Telegram as an Alternative Platform for Marginalised Groups: Identification of Trending Topics

Justina Mandravickaitė, Eglė Kankevičiūtė, Bohdan Zhyhun, Danguolė Kalinauskaitė, Tomas Krilavičius
Vytautas Magnus University
Centre for Applied Research and Development
LITHUANIA

tomas.krilavicius@vdu.lt

ABSTRACT

AI/ML tools facilitate the identification of trending topics in different media sources and exploratory visual analytics tools leverage means to determine the prevalence of topics and their dynamics. We discuss ongoing research on the application of such an approach to 6 Lithuanian Telegram channels. Telegram is one of the most used instant messaging platforms. It has become popular for its focus on user privacy. On the other hand, there are concerns that this has led to an increase in radical groups on this platform. The selected 6 Lithuanian Telegram channels are focused on alternative positions, news, and opinions in terms of important events and happenings in society. They were chosen by a group of experts based on their content and community activity. The whole dataset consists of ~ 52 thousand posts. Trending topics were identified and explored by keyword, and sentiment analysis as well as by applying BERTopic for topic analysis. BERTopic uses pre-trained language models and class-based TF-IDF to create dense clusters of words which allows us to use them to interpret topics. This enables visual exploration of identified topics as well as topics over time via static or interactive graphs which facilitates monitoring, understanding, and countering of IE threats.

1.0 INTRODUCTION

Texts on social media not only shape public opinion but also reflect it. Technological solutions such as Natural Language Processing as well as Visual Analytics have come in handy in analyzing this source of data to better understand the current topics that dominate on different social media and news sources, as well as the narrative techniques involved in shaping public opinion. Topic analysis has been an increasingly efficient solution to identify prominent public opinion and its sources. The method of topic analysis is particularly useful in the era of misinformation and fake news, hate speech, and political division. This technique analyses textual data to classify them by their topics. Topic analysis can not only help locate the sources of misinformation, but also simply help classify the trending topics in different sources of media, and identify how certain strands of rhetoric shape public opinion.

In recent years, narrative and topic analysis has been used extensively to analyze textual data in social media. The research has helped shed light not only on the common attributes of certain types of text (e.g. [1]) but also on how different types of narratives target audiences (e.g. [2]). Moreover, topic analysis via topic modeling was used to study clones and the fake channels on Telegram [3], information flow during the Coronavirus outbreak [4], detection of depression [5] and cyberbullying [6], and studying fake news [7], etc.

Telegram is an instant messaging platform that has lately seen an increase in its usage. In recent years, users have been relocating from other platforms to Telegram due to its focus on the privacy of the users' data [8]. It has become a popular messaging platform in Lithuania, but concerns have also been raised that the platform has attracted radical groups and has been convenient for further radicalization of users [3]. 6 Lithuanian Telegram channels that focus on 'alternative' news were chosen by experts for the study of trending topics.

The paper at hand employs not only well-established methods of topic analysis such as analysis by keywords but also employs BERTopic. Using pre-trained language models, BERTopic is used for clustering and allows visual analysis of the textual data. Sentiment analysis complements topic analysis in giving "tone" to textual data, expressing a general opinion regarding the topics, events, persons, etc.

This paper analyses a dataset of ~52 000 Telegram posts from 6 Lithuanian channels to study trending topics during the time frame of the last 2 years. It uses the methods of topic identification by keyword and sentiment analysis, as well as the BERTopic method, which invokes pre-trained language models and c-TF-IDF[1] to create clusters of interpretable topics. The variety of methodology helps to achieve broad results addressing different sides (lexical and semantic) of the topic analysis, while applied visualization techniques facilitate topic monitoring, interpretation, and understanding.

The paper is structured as follows: Section 2.0 presents data used for the study, Section 3.0 briefly introduces the methods used for the study, Section 4.0 describes the results achieved while Section 5.0 finishes with discussions and conclusions.

2.0 DATA

Telegram[2] is a well-known messenger where people can share their outlook on life, situations, etc. Since the function of anonymous messages is possible in Telegram [9] and therefore this platform has become popular among radical groups [3], [8], the analysis of the data of these messages becomes a relevant task. For analysis, data from the 6 most active Lithuanian groups and channels focused on the news were exported. The channels are the following: *Diskusijos ir Naujienos @Patriotai* (Discussions and News @Patriots)[3], *Geoengineering @Patriotai*[4], *Naujienos Lietuvoje* (News in Lithuania)[5], *Savas kampas LT* (One's own corner LT)[6], *Slepiamos technologijos @Patriotai* (Hidden Technologies @Patriots)[7]*, Žmonių žiniasklaida* (People Media)[8]. Three of the selected channels belong to a group of channels called *Patriotai* (Patriots)[9], which focuses on alternative positions, news, and opinions in terms of important events in the world and society as well as interpretations of these events via conspiracy theories. Export of the data from selected channels was carried out using the Telethon[10] library, which made the interaction with Telegram easier and made it possible to get the data in a clean JSON format. From each channel, approximately 10 000 messages in the different time frames, encompassing the last 2 years, were extracted, except one of them. Some channels regularly clear their history, and as *Žmonių žiniasklaida* (People Media) was one of such channels, it was not possible to scrape 10 000 messages from it.

[1] Class-based term frequency-inverse document frequency

[2] Available at https://telegram.org/

[3] Available at https://t.me/joinchat/xz2Yqp8hkkwyNDU5

[4] Available at https://t.me/joinchat/AAAAAFjdB0j3cvej1sW4_w

[5] Available at https://t.me/Naujienos_Lietuvoje/3

[6] Available at https://t.me/s/savaskampas

[7] Available at https://t.me/joinchat/LUbPMCuFp_M5MWJh

[8] Available at https://t.me/s/ZmoniuZiniasklaida

[9] Available at https://t.me/s/Patriotai

[10] Available at https://pypi.org/project/Telethon/

For our study, we used only text data (post/message, comment), meta-data of the textual data included additional variables, e.g. message ID, channel ID, media (whether photo/video was included in the post/comment), number of replies, number of reactions, etc. It should be noted that we did not collect personal information of the users, e.g. names. For analysis of trending topics via different methods (keywords and key phrases, topic modeling, and sentiment analysis), careful pre-processing in terms of cleaning our dataset up was performed as we did not want to lose important information, presented in non-standard language which is characteristic for social media texts [10]. Therefore, only tokenization was performed and non-text snippets, such as HTML bits, were removed. Moreover, multiple spaces between words, leading and trailing spaces of the lines were cleaned up and duplicates were removed. Finally, all texts were lower-cased. However, texts were lemmatized for keyword extraction and topic modeling as Lithuanian is a morphologically rich language with a high number of inflective forms, which would have resulted in a sparse data problem [11].

3.0 METHODS

In this section methods used for the study are shortly described. For identification and exploration of trending topics in the selected Lithuanian Telegram channels, keyword extraction and analysis, topic modeling and analysis, and sentiment analysis were applied. The tools, libraries, and methods are discussed in the following subsections.

3.1 Keyword extraction

Keywords and key phrases offer a compact summary of a document. Keywords are independent of any corpus and can be used across several corpora and information retrieval systems. For many different big data activities, including classification, clustering, indexing, and data analysis, it is crucial to extract meaningful keywords. For our study, we chose YAKE[11] (**Y**et **A**nother **K**eyword **E**xtractor) algorithm for the extraction of keywords and key phrases from Telegram messages. It was chosen due to its unsupervised and language-independent nature as well as the high accuracy of results in comparison to other popular keyword extraction tools [12]. YAKE finds potential keywords via statistical features of the text, therefore it does not need additional data, such as a dictionary or thesaurus.

The architecture of YAKE consists of 4 components – pre-processing, feature extraction, term weighting, and generation of candidate keywords [13]. Pre-processing step consists of tokenization, i.e. text is split into words (terms). During feature extraction, 5 types of statistical features are considered: a casing (evaluation of the proportion of particular words starting with uppercase letter in the text), word position (evaluating word importance by position in the sentence, i.e. words at the beginning of the sentence are more important), word frequency, word relatedness to context (evaluation of word’s likeness to a stop word), word different sentence (evaluation of how frequently a word occurs in different sentences) [14]. For term weighting, values of all the statistical features are combined into a single measure. The final step is candidate keyword generation. During this step it is considered that keywords may consist of a group of words, therefore final score for candidate keywords is considered not only for single keywords but for key phrases as well. However, in our study, we decided to focus on single-word keywords. See the 4.1 subsections for the results of the keyword extraction and analysis of selected Lithuanian channels with YAKE.

3.2 Topic modeling

A statistical modeling technique called topic modeling can be used to identify the general topics that appear in a group of texts. This method is helpful for textual data, especially for automating the reading of document collection at the beginning of the research process [15]. Among popular topic modeling, approaches are Latent Dirichlet Allocation (LDA) [16] and Non-Negative Matrix Factorization (NMF) [17], which view

[11] YAKE is available at https://github.com/LIAAD/yake

each document as a group of latent themes and characterize a text as a bag of words. However, these methods have several limitations, e.g. when using bag-of-words representation, they ignore the semantic connections between words, so these representations do not take word context into account [18]. For that reason, text embedding approaches have quickly gained popularity in the field of natural language processing (NLP) as a solution to this problem, especially Transformers' Bidirectional Encoder Representations (BERT) [19].

In this study, we used BERTopic, which is a topic model that uses class-based term frequency-inverse document frequency (TF-IDF) and clustering algorithms to produce coherent topic representations [20]. To get document-level information, it first builds document embeddings using a pre-trained language model. Then the model reduces the dimensionality of document embeddings and creates clusters of semantically related texts each representing a different topic. In the last step, the model creates a class-based version of TF-IDF to extract the topic representation, in this way overcoming the centroid-based perspective [21]. These 3 separate phases make it possible to create a versatile topic model that can be applied to a range of use scenarios, including dynamic topic modeling [18]. See the 4.2 subsections for the results of the topic analysis of selected Lithuanian channels with BERTopic.

3.3 Sentiment analysis

Sentiment analysis automates the extraction or categorization of sentiment (or "tone" – positive, negative, or neutral) from texts using NLP, text analysis, and computational approaches. The main goal of sentiment analysis is to analyze the texts and attribute them to the scores of sentiments [22]. It is used to study people's opinions, attitudes, and emotions toward a particular entity, which can represent individuals, events, or topics [23]. There are dictionary-based and AI/ML-based approaches for sentiment analysis [24]. The former relies on polarity lexicons, which are language-dependent. The latter depends on annotated data for training the model.

For sentiment analysis, we used a sentiment analysis module from Polyglot[12] – a natural language processing toolkit that supports multilingual applications. Polyglot sentiment module has polarity lexicons for 136 languages, including Lithuanian. The sentiment is determined by linking words to their counterparts and related words in covered languages, starting with English sentiment words [25]. The word context is not taken into account in this method, therefore, this implies that while creating scores, negations or modifiers are ignored [26]. However, Polyglot was chosen due to its wide coverage of languages, especially covering Lithuanian (Lithuanian still lacks terms of resources for sentiment analysis) and being easy to use. The scale of the words' polarity consists of +1 for positive words, and -1 for negative words. Neutral words will have a score of 0. See the 4.3 subsection for the results of the sentiment analysis of selected Lithuanian channels with Polyglot.

4.0 RESULTS

In this section results of trending topics extraction and analysis in the selected Lithuanian Telegram channels are presented. Results achieved by different Natural Language Processing and Visual Analytics methods are discussed separately. This allowed visual exploration of identified trending topics via graphs which facilitates monitoring and understanding of relevant topics.

4.1 Keyword extraction results

Keyword extraction results by YAKE tool demonstrate important single-word keywords for separate Telegram channels selected for this study. Figures 1-6 show the 20 most relevant keywords for each channel.

[12] Polyglot is available at https://github.com/aboSamoor/polyglot

The more relevant the keyword is, the lower the score it has, i.e., the most relevant keywords reside on the left side of the Figures.

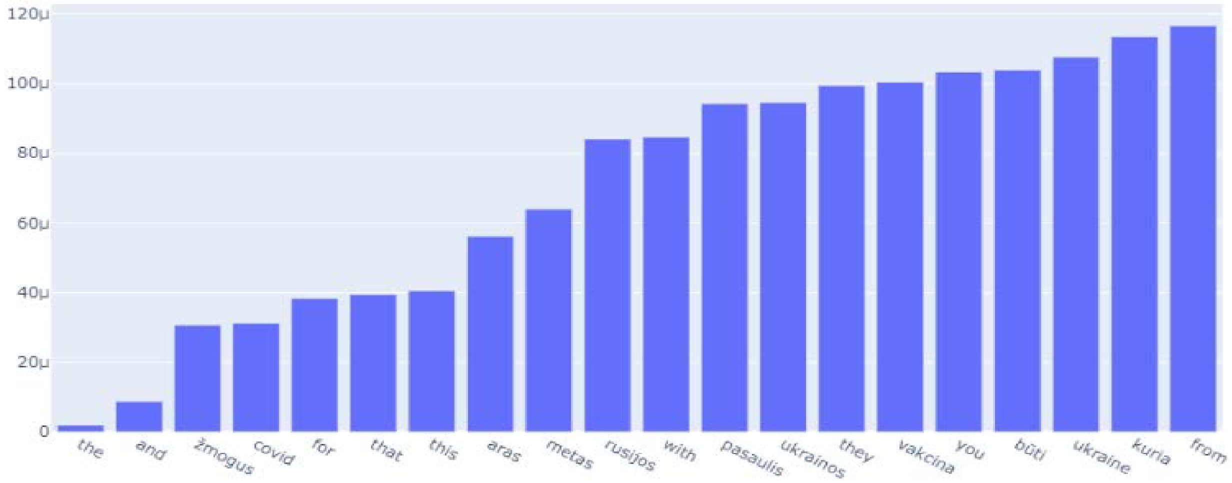

Figure 1: Telegram Channel *Diskusijos ir Naujienos @Patriotai*: Keywords (2020 July and 2022 April)

Figure 1 shows the 20 most relevant keywords for *Diskusijos ir Naujienos @Patriotai* (Discussions and News @Patriots) channel. It reveals that the content of this channel has Lithuanian mixed with English, therefore some English stop words (‘the’, ‘and’, ‘for’, ‘that’, ‘this’, ‘with’, ‘they’, ‘you’, ‘from’) were attributed to significant relevance as we expected the content to be in Lithuanian only and therefore, we defined language as Lithuanian before running YAKE. Also, we can see the words *Ukraine*. Among the most relevant keywords, besides the already mentioned, were *žmogus* (human), *covid*, *aras* (Lithuanian Police Anti-Terrorist Operations Unit), *metas* (time/period), *rusijos* (Russia), *pasaulis* (world), *ukrainos* ((of) Ukraine), *vakcina* (vaccine), *būti* (to be), *kuria* (either present tense of the verb ‘to create’ or instrumental case of the pronoun ‘which’; it is not clear without the context). So, these keywords hint at 2 important topics – Covid-19 vaccines and war in Ukraine.

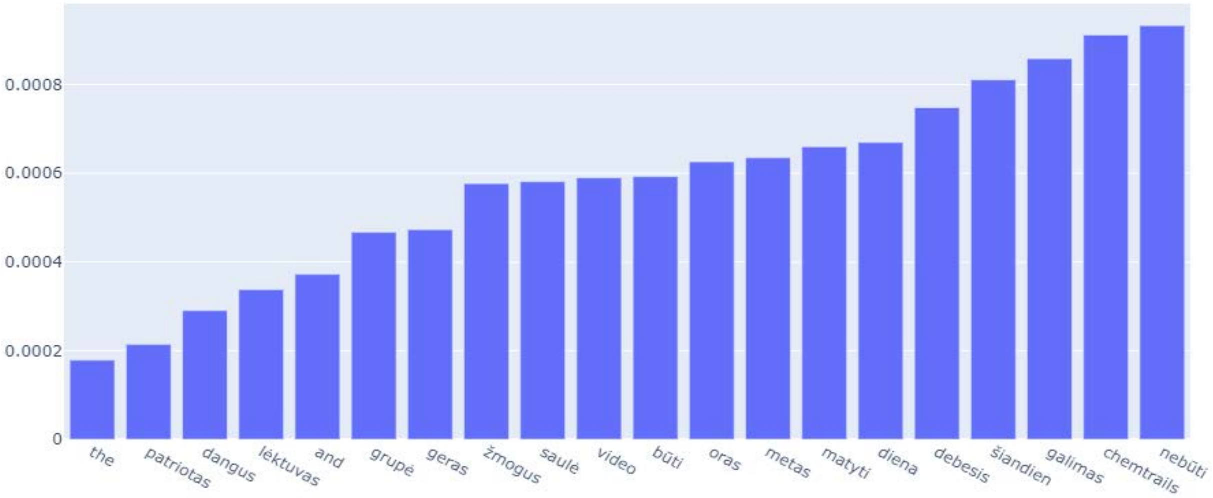

Figure 2: Telegram Channel *Geoengineering @Patriotai*: Keywords (2020 July and 2022 April)

The most relevant keywords for the *Geoengineering @Patriotai* channel contain 3 English words ‘the’ and ‘and’, and ‘chemtrails’, showing that in this channel some English was mixed with Lithuanian (see Figure 2). Other most relevant words are *patriotas* (patriot), *lėktuvas* (airplane), *grupė* (group), *geras* (good), *žmogus* (human), *saulė* (sun), *video*, *būti* (to be), *oras* (air or weather; it is not clear without the context), *metas* (time/period), *matyti* (to see), *diena* (day), *debesis* (cloud), *šiandien* (today), *galimas* (possible/potential), *nebūti* (not to be). One can observe that there is a number of keywords, related to weather conditions, which correlates quite well with the information in the channel’s description, emphasizing, that *Geoengineering @Patriotai* is for discussions about chemtrails, stratospheric aerosol injection, cloud seeding, etc.

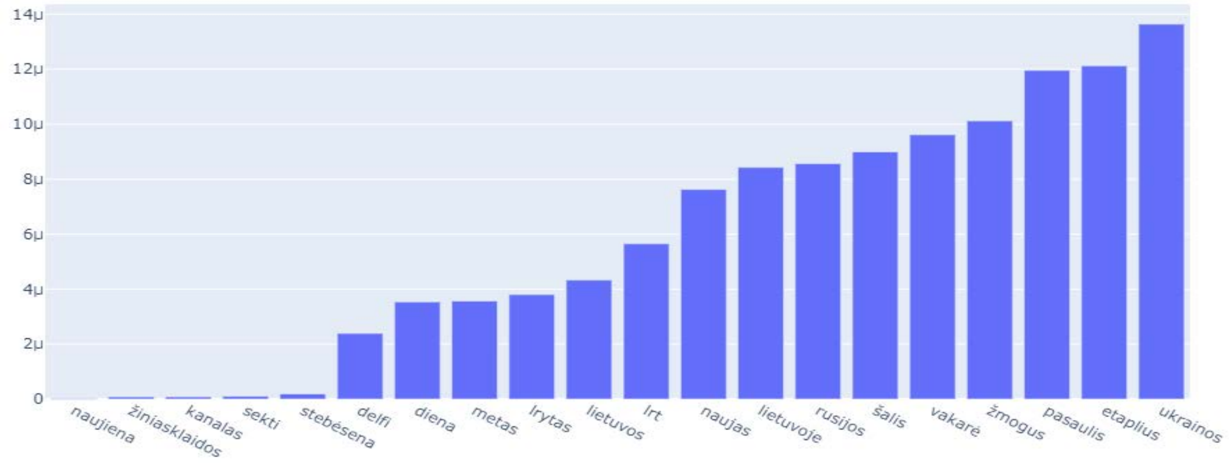

Figure 3: Telegram Channel *Naujienos Lietuvoje*: Keywords (2020 July and 2022 April)

Among the most relevant keywords for *Naujienos Lietuvoje* (News in Lithuania) there were no English words, so this channel was wholly in Lithuanian (see Figure 3). The most relevant keywords are *naujiena* (news), *žiniasklaidos* ((of) media), *kanalas* (channel), *sekti* (to follow), *stebėsena* (monitoring), *delfi* (Lithuanian news portal), *diena* (day), *metas* (time/period), *lrytas* (Lithuanian news portal), *lietuvos* ((of) Lithuania), *lrt* (Lithuanian National Radio and Television), *naujas* (new), *Lietuvoje* ((in) Lithuania), *Rusijos* ((of) Russia), *šalis* (country), *vakare* ((in the) evening), *žmogus* (human), *pasaulis* (world), *etaplius* (Lithuanian regional news platform), *ukrainos* ((of) Ukraine). So, the most relevant keywords for this channel fall mainly into 2 categories – media-related and Ukraine and Russia-related, due to the content covering the Russian war on Ukraine.

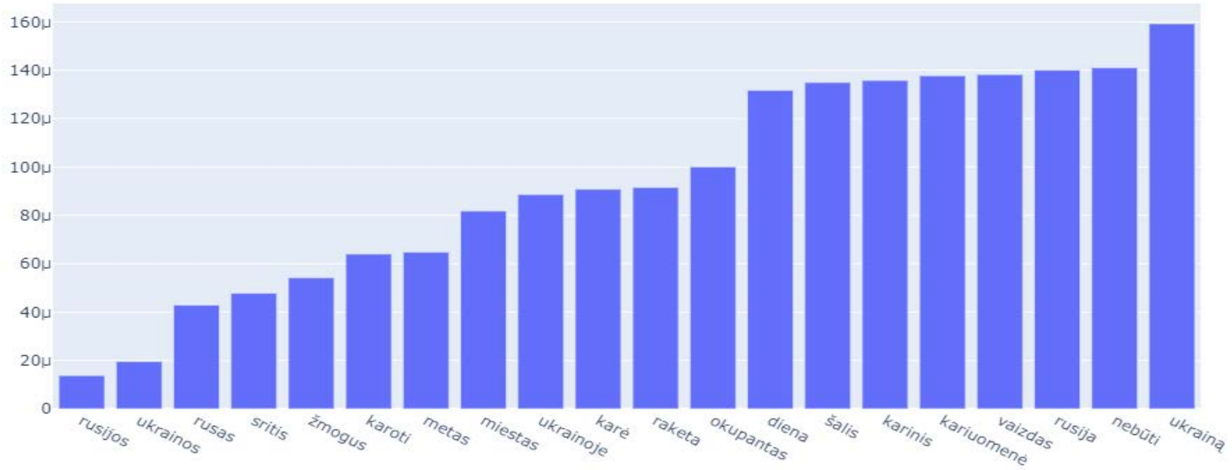

Figure 4: Telegram Channel *Savas kampas LT*: Keywords (2020 July and 2022 April)

Figure 4 presents 20 the most relevant keywords for *Savas kampas LT* (One's own corner LT) channel. All of them were related to the Russian war on Ukraine: rusijos ((of) Russia), ukrainos ((of) Ukraine), *sritis* (region), žmogus (human), *karoti* (incorrect lemma for *kariauti* – to wage war), *metas* (time/period), *miestas* (city/town), *ukrainoje* ((in) Ukraine), *karė* (incorrect lemma for *karas* – war), *raketa* (rocket), *okupantas* (occupier), *diena* (day), *šalis* (country), *karinis* (military), *kariuomenė* (armed forces), *vaizdas* (view), *rusija* (Russia), *ukrainą* (Ukraine). So, in comparison to other selected Telegram channels, here the topic of the Russian war on Ukraine was dominant.

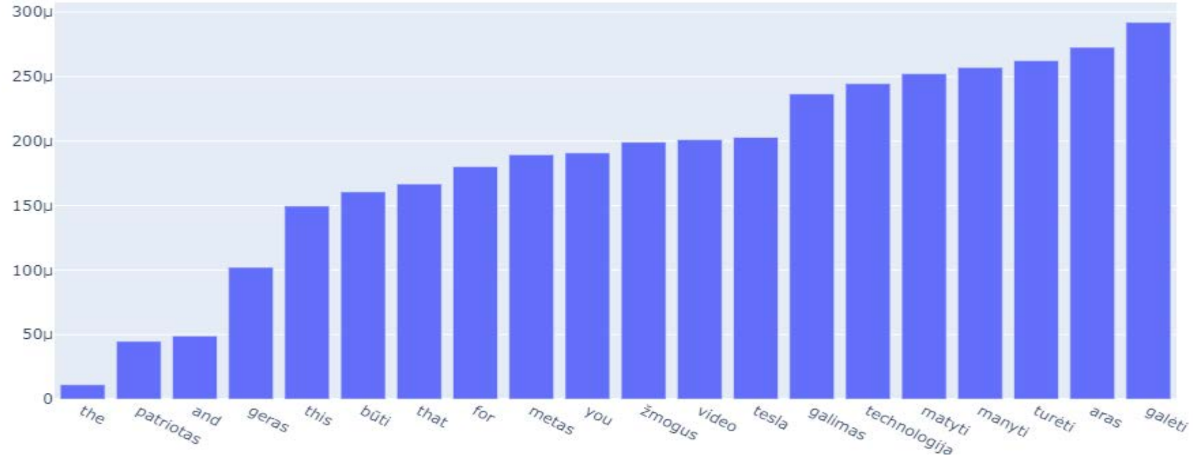

Figure 5: Telegram Channel *Slepiamos technologijos @Patriotai*: Keywords (2020 July and 2022 April)

For *Slepiamos technologijos @Patriotai* (Hidden technologies @Patriots) channel again some English words were among the most relevant keywords extracted by YAKE (see Figure 5). This, as it was already mentioned, indicated that this channel had some content in English, although the defined language was Lithuanian before running the tool. Besides English words, among the most relevant keywords can be found the following ones: *patriotas* (patriot), *geras* (good), *būti* (to be), *metas* (time/period), *žmogus* (human), *video*, *tesla* (either name of engineer and inventor Nikola Tesla or automotive company; it is not clear without the context), *galimas* (possible/potential), *technologija* (technology), *manyti* (to believe/think), *turėti* (to have), *aras* (Lithuanian Police Anti-Terrorist Operations Unit), *galėti* (can/to be able). This list of keywords seemed to correlate with the information in the channel's description, stating that *Slepiamos technologijos @Patriotai* (Hidden technologies @Patriots) is for discussions about technologies based on "vibrations, frequencies and energy".

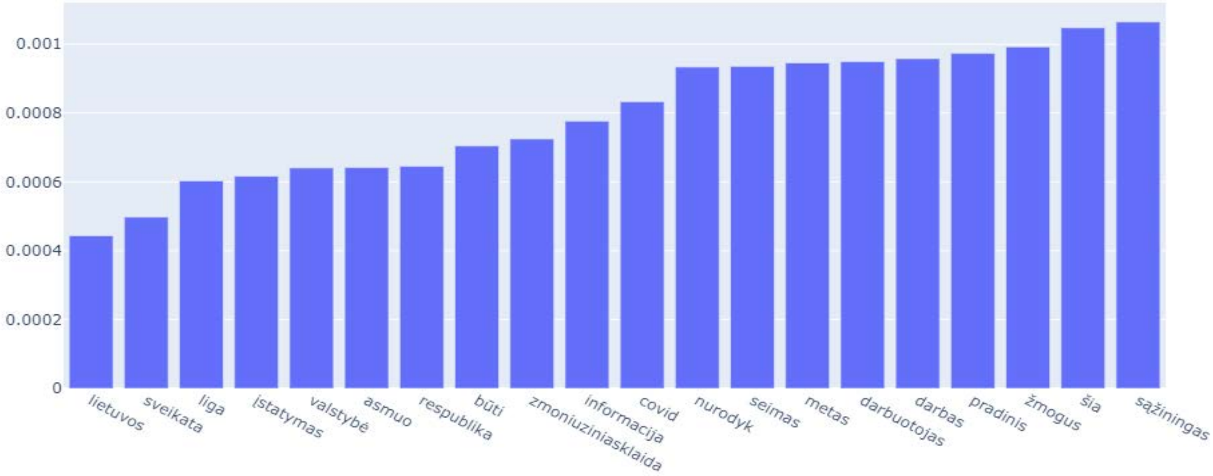

Figure 6: Telegram Channel *Žmonių žiniasklaida*: Keywords (2020 July and 2022 April)

Among the most relevant keywords for *Žmonių žiniasklaida* (People media) channel, only Lithuanian words were observed (see Figure 6). They were the following ones: *lietuvos* ((of) Lithuania), *sveikata* (health), *liga* (illness), įstatymas (law), valstybė (state (country)), asmuo (person), respublika (Republic), būti (to be) *zmoniuziniasklaida* (hashtag of *Žmonių žiniasklaida* – People media), *informacija* (information), *covid*, *nurodyk* (specify (smth)), *seimas* (Parliament), *metas* (time/period), *darbuotojas* (employee), *darbas* (work/job), *pradinis* (initial), *žmogus* (human), *šia* (this (in instrumental case)), *sąžiningas* (honest). These keywords represent quite well the description of the channel, stating that its purpose is discussions about "legal and other relevant information clearly and simply".

4.2 Topic analysis results

Topic modeling was made by using 50 768 messages from the Telegram platform. Because the data were collected with timestamps, we were able to determine which topics are the most popular each month. We also verified the word scores of the topics, and built Intertopic Distance Maps, Similarity Matrixes, Hierarchical Clustering, and Topic Probability Distributions for visual inspection.

The graph below (Figure 7) illustrates topic trends in the Telegram platform between 2020 July and 2022 April. It presents ten different topics: vaccines (Topic 0), cities and names (Topic 1), Vladimir Putin (Topic 2), media (Topic 3), inflation (Topic 4), flights (Topic 5), Russia (Topic 6), basketball (Topic 7), tennis (Topic 8) and random words about not understanding something (Topic 9).

In 2022 January, the most popular topic was about Covid-19 and vaccines. It was mentioned more than 300 times. On 2022 February, when the war in Ukraine started, the topic frequency about Putin increased and topics about vaccines and Covid-19 began to fall. In 2022 April we can see a sharp rise in people who started worrying about inflation (Topic 4). The frequency of the inflation topic rose from zero to ~500 in about two months. This was slightly higher than news and media topic, which frequency seeks ~400 (Topic 3). Also, coming closer to the beginning of the summer, people started to talk more about basketball and tennis (Topics 7 and 8).

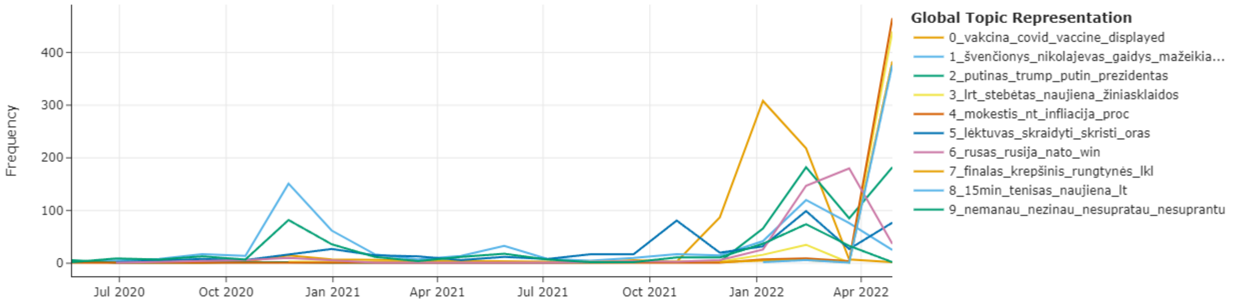

Figure 7: Telegram Topics over time between 2020 July and 2022 April

The bar charts below (Figure 8) present information about the most common words in the first 8 topics. These bar charts were generated from the c-TF-IDF scores for each topic representation. It enables us to quickly compare topic representations to one another. For instance, Topic 0 is about Covid-19 and the word *vakcina* (vaccine) is used the most frequently with a score of ~0.0332, but there is no significant difference in the scores of the other words because every word has a ~ 0,02 score. However, when we compare Topic 6, we can see a significant difference in the scores of the words *rusas* (Russian), which has a higher than ~0,0897 score, and *Rusija* (Russia), which has a score of ~0,0232.

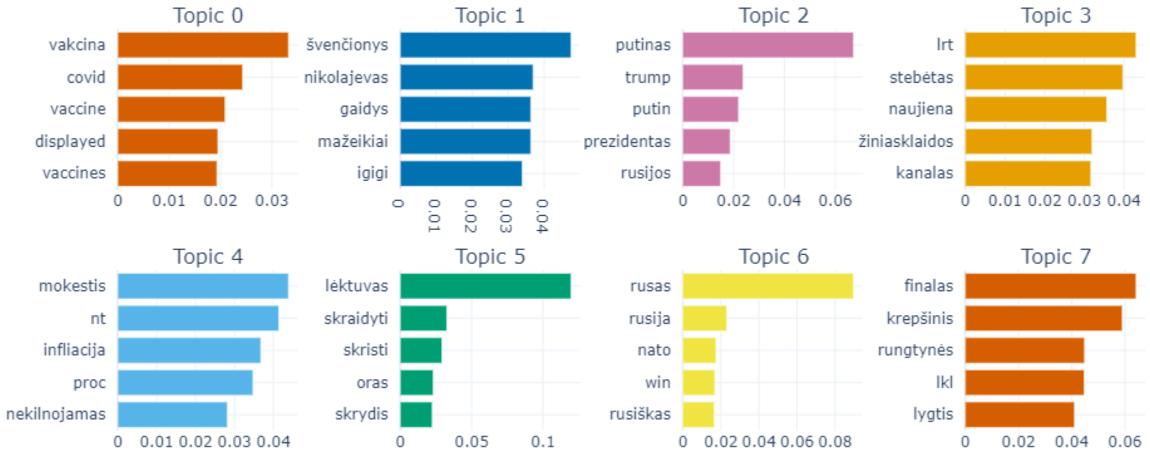

Figure 8: Topic word scores for the first eight topics

To see how closely related the topics are to each other, the Intertopic Distance Map (see Figure 9) was used. It shows 2D visualization, which can lack in accuracy for seeing the relationships between different topics when the dataset is large, and the number of topics is high. However, we used it to see groups of overlapping topics and, therefore, 100 topics were represented as circles. The centers of overlapping topics were determined by a computed topic distance. As Figure 9 shows, 18 macro-clusters were separately identified from the given topics. This visualization by its nature is interactive and by scrolling the slide at the bottom of the diagram from left to right separate topics in the clusters can be identified. This information is not available in static diagram, presented in Figure 9.

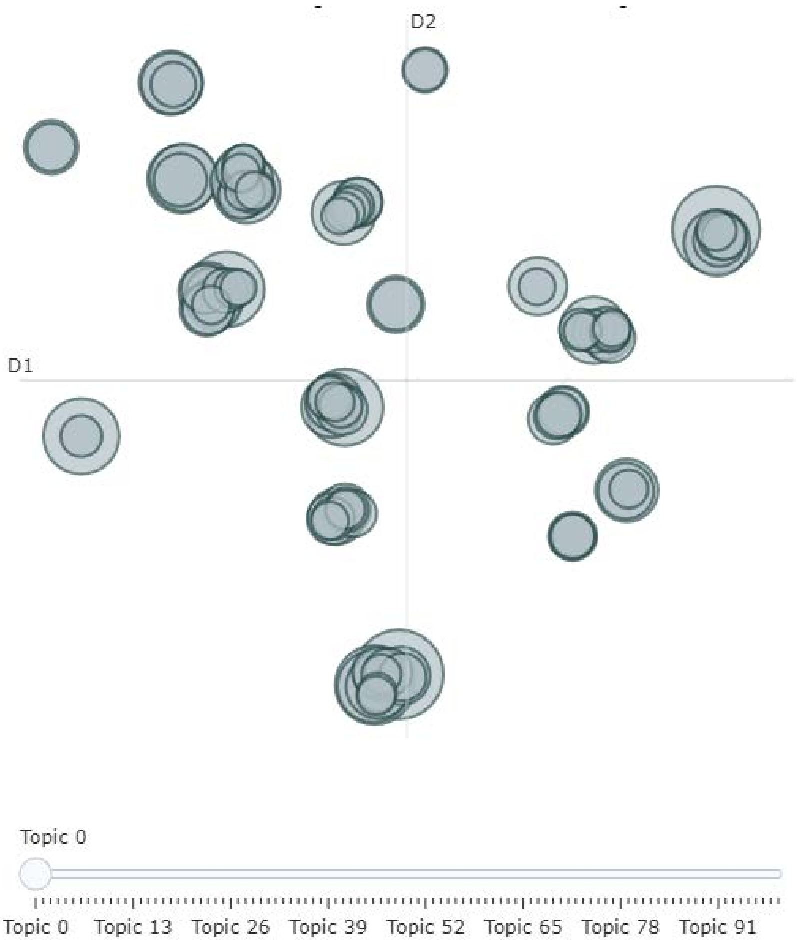

Figure 9: Intertopic Distance Map

Also, we used a similarity matrix and heatmap visualization technique to see how similar certain topics are in comparison to each other. The graph below (see Figure 10) represents the similarity matrix of the topics. The darker the blue color between clusters, the higher the similarity score between the topics, while the light green represents the lowest similarity between the topics.

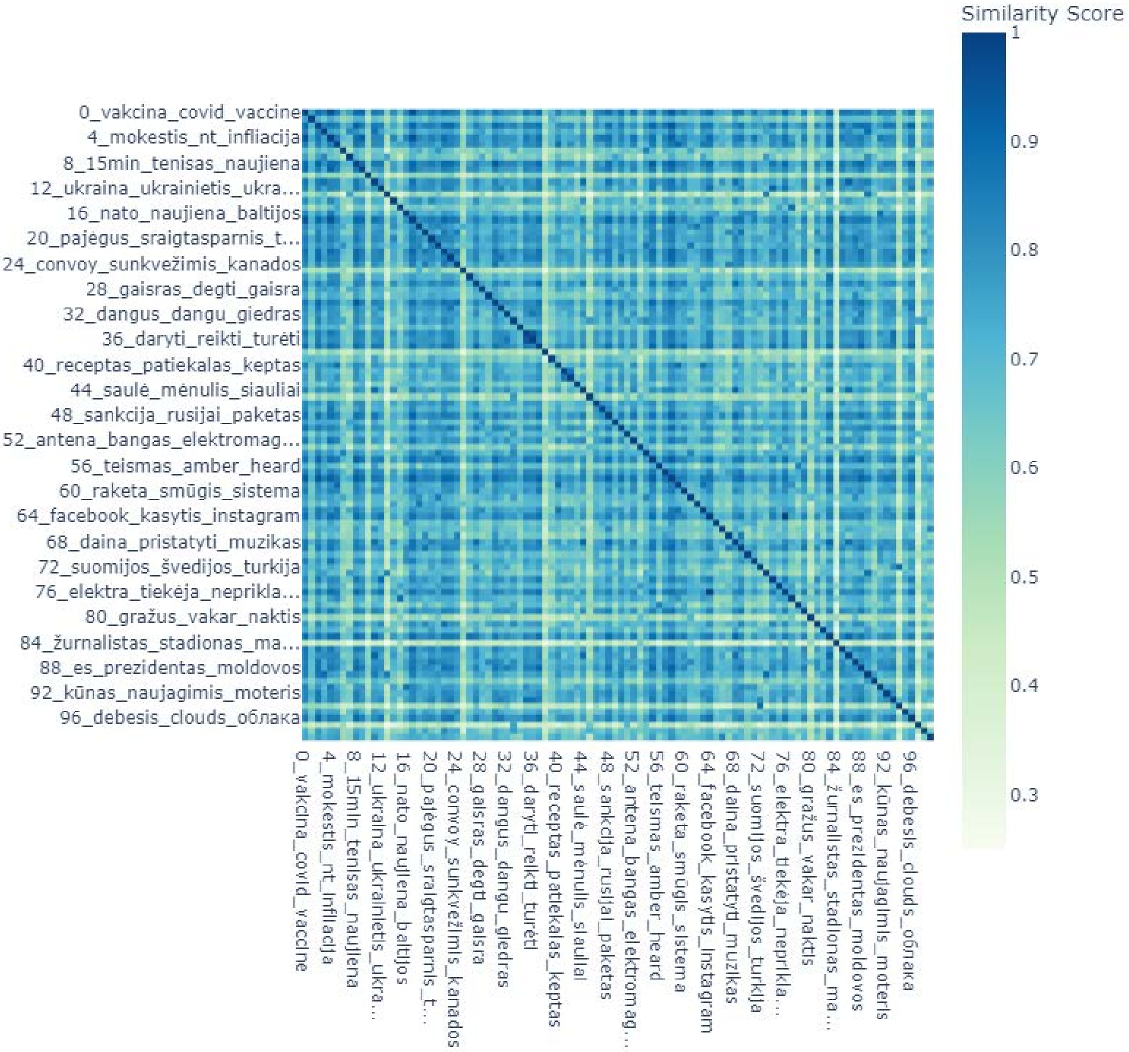

Figure 10: Similarity matrix of the topics

As we could see from the graphs above (Figure 9 and Figure 10) some topics are very close. To choose the right number of topics, they can therefore be organized hierarchically to reduce the number of relevant topics. Understanding how they relate to one another is made easier by the means of visualization. Figure 11 presents the hierarchical clustering of the 100 topics and systematically pairs them in clusters based on the similarity matrix. We noticed that cluster of Topics 23 and 66 created a unique cluster that we could label as the *music events* (marked with an arrow in Figure 11). It's important to mention that because of BERTopic limitations, not all topic clusters were formed well (for example, we can assign Topic 68 to the cluster of *music events* (marked with an arrow in Figure 11)).

For each message, we can also visualize the probability that the message belongs to each possible topic. To do so, we use the variable probabilities after running BERTopic to understand how confident the model is for that instance. Since there are too many topics to visualize, we visualize the probability distribution of the most probable topics (see Figure 12). We can see that topic 0 is more significant than the others. The relevant keywords from this topic are *vakcina* (vaccine), *covid*, *vaccine*, *displayed*, and hinting about Covid-19 vaccines.

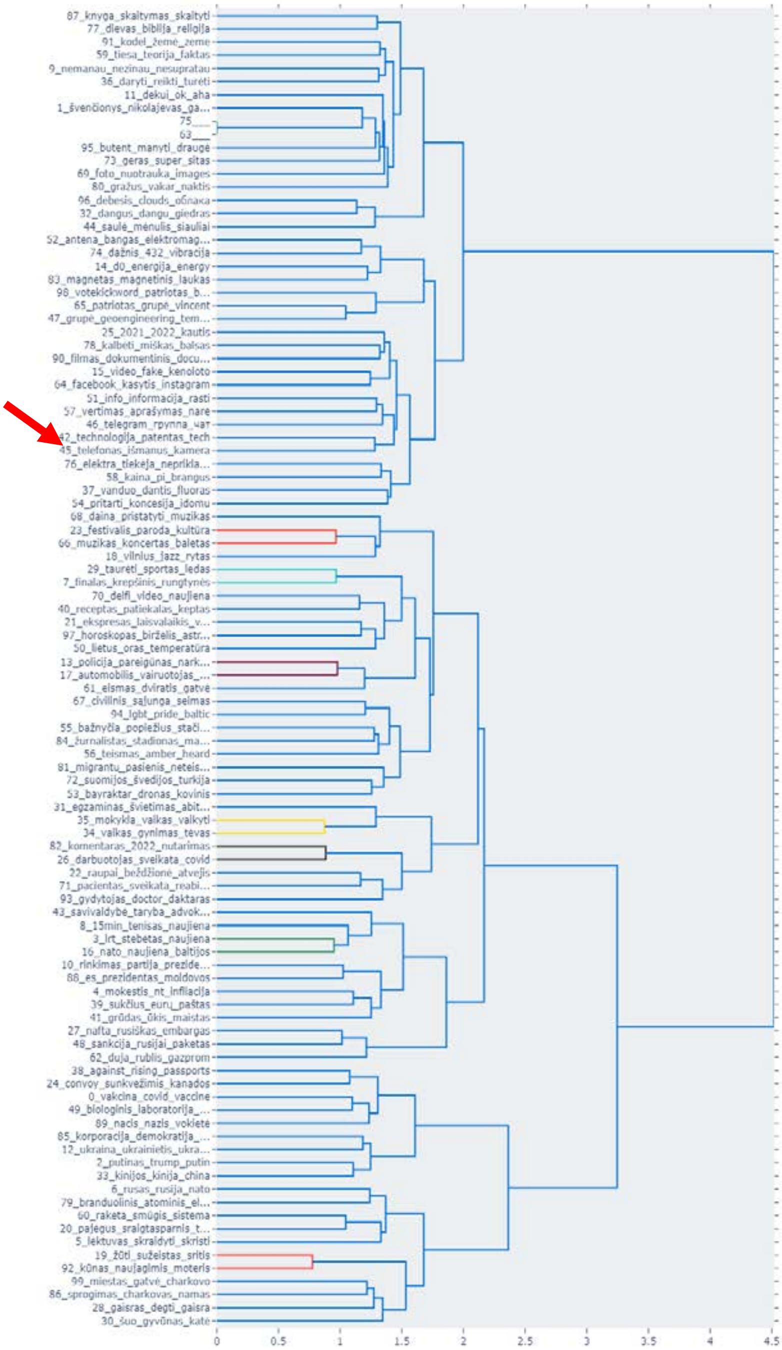

Figure 11: Hierarchical clustering of the topics

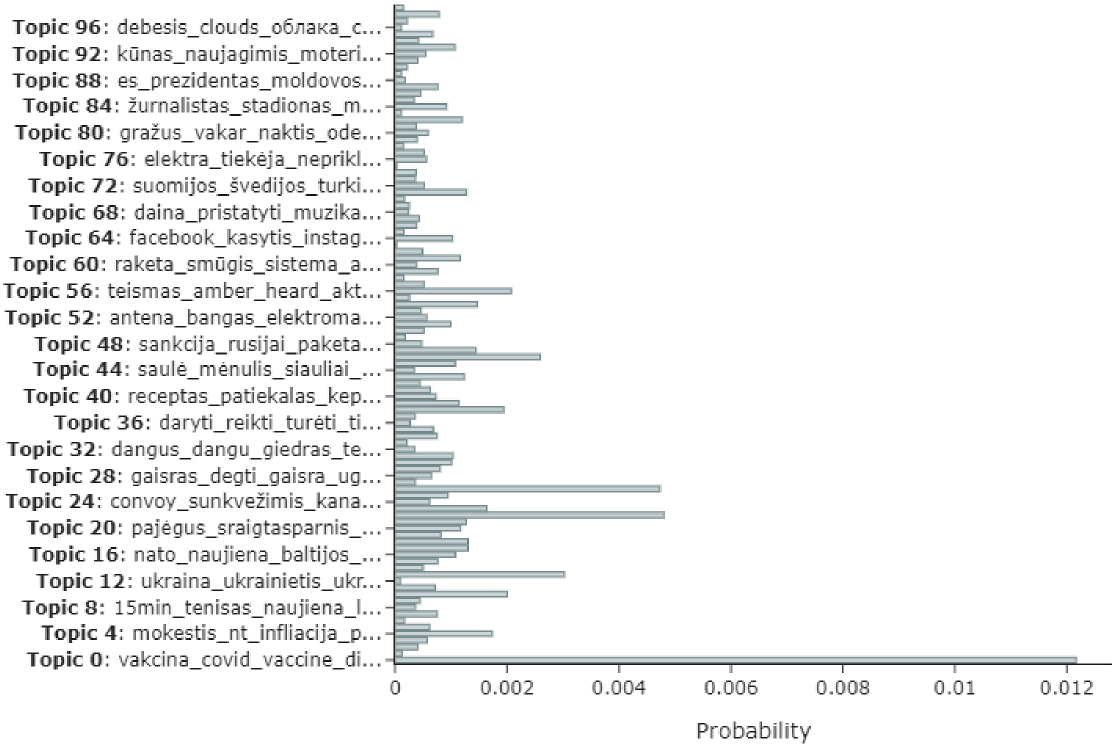

Figure 12: Topic Probability Distribution

Topic modeling allows quick inspection of the large collection of documents. The selected BERTopic extracts trending topics and facilitate their inspection, analysis, and interpretation of the results via visualizations. The next subsection (4.3) presents the results of sentiment analysis, which complements topic modeling with text “tone” exploration related to the expression of opinion.

4.3 Sentiment analysis results

Sentiment analysis for the Telegram dataset was out using Polyglot library for Python. It has polarity lexicons for 136 languages and can be used in various analysis and broad language coverage applications by marking texts as negative (-1), positive (+1), or neutral (0). The results of sentiment analysis showed that all studied channels had negative polarity (see Figure 13). It indicates that neutral or negative words are preferred over positive ones.

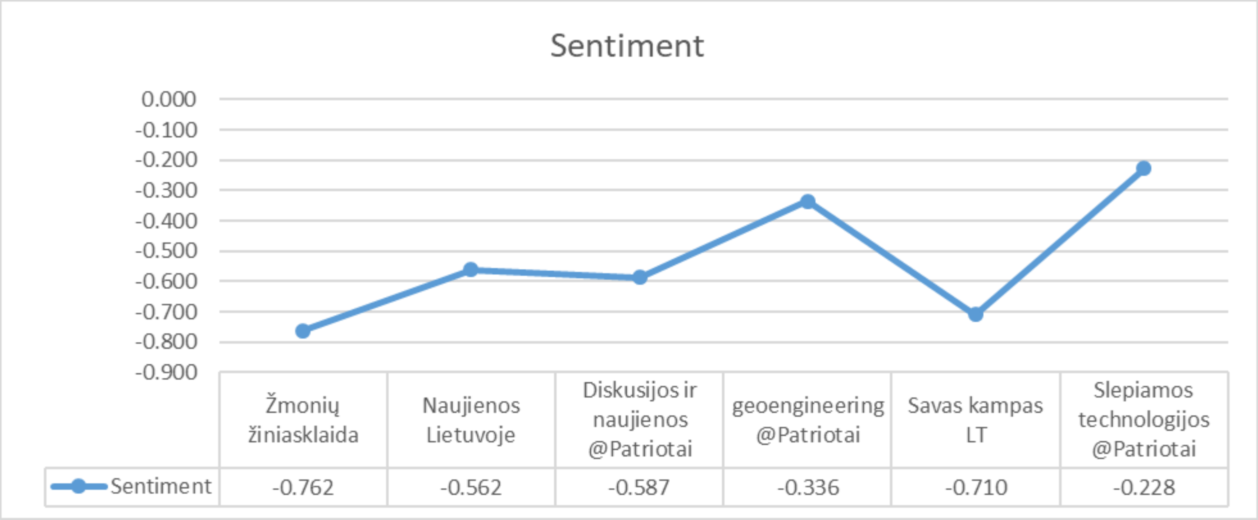

Figure 13: Sentiment Values for Selected Lithuanian Telegram Channels (2020 July and 2022 April)

The most negative channels were *Žmonių žiniasklaida* (People media) and *Savas kampas LT* (One‘s own corner LT) with sentiment scores of -0,762 and -0,710 accordingly. The least negative and closest to the neutral tone was *Slepiamos technologijos @Patriotai* (Hidden technologies @Patriots). The sharp difference between the polarity directly depends on the main topic of the channel and the number of opposing opinions.

Sentiment analysis studies the “tone” of the text in terms of being negative, positive, or neutral. It allows to evaluate the opinion of the audience towards an event, person, or idea. Sentiment analysis of selected Lithuanian Telegram channels revealed that they favored a negative tone, signaling a generally negative view towards happenings in the world and society, presented in the content of these channels in the time frame of the last 2 years.

5.0 CONCLUSIONS

In this study, we demonstrated the application of NLP, including AI/ML, and Visual Analytics to 6 Lithuanian Telegram channels. This social media platform has gained popularity due to its emphasis on consumer privacy while raising worries about radical groups using it. The chosen approach made it easier to identify trending topics in selected media sources (NLP) and to ascertain the prevalence of themes and their dynamics (Visual Analytics).

For our study, we used a dataset, made of ~52 thousand posts, which were scraped from the selected Telegram channels. By employing BERTopic for topic analysis as well as keyword and sentiment analysis, trending topics covering the last 2 years of the content in the selected channels were discovered and investigated. A combination of AI/ML and Visual Analytics makes it easier to monitor, comprehend, and mitigate IE dangers since it enables the visual study of both recognized subjects and topics across time via diverse types of graphs.

We plan to extend this research in terms of the diversity of media sources, while at the same time focusing on topic dynamics in time in greater detail as well as observing the relations between topics and media channels. Our plans also include a comparison of the presentation of the same events, people, and topics in media sources from different countries and different ideologies.

REFERENCES

[1] R. Albalawi, T. H. Yeap, and M. Benyoucef, “Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis,” *Frontiers in Artificial Intelligence*, vol. 3, Jul. 2020, doi: 10.3389/frai.2020.00042.

[2] M. Holmstrom, “The narrative and social media,” *Defence Strategic Communications*, vol. 1, no. 1, pp. 118–132, Dec. 2015, doi: 10.30966/2018.riga.1.7.

[3] M. La Morgia, A. Mei, A. M. Mongardini, and J. Wu, “Uncovering the Dark Side of Telegram: Fakes, Clones, Scams, and Conspiracy Movements,” *arXiv preprint arXiv:2111.13530*, 2021.

[4] P. Kaila, and A. V. Prasad, “Informational flow on Twitter–Corona virus outbreak–topic modelling approach,” *International Journal of Advanced Research in Engineering and Technology (IJARET)* 11, no. 3, 2020.

[5] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, “Detection of Depression-Related Posts in Reddit Social Media Forum,” *IEEE Access*, vol. 7, pp. 44883–44893, 2019, doi: 10.1109/access.2019.2909180.

[6] C. Van Hee *et al.*, “Automatic detection of cyberbullying in social media text,” *PLOS ONE*, vol. 13, no. 10, p. e0203794, 2018, doi: 10.1371/journal.pone.0203794.

[7] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, “FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media,” *Big Data*, vol. 8, no. 3, pp. 171–188, Jun. 2020, doi: 10.1089/big.2020.0062.

[8] S. Walther and A. McCoy, “US extremism on Telegram,” *Perspectives on Terrorism* vol. 15, no. 2, pp. 100–124, 2021.

[9] A. Urman and S. Katz, “What they do in the shadows: examining the far-right networks on Telegram,” *Information, Communication & Society*, pp. 1–20, Aug. 2020, doi: 10.1080/1369118x.2020.1803946.

[10] B. Plank, “What to do about non-standard (or non-canonical) language in NLP,” In *KONVENS 2016, Ruhr-University Bochum*, 2016.

[11] S. Grabus and J. Greenberg, “Computational Curation and the Application of Large-Scale Vocabularies,” In *2021 IEEE International Conference on Big Data (Big Data)*, IEEE, pp. 2220-2223, 2021.

[12] R. Campos *et al.*, “Yake! collection-independent automatic keyword extractor,” In *European Conference on Information Retrieval*, Springer, Cham, pp. 806-810, March 2018.

[13] R. Campos *et al.*, “A text feature based automatic keyword extraction method for single documents,” In *European conference on information retrieval*, Springer, Cham, pp. 684-691, March 2018.

[14] R. Campos *et al.*, “YAKE! Keyword extraction from single documents using multiple local features,” *Information Sciences*, vol. 509, pp. 257-289, 2020.

[15] N. Kolla, R. Gupta, and T. Fletcher, “Guide to Topic Modeling for Qualitative Research – Building the Model (Part 1),” May 13, 2019, https://epar.evans.uw.edu/blog/guide-topic-modeling-qualitative-research-building-model-part-1 (accessed Sep. 06, 2022).

[16] H. Jelodar *et al.*, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15169-15211, 2019.

[17] R. Vangara *et al.*, "Finding the Number of Latent Topics With Semantic Non-Negative Matrix Factorization," *IEEE Access*, vol. 9, pp. 117217–117231, 2021, doi: 10.1109/access.2021.3106879.

[18] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[19] J. Devlin *et al.*, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[20] Y. Meng *et al.*, "Topic discovery via latent space clustering of pretrained language model representations," In *Proceedings of the ACM Web Conference 2022*, pp. 3143-3152, 2022.

[21] C. Vladescu *et al.*, "What are the Latest Cybersecurity Trends? A Case Study Grounded in Language Models," In *2021 23rd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, IEEE, pp. 140-146, 2021.

[22] D. M. E. D. M. Hussein, "A survey on sentiment analysis challenges," *Journal of King Saud University-Engineering Sciences*, vol. 30, no. 4, pp. 330-338, 2018.

[23] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014, doi: 10.1016/j.asej.2014.04.011.

[24] P. Mehta and S. Pandya, "A review on sentiment analysis methodologies, practices and applications," *International Journal of Scientific and Technology Research*, vol. 9, no. 2, pp. 601-609, 2020.

[25] E. Rudkowsky, M. Haselmayer, M. Wastian, M. Jenny, Š. Emrich, and M. Sedlmair, "More than Bags of Words: Sentiment Analysis with Word Embeddings," *Communication Methods and Measures*, vol. 12, no. 2–3, pp. 140–157, Apr. 2018, doi: 10.1080/19312458.2018.1455817.

[26] M. Boukes, B. van de Velde, T. Araujo, and R. Vliegenthart, "What's the Tone? Easy Doesn't Do It: Analyzing Performance and Agreement Between Off-the-Shelf Sentiment Analysis Tools," *Communication Methods and Measures*, vol. 14, no. 2, pp. 83–104, Oct. 2019, doi: 10.1080/19312458.2019.1671966.