LiveJournal topic models and their improvement with contextualized representations for creating a model of hidden communities

Ivan D. Mamaev, Olga A. Mitrofanova

Abstract— **Social networks reflect contemporary tendencies in our society. These tendencies allow users to form communities that have both explicit and hidden links. The latter one is of current interest among scholars. Despite the effectiveness of modern algorithms, they do not take linguistic parameters of datasets into account. This gap can be filled by an algorithm that combines linguistic and quantitative data analysis. The purpose of the study is to detect hidden links among users' posts of the Russian segment of LiveJournal with the help of topic modeling procedures. The current size of the corpus is more than 95,490 posts (132 users). The procedure for constructing a model of hidden communities contains several stages. The first step is to process the corpus data using the Stanza library, which provides a single process of tokenization and lemmatization of social network posts and the removal of manually selected stopwords. The second step is creating contextualized topic models and their manual annotation. The final step is to build a semantic network of users using Easy Linavis and Gephi. The resultant model of hidden communities is represented as a group of vertices connected by edges. The results of the study provide new information about possible social groups in the Russian segment of social networks that can further be analyzed in terms of linguistics.**

Keywords— **Computational Linguistics, Hidden Communities, Topic Modelling, Semantics**

I. INTRODUCTION

The internal structure of social networks has attracted media communication specialists, sociologists and linguists for a long time, and the main research is aimed at describing opaque connections. A detailed description of the structure of communities helps to identify the most important parameters of communication and to see the whole picture of networks. As a results, scholars introduced such an important concept as hidden communities. Although F. Santo claims that it is rather difficult to formulate a general definition of a community due to the heterogeneity of systems, datasets and analyzed properties [26], however, these parameters can be considered special cases, which are described by each individual researcher. In a narrow sense, a hidden community is a clique, a subgraph whose vertices are adjacent to each other [26]. Moving on to understanding the concept in a broad sense, hidden communities are groups of social network users with common interests, these users have unstable links. Web hidden communities can be compared to real communities with a non-standard organizational structure: socially dangerous groups of drug addicts, secret organizations, etc. Unlike families, colleagues or best friends, the connections among members of real-life hidden communities are not obvious [9, 13]. Hidden communities contain independent layers within a graph. It means that understanding structures of hidden groups, as well as their organization, is crucial for a more detailed description of developing the society.

The current study is going to focus on the dataset of 2020-2022 Russian LiveJournal posts, which was developed in course of our research, and create its model of hidden communities with the help of topic modeling procedures. Although LiveJournal isn't popular among users as other social networks in Russia (VK, Instagram, etc.), it has some reasons to be analyzed. First of all, its inner structure is not as difficult as the structure of other sites like Facebook or Instagram so it can be web-scraped with contemporary libraries. Moreover, studies [3, 20] prove deep interest of researchers to LiveJournal linguistic data, thus, this social network is of current importance among researchers. It is also worth mentioning that the Russian segment of LiveJournal is characterized by a specific set of communication practices that distinguish it from other social networks. Its posts contain more textual information, which is necessary for conducting experiments in the field of topic modeling, while posts on other social networks tend to use a lot of audio and video content, as a consequence, textual information is likely to be far from being presented in full.

II. LITERATURE OVERVIEW

As it was mentioned in the previous section, the concept of hidden communities plays an important role in the network analysis. Unfortunately, depending on the purpose of each study, scholars use unique approaches and modern technologies, therefore, it is a difficult task to classify these

Manuscript received September 25, 2022.

I. D. Mamaev is with Baltic State Technical University "Voenmeh" named after D.F. Ustinov, Department of Theoretical and Applied Linguistics, Russia, 190005, Saint Petersburg, 1-ya Krasnoarmeyskaya street, 1 (e-mail: mamaev_id@voenmeh.ru) and Saint Petersburg State University, Faculty of Philology, Russia, 199034, Saint Petersburg, Universitetskaya emb. 11 (e-mail: i.mamaev@spbu.ru), corresponding author.

O. A. Mitrofanova is with Saint Petersburg State University, Faculty of Philology, Russia, 199034, Saint Petersburg, Universitetskaya emb. 11 (e-mail: o.mitrofanova@spbu.ru).

algorithms. However, according to [10], existing algorithms can be divided into three main groups: graph approaches with corresponding mathematical analysis, cluster approaches, and hybrid approaches.

The first classification group is a graph one. A graph in the broad sense of the term is usually understood as a set of interconnected nodes. The development of graph theory begins in 1736, when L. Euler proposed a solution to the problem of Königsberg Bridges [15]. In the 20th century, graph analysis has become one of the popular methods for formalizing data from various domains. For example, real-life social networks and Internet communities can be formally represented as a graph, but the rapid growth of such a graph becomes a problem of network analysis. Thus, there is an urgent need for a fundamental rethinking of graph data representations with the help of modern tools [16, 24], including linguistic resources and processors. In this regard, a large number of heterogeneous methods can be used for the analysis of big data: the shortest path method, the minimum spanning tree search method, etc.

Machine learning methods also gained much popularity as regards the task of detecting hidden communities. One of the main approaches is to use cluster analysis, i.e., ordering certain objects into relatively homogeneous groups on the basis of their similarity with further interpretation [1]. This approach is one of the fundamental ones in modern sociological experiments. For example, papers [2, 19] describe methods of detecting dangerous groups like people with antisocial or suicidal behaviour.

Since 2020 it has become pivotal to detect communities that are closely related to COVID-19 issues. L. Chaudhary created a corpus of texts on COVID-19 (since January 2020 till August 2020) using the official website of Johns Hopkins University. Such methods as principal component analysis and k-means were used to detect clusters that reflect the hidden structure of the identified communities of countries. Countries and regions that became a part of the same community can provide similar assistance to each other in taking preventive measures to avoid the worst-case scenarios for COVID-19. The hidden communities and data about them might be useful in the healthcare sector when shaping further political decisions [14].

It is also pivotal to emphasize that a clustering proper approach might have poor efficiency compared to modified graph approaches. In [27], the authors state that standard clustering algorithms (for example, mixed Gaussian models, k-means, etc.) are not applicable to work with overlapping communities, since they have tight connections in real life. The degree of stability of connections that were obtained automatically is much less, that is why the Community Detection with Generative Adversarial Nets (GAN) procedure was implemented and tested [27]. It is based on generative adversarial networks. These networks use two neural networks. The first one generates samples (probable communities), the second one selects true communities. Thus, the amount of noise in the final models is reduced, and researchers get clearer information about overlapping communities. The results of experiments carried out on five datasets (based on YouTube, Amazon, etc.) made it possible to state that CommunityGAN can be used for further experiments.

The third type of algorithms is a hybrid one. It combines both previously described methods and additional ones. For instance, in [6], the focus of the researchers' attention was a hidden community of cybercriminals. This marginal group carries out uncontrolled illegal activities, including cyberbullying, cyberfraud and the sale of drugs. The authors developed a criminal information mining platform based on the WordNet linguistic thesaurus to identify and extract information stored in chats that is important for forensic examination. After processing the suspects' chat, the cliques and topics in each conversation are identified [23]. Two nodes are connected if each user's keywords have a common hypernym in the WordNet thesaurus. During the experimental evaluation, it was found that the developed approach allows one to extract cliques and the semantics of the conversation among clique members. The accuracy of detecting groups and their topical component in the dataset has increased by 10-20% compared to other modern algorithms for detecting hidden communities.

Further studies were focused on combinations of graph approaches and topic models. In [9], the proposed solution is based on the idea of quasi-author-topic models which implies creating basic LDA topics of the Russian VK posts with further addition of the authorship parameter. To simplify the creation of a resultant model, it was decided to label topics with the help of external and internal sources. The term *quasi-author-topic* model is not introduced by the authors, however, in course of experiments, it was found that the distribution of topics over authors is set manually. Since the experiment was conducted in the first wave of the COVID-19 pandemic, the leading topic was the topic of health.

III. Experimental design

A. Collecting the LiveJournal corpus

The corpus of LiveJournal Russian posts includes texts that are downloaded from LiveJournal social network. At the initial stage, using the continuous sampling method, we selected more than 1500 LiveJournal user profiles. Then we developed a parser, which was based on such libraries as *beautifulsoup4* [3] and *requests* [25]. The input was a list of users, for each individual user at the beginning of the list, a check was made with the tail of the list. It was necessary to see whether the target user and the user from the tail are friends on the social network. If they were friends, then the tail user was removed from the list. As a result, the biggest losses of real friends were at the beginning of the process, there were no serious losses in the middle of the checking process (See Figure 1).

The step of filtering friends was the most pivotal one. If two users were friends on social networks, they could not be united by latent links and form a model of a hidden community. This situation conflicted with the term of hidden communities, the links between such users were considered to be obvious.

The next step included collecting posts with the help of *lj-crawler* [8] since the beginning of 2020 till August 2022. During this step some of the users were also let out of account as they published non-textual information. The resultant number of users turned out to be 132. On average, each of the users published about 723 posts within 2020, 2021, and 2022, and the average length of each text was 334 words.

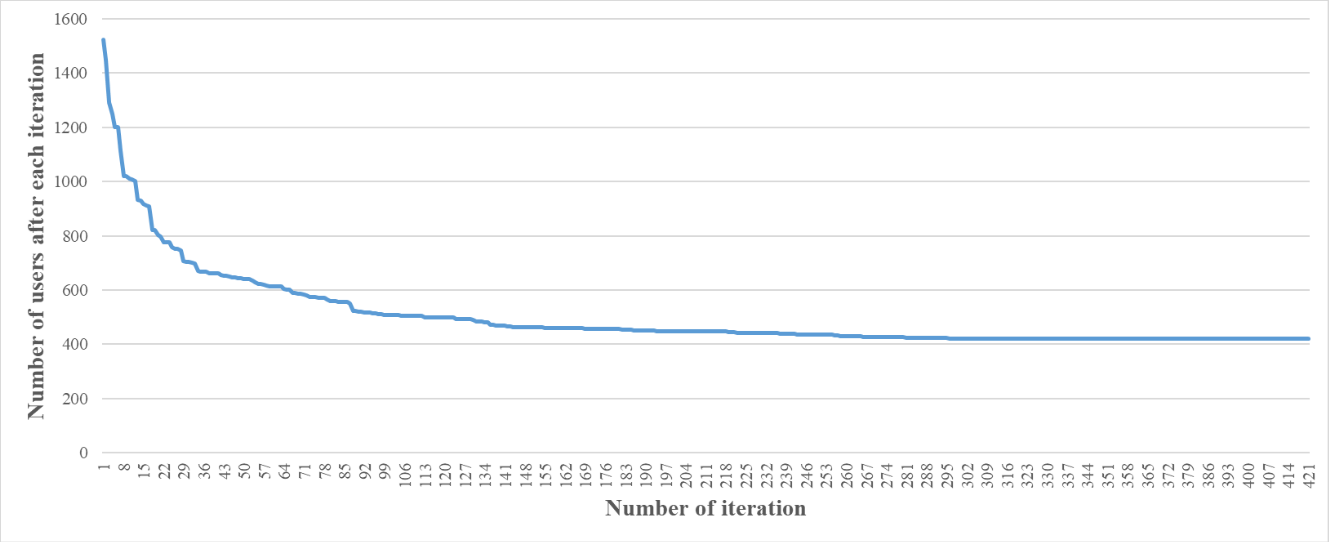

Figure 1. Filtering LiveJournal users

The final step implied corpus preprocessing. The *stanza* library [22] was chosen for this purpose as it allows creating a non-stop pipeline in a single code environment. We also used a stop-list during lemmatization to check each token, the stop-list includes prepositions, conjunctions, particles, interjections, symbols of various alphabets, obscene vocabulary, abbreviations, etc. The stop-list is based on a Frequency Dictionary of Contemporary Russian by O. N. Lyashevskaya and S. A. Sharoff [21], as well as words and expressions that were included after checking topic models of the first preliminary procedures: expressions of laughter like *ха (hah)*, graphical representations of emoticons, etc. The total number of stop words is more than 1400.

B. Creating contextualized topic models

Standard topic modeling procedures like Latent Dirichlet Allocation and Latent Semantic Indexing allow one to extract important topical words from both structured and unstructured texts [5]. Unfortunately, such models don’t take context into account, that’s why some semantic features of topical sets are unlikely to be mentioned. Nowadays, pre-trained language models like BERT fill in this gap, they are used in numerous NLP applications, topic modeling isn’t an exception. One of such implementations is BERTopic that is an approach that uses transformers and c-TF-IDF to create dense clusters for interpretable topics, it allows keeping important words in the topic descriptions [17]. The algorithm consists of three stages: creating document embeddings, predicting semantic clusters and printing topic representation from clusters. The c-TF-IDF compares the importance of lexical units to a specific cluster and reveals the most significant lexical units in a topic. It is calculated according to (1):

$$c - TF - IDF = \frac{f_i}{wd_i} \times log \frac{m}{\sum_j^n f_j}.$$

(1)

The frequency for each word f is extracted from each particular cluster i and then divided by the total number of lexical units wd of a cluster i. It is a way of normalizing the frequency of words in each cluster. Then the number of clusters m is divided by the total frequency of the word f across all the clusters. After generating the c-TF-IDF representations, a set of lemmata, that describe a collection of texts, is obtained. Of course, it does not mean that these words describe a coherent topic. To improve topic coherence, BERTopic uses Maximal Marginal Relevance to find the most coherent words with little overlap. This action results in removing words that do not contribute to a particular topic. The architecture of BERTopic is presented in Figure 2.

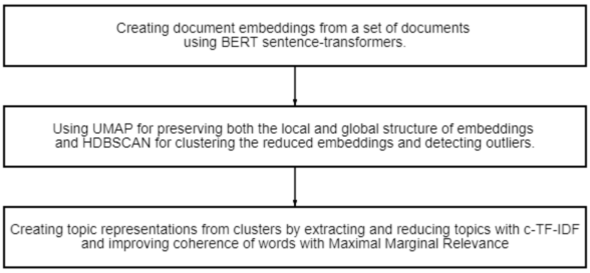
Creating document embeddings from a set of documents using BERT sentence-transformers.
Using UMAP for preserving both the local and global structure of embeddings and HDBSCAN for clustering the reduced embeddings and detecting outliers.
Creating topic representations from clusters by extracting and reducing topics with c-TF-IDF and improving coherence of words with Maximal Marginal Relevance

Figure 2. The architecture of BERTopic

Before obtaining topic models, it is necessary to develop a system of linguistic and technical filters to produce correct and interpretable results. The first stage is model tuning. The posts are distributed among all the authors so that it would be possible to assign particular topics to each author. As we manually include the authorship as a pivotal parameter for the model, the basic BERTopic models are transformed into contextualized quasi-author-topic models. For the contextualized BERTopic model, we choose the topic size that is equal to 10 lemmata. As this number of lemmata is usually set as a standard one in most topic modeling techniques, we decided to use the following filter: if the number of lemmata is less than 10, a topic will not be included in the resultant model and will not be seen in the output. The next stage is to choose the language. As the Russian model is absent in the current library, we choose the multilingual BERT model as it is also trained on some Russian data. Finally, the different posting activity of LiveJournal users resulted in great variation in the number of texts published. To smooth out such heterogeneity and output approximately the same number of topic models for each user, it was decided to take into account additional parameters. The *min_topic_size* parameter allows filtering topics that cover the smallest part of a user's subcorpus. It was also found the minimum value of the parameter should be equal to 20 for the corpus. However, when setting this parameter, there were still situations when the user had more than 500 topics. Papers [11, 18] indicate that the optimal number of relevant topic models varies between 10 and 20, so in this experiment, the first ten sets were considered significant topics for the user. From the point of view of linguistics, it is also important to note that the first topical set can be a hypernym in relation to the co-hyponym topics that appear in lower positions in the final output. In this regard, hypernymic topics are more likely to describe the interests of the user, that is why they are important for creating a resultant model of hidden communities.

Some of the results are presented in Table 1 that includes full lists of topics for each user. As we use additional filters, some preliminary topics were not printed in the output. This is the reason of omitting some topical indices in Table 1.

TABLE 1. BERTOPIC TOPICS

User	Topic index	BERT lemmata
tcaagan-sar	tcaagan-sar_0	*человек, система, государство, мера, экономический, бюрократия, власть, новый, цк, контроль (person, system, state, measure, economic, bureaucracy, power, new, central committee, control)*
	tcaagan-sar_1	*жизнь, рождаться, умереть, жить, пойти, поддерживать, понимать, кпсс, делать, человек (life, be born, die, live, go, support, understand, communist party of the soviet union, do, man)*
	tcaagan-sar_2	*церковь, церковный, государство, священник, государственный, общество, бюрократия, положение, оборона, католик (church, ecclesiastical, state, priest, state, society, bureaucracy, position, defense, catholic)*
	tcaagan-sar_4	*верующий, религия, религиозный, дело, совет, общественный, объединение, социальный, организация, отношение (believer, religion, religious, business, advice, public, association, social, organization, relation)*
	tcaagan-sar_5	*советский, православный, россия, мусульманин, горбачев, религия, церковь, посол, русский, россиянин (soviet, orthodox, russia, muslim, gorbachev, religion, church, ambassador, russian, russian resident)*
bogun-333	bogun-333_0	*перезагрузка, великий, глобальный, являться, свобода, новый, процветание, мировой, план, байден (reboot, great, global, be, freedom, new, prosperity, world, plan, biden)*
bogun-333	bogun-333_1	*сердечный, помощь, заболевание, проблема, штат, отделение, больница, неотложный, умереть, пациент (cardiac, help, disease, problem, staff, department, hospital, emergency, die, patient)*
	bogun-333_2	*смотреть, объяснить, ответ, человек, акцент, позволить, жилет, короткий, вопрос, ключевой (watch, explain, answer, person, accent, allow, vest, short, question, key)*
	bogun-333_3	*коронавирус, пандемия, мир, называть, вакцина, вакцинация, новый, обратный, представлять, грипп (coronavirus, pandemic, world, name, vaccine, vaccination, new, reverse, represent, flu)*

C. *Creating a model of hidden communities*

The resulting sets had a large lexical diversity, and therefore, the unification of topical lemmata became pivotal. Topics were marked manually. Some labeling features are worth mentioning. For example, despite describing opposed situations, antonymous topics are united by a common semantic component, and they can be assigned a single label. For such opposite topics as *умереть, сократиться, количество, число, уменьшиться, процент, смерть, умирать, страна,* *больной* (*die, shrink, quantity, number, decrease, percentage, death, perish, country, sick*) and *здоровье, долголетие, ребенок, тайна, пятница, суббота, человек, мир, здравствовать, привет* (*health, longevity, child, secret, Friday, Saturday, man, world, long live, hello*) we selected a single label of health. Also, in topical sets, all the lexical units could have broad semantics, which indicated that it was impossible to create a specific label. Such sets were about everyday problems, so we decided to choose a single label of everyday life: *уберечь, держаться, помочь, дело, хватить, ссылка, часть, делать, процесс, верх (save, keep, help, business, enough, link, part, do, process, top).*

The labels allowed uploading the obtained data to the Easy Linavis [4] application. Initially, the application was intended to visualize the relationship between the characters of literary works. Later it became widespread for the visualization of linguistic data. To create a model of hidden communities based on the LiveJournal corpus, we followed the following syntax: the # symbol was followed by the name of the topic, then usernames, who were interested in the topic, were written down on each individual line. Following the basic rules of such syntax, we built an elementary semantic network, which was modified in the Gephi application. The resultant model is in Figure 3.

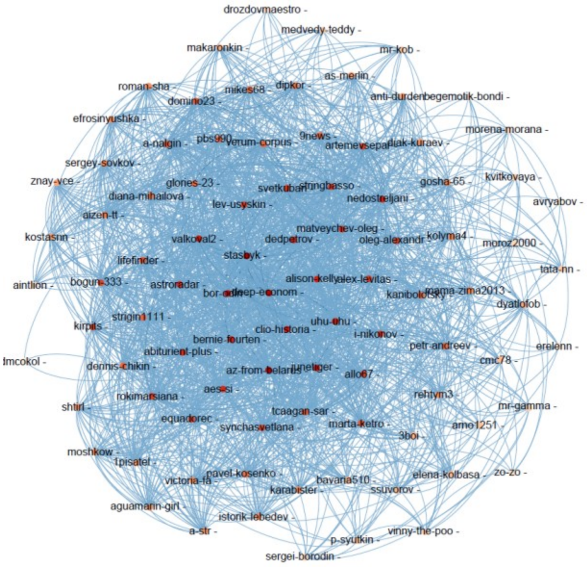

Figure 3. The LiveJournal hidden communities

IV. Discussion

The initial analysis of the graph (Figure 3) show that the center of the model has a dense structure. This feature is explained by the linguistic fact that users, who have a great number of common topics and are members of several hidden communities, are concentrated in the center of the graph. The graph periphery is sparse, and it means those users have fewer common topics with other users.

As for the mathematical properties of the graph, it is undirected and consists of 92 nodes and 1910 edges. Each edge has its own weight in the range from 1 to 5, which directly corresponds to the number of common interests between two users. Table 2 provides a brief description of the edges in terms of their weights.

Table 2. Weights of edges

Weight	Number of edges
1	1269
2	476
3	137
4	27
5	1

It is also important to mention the various options for assigning tags to topics, which might improve the quality of labelling resultant sets in the following experiments. First, for some implementations of topic models, the labelling procedure is not needed. For example, when one works with a real implementation of author-topic models, which is implemented in the *gensim* library [7], at the first stage a single set of topical sets is created, which are subsequently distributed among all the authors in the corpus. Then a certain percentage value is set for each of the topic. The higher the value is, the more important this topic is for the user. Topical sets themselves are already labels, and one can build a graph on their basis. Secondly, the procedure for labelling topics can be automated. For these purposes, one can use both internal sources (a vector model of the analyzed dataset) and external sources (vector models of other datasets or data obtained from search engines).

V. Conclusion

In this study, we have run several experiments to create a model of LiveJournal hidden communities. In recent articles, it is noted that hidden communities can be considered as separate cores, cliques, or clusters. We propose a hybrid algorithm that includes graph methods and contextualized topic representations. From a practical point of view, this algorithm can be further implemented in an application for finding potentially dangerous social groups. Such a method can also be used for tracking trends on Russian social networks. Further research will be aimed at linguistic description of the obtained hidden communities in order to identify common parameters for the formation of texts on these topics.

References

[1] A. V. Kutyrkin, A. V. Syomin, Klasternyj analiz: Metodicheskie ukazanija, Pereizdanie, M.: MIIT, 2009, 22 p.

[2] A. Wong, C. Lai, A. K. Shum, and P. S. Yip, “From the hidden to the obvious: classification of primary and secondary school student suicides using cluster analysis”, in *BMC public health*, vol. 22(1), 2022, pp. 1-7.

[3] Beautiful Soup Documentation [Online]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[4] C. Milling, E. S. Shlosman, and D. A. Skorinkin, “Easy linavis (simple network visualisation for literary texts)”, in *Informacionnye tekhnologii v gumanitarnykh naukakh*, 2017, pp. 104-107.

[5] F. Bianchi, S. Terragni, D. Hovy, D. Nozza, and E. Fersini, “Cross-lingual contextualized topic models with zero-shot learning”, in: *arXiv preprint arXiv:2004.07737*, 2020.

[6] F. Iqbal, B. C. Fung, M. Debbabi, R. Batool, and A. Marrington, “Wordnet-based criminal networks mining for cybercrime investigation”, in *IEEE Access*, vol. 7, 2019, pp. 22740-22755.

[7] Gensim. *Author-topic models* [Online]. Available: https://radimrehurek.com/gensim/models/atmodel.html

[8] Github, *lj-crawler 0.9* [Online]. Available: https://github.com/roman-lugovkin/lj-crawler

[9] I. Mamaev, and O. Mitrofanova, “Automatic Detection of Hidden Communities in the Texts of Russian Social Network Corpus”, in *Conference on Artificial Intelligence and Natural Language*, Springer, Cham, 2020, pp. 17-33.

[10] I. Mamaev, and O. Mitrofanova, “Hidden Communities in the Russian Social Network Corpus: a Comparative Study of Detection Methods”, in *CMCL*, 2020, pp. 69-78.

[11] J. Gan, and Y. Qi, “Selection of the Optimal Number of Topics for LDA Topic Model—Taking Patent Policy Analysis as an Example”, in *Entropy*, vol. 23(10), 2021, pp. 1-45.

[12] K. He, S. Soundarajan, X. Cao, J. Hopcroft, and M. Huang, Revealing multiple layers of hidden community structure in networks, in *arXiv preprint arXiv:1501.05700*, 2015.

[13] K. He, Y. Li, S. Soundarajan, and J. E. Hopcroft, “Hidden community detection in social networks”, in *Information Sciences*, vol. 425, 2018, pp. 92-106.

[14] L. Chaudhary, and B. Singh, “Community detection using unsupervised machine learning techniques on COVID-19 dataset”, in *Social Network Analysis and Mining*, vol. 11(1), 2021, pp. 1-9.

[15] L. Euler. “Solutio problematis ad geometriam situs pertinentis”, in *Commentarii academiae scientiarum Petropolitanae*, 1741, pp. 128-140.

[16] M. E. Newman, “The structure and function of complex networks”, in *SIAM review*, vol. 45(2), 2003, pp. 167-256.

[17] M. Grootendorst, BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics, 2020, doi: 10.5281/zenodo.4381785.

[18] M. Hasan, A. Rahman, M. Karim, M. Khan, S. Islam, and M. Islam, “Normalized approach to find optimal number of topics in Latent Dirichlet Allocation (LDA)”, in *Proceedings of International Conference on Trends in Computational and Cognitive Engineering*, Springer, Singapore, 2021, pp. 341-354.

[19] N. E. Lyapin, and M. E. Abramov, “Instruments and technologies for automated assessment of expression of personal characteristics of social network users” [Instrumenty i tekhnologii dlya avtomatizatsii otsenki vyrazhennosti lichnostnykh osobennostej pol'zovatelej social'nykh setej], in *Regional Informatics (RI-2020). XVII St. Petersburg International Conference "Regional Informatics (RI-2020)" [Regional'naya informatika (RI-2020). XVII Sankt-Peterburgskaya mezhdunarodnaya konferentsiya «Regional'naya informatika (RI-2020)»]*, vol. 2, 2020, pp. 253-255.

[20] O. Koltsova, S. Alexeeva, S. Pashakhin, and S. Koltsov, “PolSentiLex: Sentiment Detection in Socio-Political Discussions on Russian Social Media”, in *Conference on Artificial Intelligence and Natural Language*, Springer, Cham, 2020, pp. 1-16.

[21] O. N. Lyashevskaya, and S. A. Sharoff, *Frequency dictionary of modern Russian based on the Russian National Corpus [Chastotnyj slovar’ sovremennogo russkogo jazyka]*, Azbukovnik, Moscow, 2009.

[22] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. “Manning, Stanza: A Python natural language processing toolkit for many human languages”, in *arXiv preprint arXiv:2003.07082*, 2020.

[23] R. D. Alba, “A graph - theoretic definition of a sociometric clique”, in *Journal of Mathematical Sociology*, vol. 3(1), 1973, pp. 113-126.

[24] R. Pastor-Satorras, and A. Vespignani, *Evolution and structure of the Internet: A statistical physics approach*. Cambridge University Press, 2004.

[25] Requests: HTTP for Humans [Online]. Available: https://requests.readthedocs.io/en/latest/

[26] S. Fortunato, Community detection in graphs, in *Physics reports*, vol. 486(3-5), 2010, pp. 75-174.

[27] Y. Jia, Q. Zhang, W. Zhang, and X. Wang, X. “CommunityGAN: Community Detection with Generative Adversarial Nets”, in *The World Wide Web Conference*, 2019, pp. 784-794.