is structured as follows. Section 2 mentions related work and summarizes it by indicating contributions of this paper. It will be complementary to motivations from introduction. Section 3 presents the data. Section 4 discusses the methodological approach. Section 5 is focused on results. Section 6 summarizes the work and debates about further development.

2 Related work

Twitter as data set for Russo-Ukrainian war investigation By using Twitter, one is able to draw attention of thousands or even millions of people, all thanks to its popularity. Usage of hashtags might be a valuable component of every campaign. That was the case for #SaveDonbassPeople hashtag against military operation in Eastern Ukraine. Usage of this hashtag was initiated by anti-government activists, but soon became contested by supporters of the Ukrainian government. An information war has begun, and both sides tried to use Twitter to propagate their view on conflict. However, hashtag was also used by bystanders to pass relevant information and discuss the conflict rather than producing emotional and political messages[Makhortykh and Lyebyedyev, 2015]. Twitter can be used to examine how antiwar activists and their opponents framed a specific event in time[Nikolayenko, 2019]. This gives opportunity for AI methods of analysis to leverage experiments in social sciences. Twitter is also an opportunity to examine real-time effects of the escalating Ukrainian crisis and correlation between economic indicators and public's perception of the events[Polyzos, 2022]. The need was identified and to facilitate other researchers the data set for investigating Russo-Ukrainian war was created [Chen and Ferrara, 2022].

Impact of Social Media on Political Decision-Making Mining the causes for political decision-making is not new for the filed of political science. May work has been done with focus on long-term policies. However, studying short-term decisions within the same topic has been also previously proposed.Jin et al. [2021].

3 Data

To start working on the project, we first had to collect the necessary data. Unfortunately, there have been problems in obtaining an academic license for the Twitter API. Also, the standard version of the API severely limited us - we could not even scrape tweets older than a week. So we had to scrape the tweets we were interested in ourselves and were vital to observe potentially the biggest shift in opinions about the conflict. We used the snscrape library for this, which is extremely easy to use and powerful, and allows you to scrape data from social networks in an incredibly efficient way.

So we wrote a scraping function, and we could move on to the actual data collection. We tried to collect tweets in many different ways. We have initially limited ourselves to tweets in English only. First, we collected Tweets regarding the various sanctions imposed on Russia. For example, we collected all tweets with the keyword 'Netflix' from the day Netflix introduced sanctions in Russia, etc. The results were not entirely promising. Conducting an exploratory data analysis, for example, we found that a lot of tweets that were intended to be about the SWIFT transfer ban were about Taylor Swift. We then decided to collect the tweets filtered by keywords denoting different geopolitical-military events. For example, we collected all tweets with the keyword 'Bucha' from the day of the Bucha genocide.

However, the problems encountered with this approach resembled those encountered with the previous approach. Tweets were not searched by keyword alone. We have also collected every tweet ever written by current MEPs. However, there were not as many tweets about Ukraine in this collection as we would have liked. In addition, we even collected a small sample of tweets in Russian language. However, we did not consider it sensible to continue with this approach at this stage - perhaps we will return to it later. This allowed us to identify a potential direction for the future. In the end, we decided to work on all tweets containing the keyword "Ukraine" from the first week of the war and three days before the war started. This approach proved to be optimal and produced the most satisfactory results in exploratory data analysis. In total, several million tweets were collected.

4 Methods

4.1 Data preparation

We used the popular spaCy library, used in Natural Language Processing task, to tokenize words. Lemmatization and removal of stop-words (in our case some stop-words were added manually because there were specific for our problem e.g 'Ukraine', 'Putin', 'war' were considered as these type), allowed us to generate word clouds (also for specific parts of speech, for example nouns or verbs).

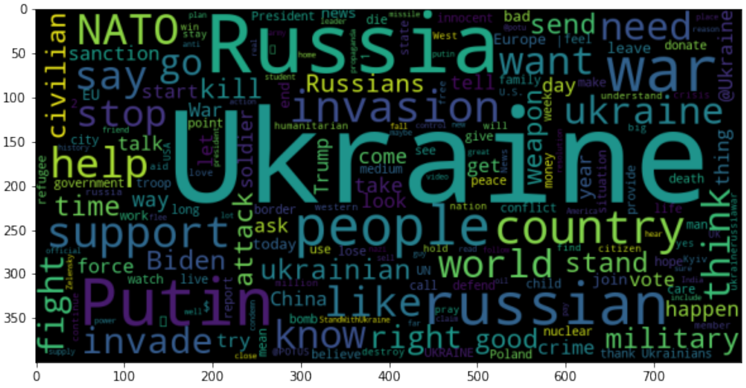

Figure 1: Wordcloud with the most popular words from Twitter on the 2nd of March (from tweets with keyword 'Ukraine').

Our initial idea to analyze the war through the tweets of MEPs, or tweets referring to specific sanctions, fell through due to the strong noise in the collected data. We then decided to change the approach to the problem and analyze tweets with the keyword 'Ukraine' in terms of changes over time.

4.2 c-TF-IDF

$$W_{x,c} = tf_{x,c} \times \log(1 + \frac{A}{f_x}), \tag{1}$$

$tf_{x,c}$ - describes frequency of word x in class c,
f_x - describes frequency of word x in all classes,
A - describes average number of words per class

Analyzing the frequency of occurrences of the most popular words, as well as the words most characteristic for a given day with c-TF-IDF (i.e. standard TF-IDF, only it was broken down by class) gave promising results. The number of occurrences of bios correlated with current events like the sending of Starlinks over Ukraine on 27 February. It is also quite easy to read specific actions like the boycott of Russian products in the USA and Canada.

4.3 BERTopic and CountVectorizer

We used BERTopic to model the themes. This is a topic modelling technique that uses, among other things, transformer-based models (BERT). In our scenario, we used as a model the default one for this task, which is $all{-}MiniLM{-}L6{-}v2$ model. It is an all-round model tuned for many problems and was trained on a large and diverse dataset. In our case, we were able to extract the 500 most significant topics after reducing the number of them. During that part we did not remove stop words yet because we would like to maintain the structure of the sentence. With short fragments of texts such as tweets, it is vital to retrieve the same collocations as n-grams.

After that, we used CountVectorizer, which at the end de-noised their titles. Besides, we used former mentioned c-TF-IDF to create dense clusters, allowing an easy interpretation of the topics, by extracting the important words

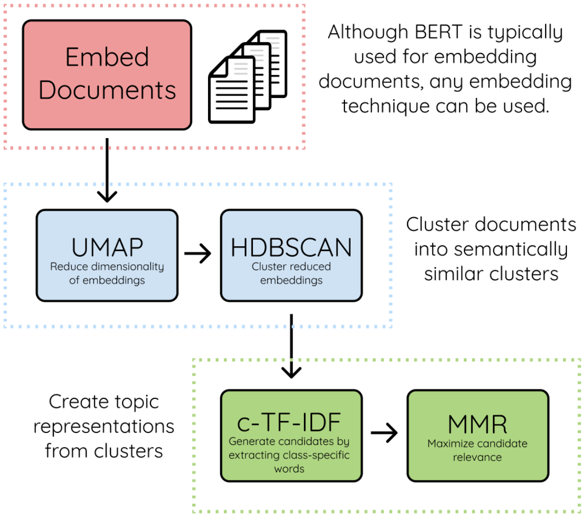
Embed Documents
Although BERT is typically used for embedding documents, any embedding technique can be used.
UMAP
Reduce dimensionality of embeddings
HDBSCAN
Cluster reduced embeddings
Cluster documents into semantically similar clusters
Create topic representations from clusters
c-TF-IDF
Generate candidates by extracting class-specific words
MMR
Maximize candidate relevance

Figure 2: BERTopic main components by Maarten Grootendorst on the Algorithm Documentation Page

described in them. Using this technique, we investigated the distribution of topics over days, and the results were extremely promising.

5 Results

The themes included war or geopolitical events. At first glance, not all of them were directly relevant to our topic. For example, among the most prominent are those about cryptocurrencies, the emergence of which is the reason for the possibility of subsidizing Ukraine with this means of payment, or the topic of China and Taiwan, where, along with the current armed conflict, another possible one has emerged, related precisely to Asian countries.

Finally, we used methods to study the distribution of topics over time, available in BERTopic. This provides a better understanding of how specific events translated into the emergence of a topic in public discussion. For a more in-depth analysis, we decided to build an interactive application.

So we created an application in the Streamlit library. It contains graphs of the most popular unigrams and bigrams over time, also with the option to plot trending words on specific days. In addition, the user can filter any word from the graph. In addition, we have added a hierarchical clustering chart, the most popular words for the given topics and, perhaps most interestingly, a chart of the dependence of topics over time. The application is so interesting that you can use basically any data spread over time on it. We have published the application online and anyone can try it out at this link.

6 Discussion

6.1 Summary

We managed to collect data by using SnScaper. Not only are the collected records the source of ongoing research, but also provides a solid database for subsequent potential projects and outcomes in the field. Concretely, we introduced to the research field the new potential tool where one can filter through the data and search for interesting phrases, topics of the Ukraine conflict, and the distributions of it in the given time.

The topic modeling performed with BERTopic allowed us to divide the dataset into clusters. Moreover, by doing so, we were able to track trends of ever-changing human behavioral responses to the world situation at hand. Perfect examples, which were mentioned in this paper, are Elon Musk's intervention to send Starlink or President Zalenksi's steady increase in popularity during the studied period.

Yet not every observed impact on the real world has been answered and reflected in tweets. Regarding the relationship between public discussion on Twitter and imposed sanctions, we had to change the approach to our research. While the methodology has potential we decided not to continue with the sanction aspect at the given time and in this project we focused on tweets with the hashtag Ukraine only.

Finally, thanks to the prepared tool, data, and other sources of knowledge, we were able to conduct an analysis that showed that the most frequent topics were relevant to the situation in Ukraine, and the words that were used in many cases were about support during a difficult time for this nationality.

6.2 Further research

The potential directions for development are endless. Maintenance and the further development of the application is the logical prolonging of our Twitter NLP project. To begin with, we could improve usability and add additional features to make the app comprise more comparable plots. The app could be faster and more user-friendly, with a more modern layout. In addition, data could be entered directly into the app, making it a true research tool in the field.

Although we used some methods to clean the dataset itself and the topics by introducing advanced techniques, our data was still noisy and the topic of "-1" (meaning stop words for the particular problem), which is not included in this paper, could be adjusted to obtain better results. Using a larger percentage (10%) of the available data results in a 10-fold increase in the number of topics compared to the baseline (1%). Therefore, this area of the project needs to be revised. An aspect that was not addressed in our work due to the fact that we performed unsupervised learning is semantic analysis. Our dataset has no labels, so this approach will require us to use special techniques. The numbers of likes and topic modeling on comments can be considered as a proxy for describing characteristics of twits.

6.3 Limitation

Some of the limitation were mentioned in the data section. The Twitter API was a barrier. However, with tremendous volume of the data, the computational power and memory resources play a vital role in terms of learning models and tuning them right.

References

Becky Sullivan. Russia's at war with Ukraine. Here's how we got here, February 2022. URL http://www.capradio.org/news/npr/story?storyid=1080205477.

Angelina Kariakina. Five years on from 'Maidan', Ukraine's small successes are its real revolution. *The Guardian*, March 2019. ISSN 0261-3077. URL https://www.theguardian.com/commentisfree/2019/mar/06/five-years-ukraine-maidan-revolution-2014.

Matt Clinch. How Russia invaded Ukraine in 2014. And how the markets tanked, January 2022. URL https://www.cnbc.com/2022/01/27/how-russia-invaded-ukraine-in-2014-and-how-the-markets-tanked.html. Section: Europe Politics.

Anonymous. Some 92% of Ukrainians have negative attitude towards Russia - KIIS opinion poll, May 2022. URL https://en.interfax.com.ua/news/general/834975.html.

Simon Kemp. The Latest Twitter Statistics: Everything You Need to Know, May 2022. URL https://datareportal.com/essential-twitter-stats.

David Sayce. The Number of tweets per day in 2020, December 2019. URL https://www.dsayce.com/social-media/tweets-day/.

Mark Scott. As war in Ukraine evolves, so do disinformation tactics, March 2022. URL https://www.politico.eu/article/ukraine-russia-disinformation-propaganda/.

Mykola Makhortykh and Yehor Lyebyedyev. #SaveDonbassPeople: Twitter, Propaganda, and Conflict in Eastern Ukraine. *The Communication Review*, 18(4):239–270, November 2015. ISSN 1071-4421. doi:10.1080/10714421.2015.1085776. URL https://doi.org/10.1080/10714421.2015.1085776. Publisher: Routledge _eprint: https://doi.org/10.1080/10714421.2015.1085776.

Olena Nikolayenko. Framing and counter-framing a Peace March in Russia: the use of Twitter during a hybrid war. *Social Movement Studies*, 18(5):602–621, April 2019. ISSN 1474-2837. doi:10.1080/14742837.2019.1599852. URL https://doi.org/10.1080/14742837.2019.1599852. Publisher: Routledge _eprint: https://doi.org/10.1080/14742837.2019.1599852.

Efstathios Polyzos. Escalating Tension and the War in Ukraine: Evidence Using Impulse Response Functions on Economic Indicators and Twitter Sentiment. SSRN Scholarly Paper 4058364, Social Science Research Network, Rochester, NY, March 2022. URL https://papers.ssrn.com/abstract=4058364.

Emily Chen and Emilio Ferrara. Tweets in Time of Conflict: A Public Dataset Tracking the Twitter Discourse on the War Between Ukraine and Russia. page 5, March 2022.

Zhijing Jin, Zeyu Peng, Tejas Vaidhya, Bernhard Schoelkopf, and Rada Mihalcea. Mining the Cause of Political Decision-Making from Social Media: A Case Study of COVID-19 Policies across the US States. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 288–301, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.findings-emnlp.27. URL `https://aclanthology.org/2021.findings-emnlp.27`.