UDC 004.62

Frequency Analysis of Russian Propaganda Telegram Channels

Kyrylo Kiforchuk[1]

[1] *National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Institute of Physics and Technology*

Abstract

On 24 February 2022 Russia launched a full-scale invasion of Ukraine. In addition to large-scale military operations against Ukraine, many information attacks were organized. These attacks began before the invasion: for several months, Russia has been actively paving the way for the war by launching different types of information operations in cyberspace. As an example, Russian propaganda media were promoting the idea of "Russian world", which calls into a question the existence of Ukraine as an independent state and justifies Russian military aggression. In this work, Russian propaganda Telegram channels were analyzed using term frequency analysis with bag-of-words technique. For this analysis, text data from Telegram propaganda channels was collected and processed. The obtained results revealed different patterns in Russian propaganda against Ukraine via Telegram channels.

Keywords: telegram channels analysis, telegram scraping, telegram data mining, frequency analysis, bag-of-words, term frequency, Russian propaganda, information warfare

Introduction

There is dramatic increase in evidence of information warfare (IW) during past decade. Numerous cases of cyber operations are confirmed all over the world. Many cyber and information incidents are claimed to be linked to Russian government. The list of notable Russian state-sponsored IW campaigns might include [1, 2]:

- Cyber attacks against the Estonian government IT infrastructure in 2007
- Campaign against Ukrainian official government websites, political IT infrastructure, media, social media, critical infrastructure and private IT companies since 2014
- Cyber operation against Warsaw stock exchange in 2014
- DDoS attack on the Bulgarian Central Election Commission in 2015
- Cyber operation against World Anti-Doping Agency (WADA) in 2015
- Influence campaign and cyber intrusions into the U.S. Democratic National Convention (DNC) during U.S. presidential election in 2016
- Attempted cyber intrusion into Norway's Labor Party in 2016
- DDoS attacks against the websites of the government of Montenegro and media in 2016
- Cyber intrusions into the election campaign of French presidential candidate Emmanuel Macron

During active phase of Russo-Ukrainian war Russian state-sponsored cyber attacks became more aggressive and evident. Besides the main goal to damage Ukraine's public, energy, media, financial, business, and non-profit sectors, some of these attacks were targeted to influence Ukrainian people: massive propaganda was launched via mass and social media. This paper aims to prove the presence of such information operation via frequency analysis of social media posts.

1. Information warfare

Despite the term "Information warfare" was defined decades ago, its concept, strategies, methods and goals are evolved drastically. IW has been used throughout history in various forms, from propaganda and disinformation

campaigns to psychological operations and cyber attacks. In ancient times, false information was spread through rumors and propaganda to influence public opinion and gain an advantage in conflicts. During World War II, both the Allies and Axis powers used radio broadcasts and leaflets to spread propaganda and disinformation.

With the advent of the internet and digital technologies, IW has become more sophisticated and widespread. Today, IW is often conducted through social media platforms, online news sites, and other digital channels. Cyber attacks and hacking have also become common tactics used in IW. Overall, IW includes different types of information activities based on its objective, target and implementation [3]:

- Electronic warfare (EW)
- Computer network operations (CNO)
- Military deception (MILDEC)
- Operations security (OPSEC)
- Psychological operations (PSYOP)

EW and CNO are used for making physical and informational damage – destroy, disrupt or delay information and/or information system and/or its components, while MILDEC, OPSEC and PSYOP are developed to cause cognitive damage – mislead or influence people. In this work, we consider methods and techniques of PSYOP – propaganda via social media is one of the most effective ways to achieve the goals of PSYOP.

2. Data mining

One of the biggest challenges in modern IW is the difficulty in identifying the source of false information or cyber attacks. With the ability to remain anonymous online, it can be challenging to hold individuals or groups accountable for their actions. In this regard, Telegram messenger can be a valuable source of text data, as it has a verification system for channels. This system allows to apply for a verified badge to confirm it is an official news feed of a business, organization, or public figure. Once a channel is verified, a blue checkmark badge will appear next to the name of the channel and will provide users with an indication of the channel's authenticity.

Telegram allows users to send messages, photos, videos, and other types of media to each other, making it a rich source of text data. Telegram data mining refers to the process of extracting, collecting, and analyzing data from Telegram. The data can come from various sources, such as public channels, groups, and individual chats. Data mining on Telegram can provide insights into user behavior, sentiment analysis, and trending topics.

Social media data mining process does not depend much on the data source and typically consists of the following steps [4]:

1. Authentication: an optional step, which aims to get access to social media platform.
2. Data collection: gathering raw data from the source.
3. Data cleaning and pre-processing: once the data is collected, it needs to be cleaned to remove any errors, duplicates, or inconsistencies in the data. This step is important to ensure that the data is accurate and reliable.
4. Modeling and analysis: the core of the data mining process, where various techniques and algorithms are used to analyze the data and identify patterns and insights.
5. Result presentation: once the patterns and insights are evaluated, they are represented in a way that can be easily understood by the end-user, such as visualizations or reports.

Figure 1 generalizes consecutive process of the data mining.

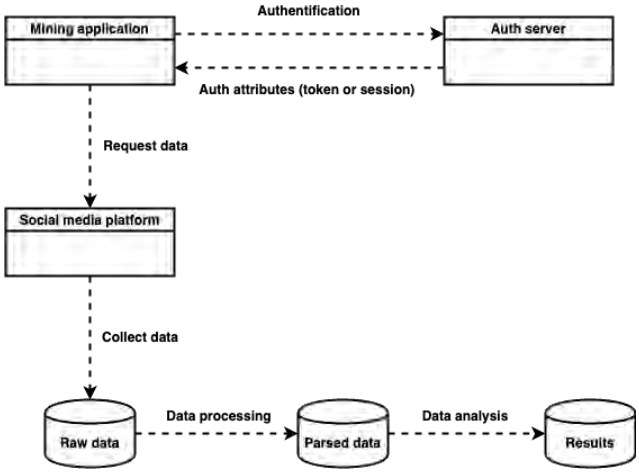

Figure 1: The overall process of social media mining

2.1. Telegram scraping

The process of collecting data from web resource is called web scraping or simply scraping. It involves extracting data from

websites, and it can be used to extract data from public channels, groups, and individual chats on Telegram. There are several ways of implementing Telegram scraping tool:

- Using Telegram application programming interface (API)
- Using third-party API libraries
- Using third-party scripts, services or tools

The last method is the simplest, but least flexible: extracted data is defined by third-party and cannot be customized. On the contrary, the first technique is the most complex and time consuming, but extracted data is fully customizable. Finally, the second method is an intermediate option between the mentioned approaches – it relies on API but provides a convenient way to request it using specific programming language. Considering the above, Telethon Python library use was chosen as the way to collect Telegram text data. Russian propaganda Telegram channels which were chosen as a data source are listed in Table 1.

Table 1
Russian propaganda channels (as of 16 April 2023)

Channel name	Number of subscriptions	Number of collected posts
margaritasimonyan	488 529	12 585
rt_russian	702 715	40 000
skabeeva	192 292	17 147
SolovievLive	1 326 029	40 000

Scraping process was aimed to collect equal amount of text data before and after Russian invasion to Ukraine on 24 February 2022. For achieving this goal, the identifier of last post before 3:00 a.m. UTC on 24 February 2022 was found for each propaganda channel. Then, 40 000 posts were requested from Telegram: 20 000 messages with id less or equal to the "boundary" id and 20 000 messages with greater id. The actual number of collected posts is shown in Table 1. For some channels this number is less than initially requested due to the channel's activity.

2.2. Text processing

For the data cleaning and pre-processing step, standard approaches were used:

- Posts with empty text were removed (photo or video posts).
- Special characters, including emojis and punctuation characters, were omitted.
- All text was transformed to lower case.

Number of posts after data cleaning is shown in Table 2.

Table 2
Number of posts after cleaning

Channel name	Number of processed posts before invasion	Number of processed posts after invasion
margaritasimonyan	9 328	2 134
rt_russian	18 759	18 318
skabeeva	7 688	7 674
SolovievLive	18 335	16 112

As it was already mentioned, the goal of scraping stage was to collect equal amounts of text data before and after 24 February 2022. As shown in Table 2, for 3 out of 4 selected channels equal data sets were extracted. For *margaritasimonyan* Telegram channel number of processed posts before Russian invasion was reduced to the number of posts after invasion, i.e., latest 2 000 posts before 24 February 2022 were taken.

3. Frequency analysis

Posts for each channel were grouped by using several time periods: messages within a day, week and month. This was done to get results at different scales. For the mentioned groups different text characteristics were calculated:

- Average number of messages in specific group is shown in Table 3.
- Time range for the collected posts is shown Table 4.
- Average number of characters per group for different scales is shown in Table 5.

Table 3
Average number of messages in different groups

Channel name	Posts per day (avg)	Posts per week (avg)	Posts per month (avg)
margaritasimonyan	5.61	36.18	152.74
rt_russian	123.8	842.66	3370.64
skabeeva	10.4	61.45	256.03
SolovievLive	114.06	782.88	3444.7

Table 4
Time range for the collected datasets

Channel name	First message date	Last message date
margaritasimonyan	09-02-2021	16-04-2023
rt_russian	13-08-2021	09-06-2022
skabeeva	23-11-2017	16-04-2023
SolovievLive	03-08-2021	31-05-2022

Table 5
Average text size per group for different time ranges

Channel name	Day	Week	Month
margaritasimonyan	2288	14 754	62 299
rt_russian	50 949	348 542	1 394 170
skabeeva	3 304	19 528	81 368
SolovievLive	41 602	285 548	1 256 416

3.1. Bag-of-words model

For using bag-of-words model, additional text processing was done [5, 6]:

- Text lemmatization using *pymorph2* python library [7].
- Stopwords removal using Russian stopwords dataset from *nltk* python library [8].

The order of applying these processing techniques does matter – lemmatization process is more time consuming if being applied before stopwords removal, but more accurate since there is no need to list all forms of stopwords, i.e., it's enough to include only normal form of a stopword to the stopwords dataset. In this work, lemmatization process was applied before stopwords removal.

For each of the selected Telegram channels all text messages were merged into a single text, which was transformed in bag-of-words model. Figure 2, Figure 3, Figure 4 and Figure 5 show 100 most frequent words for each Telegram channel. Bag-of-word model was also built for the previously mentioned scales: day, week and month.

With the use of obtained bag-of-word models it's possible to build time series for the specific word, i.e., word frequency by day, week and month. Analyzing the number of occurrences of individual words at different scales, the following meaningful patterns were distinguished:

- Occurrences of words «лнр» and «днр» in selected Telegram channels are shown in Figure 6 and Figure 7
- Occurrences of words «сша» and «нато» in *SolovievLive* and *rt_russian* channels are shown in Figure 8

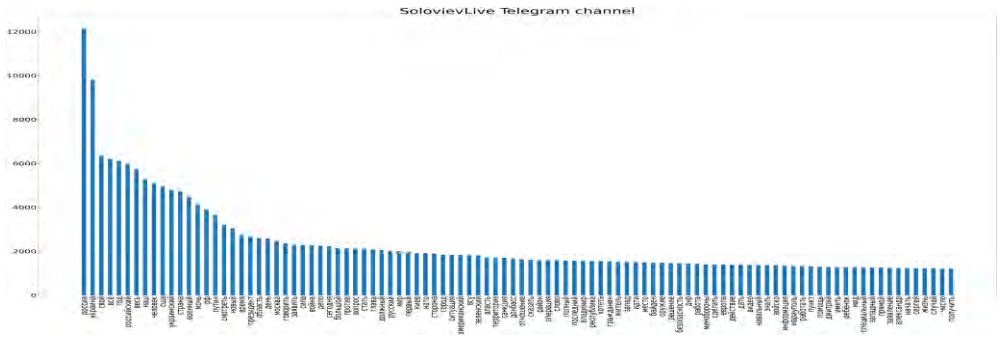

Figure 2: Most frequent words in *SolovievLive* channel during selected time range

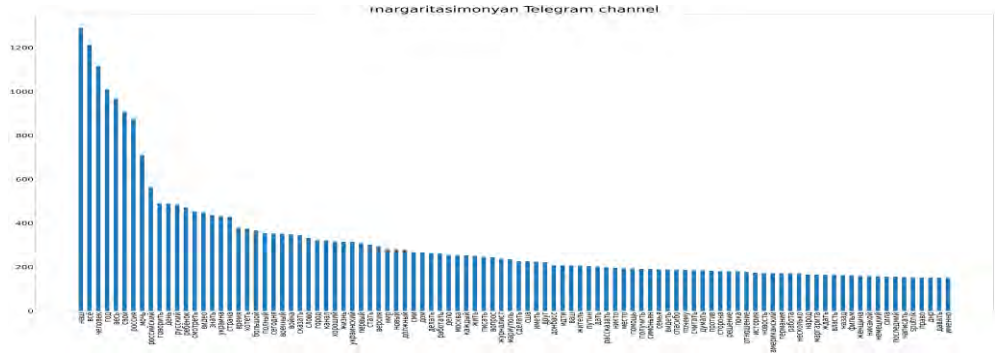

Figure 3: Most frequent words in *margaritasimonyan* channel during selected time range

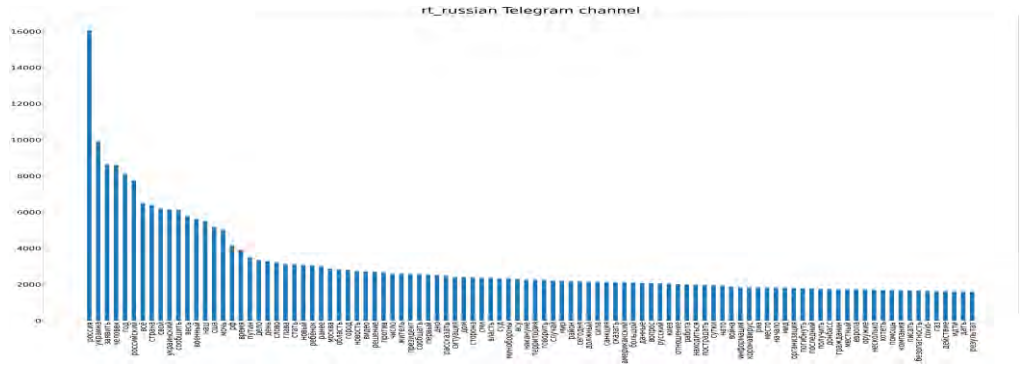

Figure 4: Most frequent words in *rt_russian* channel during selected time range

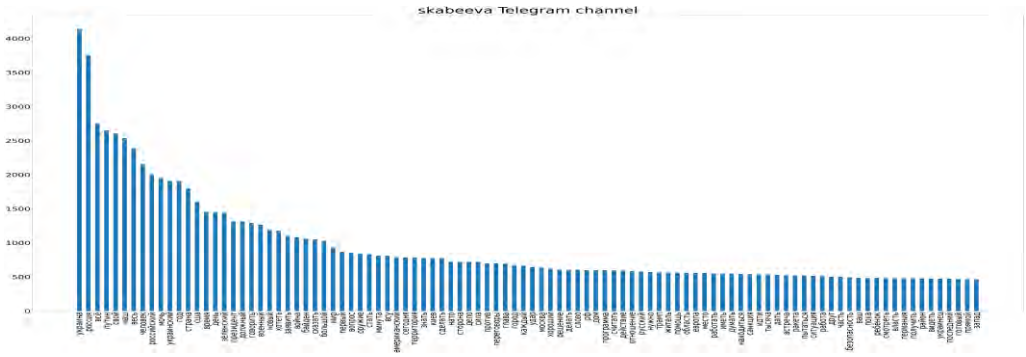

Figure 5: Most frequent words in *skabeeva* channel during selected time range

- Occurrences of «война» vs «операция, сво» words are shown in Figure 9, Figure 10, Figure 11 and Figure 12
- Occurrences of «путин» vs «война» words in *rt_russian* and *skabeeva* channels are shown in Figure 13 and Figure 14
- Occurrences of «народ» vs «власть» words in *rt_russian* channel are shown in Figure 15
- Occurrences of «русский» vs «мир» words in *SolovievLive* channel are shown in Figure 16

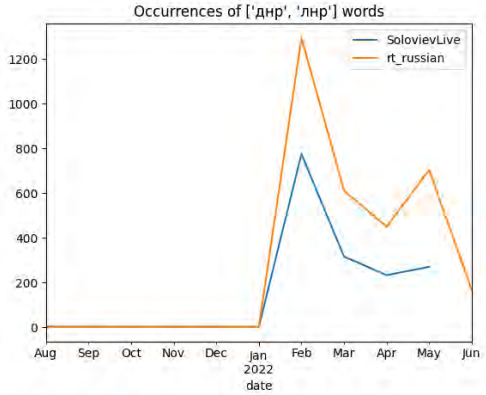

Figure 7: Occurrences of words «днр» and «лнр» in *SolovievLive* and *rt_russian* channels (month scale)

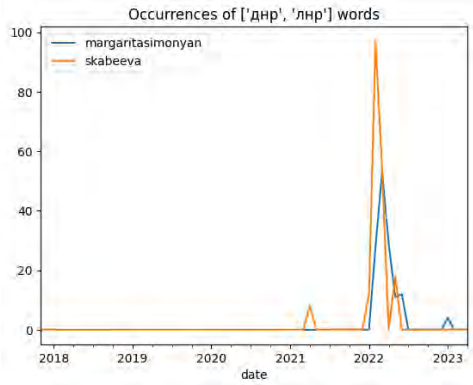

Figure 6: Occurrences of words «днр» and «лнр» in *margaritasimonyan* and *skabeeva* channels (month scale)

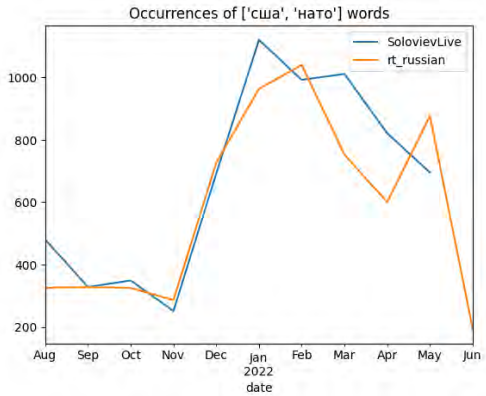

Figure 8: Occurrences of words «сша» and «нато» in *SolovievLive* and *rt_russian* channels (month scale)

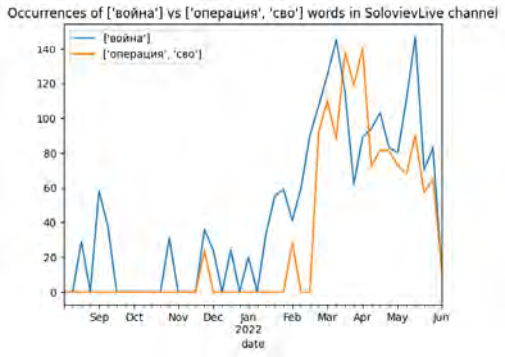

Figure 9: Occurrences of «война» vs «операция» and «сво» words in *SolovievLive* channel (week scale)

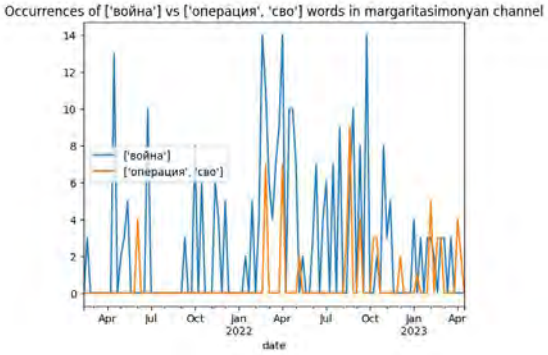

Figure 10: Occurrences of «война» vs «операция» and «сво» words in *margaritasimonyan* channel (week scale)

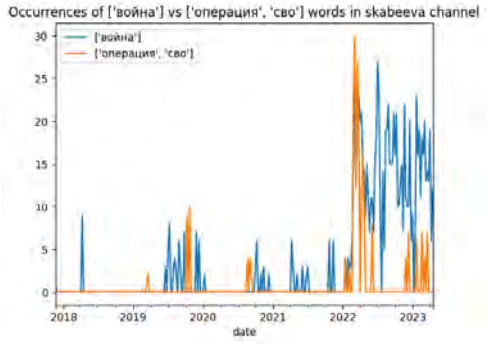

Figure 11: Occurrences of «война» vs «операция» and «сво» words in *skabeeva* channel (week scale)

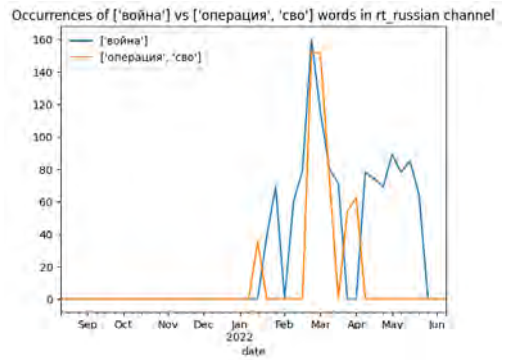

Figure 12: Occurrences of «война» vs «операция» and «сво» words in *rt_russian* channel (week scale)

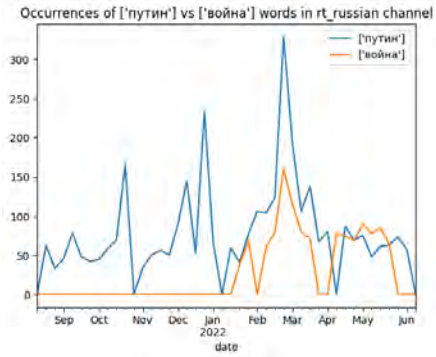

Figure 13: Occurrences of «война» vs «путин» words in *rt_russian* channel (week scale)

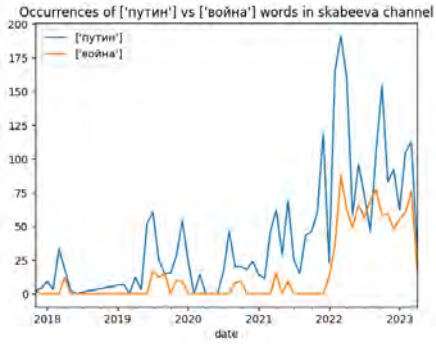

Figure 14: Occurrences of «война» vs «путин» words in *skabeeva* channel (month scale)

Figure 15: Occurrences of «власть» vs «народ» words in *rt_russian* channel (week scale)

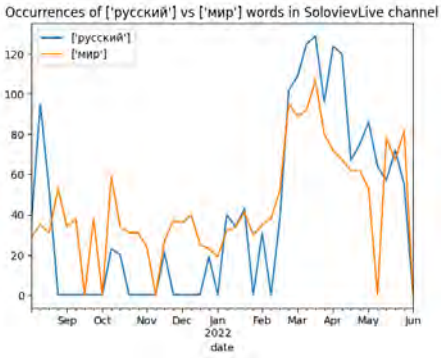

Figure 16: Occurrences of «русский» vs «мир» words in *SolovievLive* channel (week scale)

Conclusions

In this work different approaches of Telegram scraping and data mining were considered. Using selected techniques and methods data from popular Russian propaganda Telegram channels was gathered and processed. The frequency analysis with the use of bag-of-words model revealed different term frequency patterns. The following notable patterns were discovered and shown:

- Despite Russian officials' statements about the need of Russian-speaking people's protection in Donbas region, there was no mentions about these territories before Russian invasion to Ukraine. This pattern is shown in Figure 6 and Figure 7, and can be interpreted as the evidence of a formal pretext for an invasion.
- Additionally, frequency analysis of terms «нато» and «сша» shown in Figure 8, reveals that U.S. and NATO were more discussible than Donbas region.
- Despite Russian officials are avoiding the term «война» and using the term «специальная операция» or «СВО» instead, obtained results show that «война» term is more frequent among propaganda Telegram channels.
- Another example of notable pattern is the correlation in using terms «путин» and «война», «русский» and «мир», which is shown in Figure 13, Figure 14 and Figure 16.
- Figure 15 shows that authority is more discussible than people in Russian propaganda channels.

Obtained results can be used for more detailed and deep semantic analysis of Russian propaganda Telegram channels – bag-of-words model doesn't consider words relation between each other, while another frequency analysis technique called term frequency – inverse document frequency (TF-IDF) can be used on the same datasets for revealing more complex text patterns. Obtained bag-of-words models can be used as datasets features for machine learning text analysis as a part of sentiment or semantic analysis. Created Telegram scraping application can be used for retrieving any amount of data from any public Telegram channel.

References

[1] C. Whyte, T. Thrall, B. Mazanec (Eds.), Information warfare in the age of cyber conflict, Routledge studies in conflict, security and technology, Routledge, 2021.

[2] L. Bilyana, Russian information warfare: assault on democracies in the Cyber Wild West, Naval institute press, Annapolis, Maryland, 2022.

[3] Joint Chiefs of Staff, Joint publication 3-13, Information operations, Defence technical information center, 2012, Incorporating change 1, 2014.

[4] M. Bonzanini, Mastering social media mining with Python, Packt publishing, Birmingham, 2016.

[5] V. Yordanov, Introduction to Natural Language Processing for Text, 2018. URL: https://towardsdatascience.com/introduction-to-natural-language-processing-for-text-df845750fb63.

[6] R Kotubeev, Preparing Russian texts for machine learning with python, 2020. URL: https://python-school.ru/blog/russian-text-preprocessing.

[7] M. Korobov, Morphological analyzer and generator for Russian and Ukrainian languages, in: M. Khachay, N. Konstantinova, A. Panchenko, D. Ignatov, V. Labunets (Eds.), Analysis of Images, Social Networks and Texts, volume 542 of Communications in Computer and Information Science, Springer International Publishing, 2015, pp. 320–332. doi: 10.1007/978-3-319-26123-2_31.

[8] S. Bird, E. Loper, E. Klein, Natural language processing with Python: analyzing text with the natural language toolkit, O'Reilly Media Inc., 2009.