# Topic Modeling of Russian-Language Texts Using the Parts-of-Speech Composition of Topics (on the Example of Volunteer Movement Semantics in Social Media)

Anna Maltseva, Natalia Shilkina, Evgeniy Evseev,
Mikhail Matveev
Saint Petersburg State University
Saint Petersburg, Russia
{a.maltseva, n.shilkina, e.evseev, m.matveev}@spbu.ru

Olesia Makhnytkina
ITMO University
Saint Petersburg, Russia
makhnytkina@itmo.ru

*Abstract*—**The article presents a new approach to thematic modeling of texts - this is thematic modeling based on part-of-speech topics. We do not consider parts of the speech as a gnoseological concept that reflects the way in which language is formally classified. We believe that parts of speech are within the language competence of the person and are used in the process of communication, performing a certain function in the communication process. The essence of thematic modeling is seen as the creation of semantic models of the text corpus. The goal is to study the speech representation of modern movements and communities. The hypothesis is that the forums of a social movement reflect its characteristics, the nature, and activities of this movement. Three groups of the Russian social media VKontakte were chosen as an empirical object: "All for the Victory!", "Center of (City) Volunteers of St. Petersburg," "Volunteers of St. Petersburg." Topic modeling was carried out using the latent Dirichlet allocation (LDA) method, implemented in the Gensim package along with the Mallet implementation. Model quality validation was carried out using the coherence coefficient. The described approach to the analysis of web texts of volunteer semantics based on the part-of-speech composition of topics made it possible to identify signs that characterize group identity, emotionality, and joint activities of Russian volunteers.**

## INTRODUCTION

This work is devoted to the actively developing direction of probabilistic topic modeling, that is quite popular in recent years. Topic modeling is chosen as an analysis method, a way of constructing a model of a message collection that allows you to find hidden topics in the discussion. [1]

A society that is becoming more transparent opens to the public a large body of various information about its subjects, processes, and events. The interdisciplinary challenge of modern science is to extract structured information from this disparate information and interpret it. And topic modeling acts as a method of generalization, systemization of large text data, a means of identifying non-obvious meanings.

The changes taking place in all spheres of modern society are cardinal and rapid. New movements are constantly emerging, affirming new semantics of life. This actualizes research in a manner “here and now” that records the facts of social reality and social subjects at the real top of event discourse. The topic model acts as a means of generalization, systemization, and semantic search for large text collections. Such models allow the subject of texts to be defined and serve to solve the problems of classification and clustering documents based on semantic proximity of content. Thematic models solve a variety of current natural language processing challenges; for example, identifying identities, behavioral intentions, dominant emotions, etc. Typically, the number of topics found in messages is less than the number of different words in the entire set. Therefore, hidden variables - topics - allow you to present a collection of messages as a vector in the space of latent topics instead of representation in the space of words. As a result of this analysis, the message reduces the number of components, allowing for faster and more effective conclusions about its meaning content.

The essence of topic modeling is to create semantic models of the body of texts based on varieties of fuzzy clustering of vocabulary [2], [3]. The aim is to respond to the request of science and society for research on speech representation of modern movements and communities, both permanent and emerging situationally. As a rule, the purpose of topic modeling is associated with an acute social problem and its latent content. Modern publications cover a fairly wide range of fundamental and applied problems: various crisis manifestations in society, from racial divisions to overtly destructive facts [4], [5], [6]; personal and emotional problems of different groups of the population: students, migrants, survivors of tragedies and/or suffering from various diseases, etc. [7], [8], [9]; assessments of the movement of masses of people within settlement boundaries [10], [11]; identifying fake news and the popularity of various media [11], [12], [13], [14]; improving the quality of medical care through the use of social media [15], [16], [17]; study of assessments of various relevant news and services [18], [19], [20] and so on.

Social media materials: texts, symbols and reposts are very informative for learning different movements and phenomena. However, the challenge for scientists is the scarcity and imperfection of analytical methods.

Thematic modeling, having emerged at the turn of the 20th and 21st centuries, has formed several new research canons [3],

[21], [22]. However, many methodical problems of automatic text processing are still relevant today.

The unresolved task so far is to learn to ignore the insignificant differences of verbal forms and to highlight the differences that define meanings. It is necessary to analyze the various passages and understand the semantics of each in the context of the problem under study. The same concept can be expressed by any number of different terms (synonyms), whereas one term often has different meanings in different contexts (polysemy). Thus, ways must be found to distinguish between the representation of a particular concept and the specific meaning of ambiguous terms.

## II. Research Details

### *A. Related works*

The article presents a new approach to the thematic modeling of semantics on the example of complex-formed verbality - a topic modeling based on the part-of-speech composition of topics.

We believe that the idea that parts of speech are not only a gnoseological concept that reflects the way of formal classification of language is applicable. We agree with M.G. Kulikova that parts of speech are included in the linguistic competence of a person and are used in the process of communication, performing a certain function in the process of communication [23].

The reason for choosing an empirical object is a high interest in the volunteer movement. The completed study provides additional information about its internal content. Like most others, this movement is increasingly using social media to communicate and carry out activities. At the same time, the topic of learning volunteering and activism through the analysis of social media texts in the Scopus database is devoted to a relatively small number of publications - less than 40. Only 5 of them describe the use of the topic modeling method. These works are carried out in the context of studying activism in the context of neoliberal policies [24], different strategies of activity in Twitter and Facebook [25], [26] and based on texts in English. There are no scientific discussions of Russian-language volunteer forums in the international database of publications.

Our proposals to modify the case of topic modeling procedure are the following:

*1)* the range of analysis of thematic models is expanded by considering the partial composition of the identified topics in the analysis;

*2)* pre-processing of the data was supplemented by the procedure for detecting stable speech revs characteristic of the movement studied, as well as an author's list of typical stop words.

The validation of the quality of the models was based on the coherence coefficient.

The expected result of the research task is to identify in the collection of text messages volunteer communications characteristics that characterize this group.

The hypothesis is that the forums of a social movement reflect its characteristics, which are repeated from one discussion to another, and reflect the essence and activity of this movement.

Three groups of the Russian social media VKontakte were chosen as an empirical object: "All for the Victory!", "Center of (City) Volunteers of St. Petersburg", "Volunteers of St. Petersburg".

### *B. Datasets and Methods*

**Description of the dataset**. A brief description of VKontakte social media groups whose messages were considered as raw data:

"All for the Victory!" – the community of people who do care» (https://vk.com/vsezapobedu, last accessed 2021/02/13). A group focused on preserving the traditions of the country's past, as well as creating a culture of historical assessments, the group includes more than 100,000 subscribers.

"Saint Petersburg (city) volunteer center" (https://vk.com/volunteerspb, last accessed 2021/02/13). A group that aims to create and maintain an effective system of education, selection and training (training) of volunteers for quality events at any level, more than 15,000 subscribers.

"St. Petersburg Volunteers" (https://vk.com/volspbcenter, last accessed 2020/09/13). The largest group of service and event volunteering in St. Petersburg for successful, caring, and promising young people, more than 50,000 subscribers.

Corpuses of texts of the considered social groups on VKontakte were collected in May 2020 (Celebration of the 75th anniversary of Victory in the Great Patriotic War) and contain messages published between 2012 and 2020. Descriptive statistics characterize the number of messages in these three groups, the average number of comments per message, and the average number of likes and reposts per message, reflecting the level of activity in the groups and their popularity. (Table I)

TABLE I. Descriptive statistics for the dataset

| Group (transliteration of the original name/ English) | Number of posts | Average number of comments | Average number of likes | Average number of reposts |
|---|---|---|---|---|
| Vsezapobedu/ All for the victory! | 14424 | 4,98 | 94,35 | 9,58 |
| Volspbcenter/ Center of (City) Volunteers of St. Petersburg | 10456 | 4,97 | 35,15 | 2,41 |
| Volunteerspb/ Volunteers of Petersburg | 8761 | 5,08 | 31,18 | 2,36 |

So, back to the tasks set, consider how thematic modeling can act as a tool for analyzing social movements, what information can be obtained from the analysis of social networks; what effects can be unearthed. To meet this challenge, we need automated data processing.

**Data preprocessing**. Pre-processing of text data to obtain topic models required graphematic and morphological analysis.

The graphematic analysis phase included tokenizing the text to sentences and tokenizing sentences into words, except for numbers, punctuation, symbols, and "stop words", which are prepositions, alliances, binding parts of speech. Pronouns are also often "stop-words," but pronouns are retained in this study because they have values for group identification.

So, for example, the sentence "We remind you that the acceptance of works for the second nomination of the #MyHistory "Every family has heroes" contest continues" is transformed into the following set of words: remind continues to accept works nomination competition my story is the heroes of every family.

At the stage of morphological analysis, lemmatization was carried out, that is, the transformation of the word into its original form with morphological parser of Russian pymorphy2. From the example above, we get the following lemmas: "remind", "continue", "reception", "work", "nomination", "competition", "wash", "history", "hero", "everyone", "family" etc.

At this stage, we see an example of a lemmatization error. As "washing" a pronoun, important for us, it is transformed in "to wash" with distortion of meaning, it is added manually in all grammatical options "washing", "mine", "ours". Several lemmas are also possible, for example, "hero," "heroic." Lemmatization in this case has advantages over stemming, which mechanically excludes prefixes, suffixes, and endings, which is not applicable in working with pronouns important to us.

Then we proceed to create a collection of lemmas, the so-called "bag of words," presented in the form of a matrix in which each row corresponds to a message, and all the obtained lemmas correspond to columns, the values  in the "lemma-document" matrix reflect a measure of the importance of the lemma in the document. For better interpretability of models using "bag of words," combining words into stable phrases was carried out. An example of the use of pre-processing methods is given in Table II:

TABLE II. AN EXAMPLE OF PREPROCESSING A TEXT MESSAGE

| **Method** | **Results** |
|---|---|
| original text message | The Committee for Youth Policy and Interaction with Public Organizations has begun to accept applications and documents for submission to the award of the Government of St. Petersburg |
| tokenization | 'committee', 'for', 'youth', 'politics',' and ',' interaction ', 'with', 'public', 'organizations', 'begin', 'accept', 'applications',' and ',' documents', 'for', 'submission', 'to', 'award', 'government', 'saint', 'petersburg' |
| lemmatization | 'committee', 'for', 'youth', 'politics',' and ',' interaction ',' with ',' public ',' organization ',' begin ',' accept ',' application ',' and ',' document ',' for ',' presentation ',' to ',' award ',' government ',' saint petersburg ' |
| stop-words deleting | 'committee', 'youth', 'politics',' interaction ',' public ',' organization ',' begin ',' accept ',' application ',' document ',' presentation ',' award ',' government ',' saint petersburg ' |
| identification of stable phrases | 'committee', 'youth_policy', 'interaction', 'public_organization', 'begin', 'accept', 'application', 'document', 'presentation', 'award', 'government_saint-petersburg' |

**Vector data model**. Data clustering requires the documents to be expressed in numerical format. This work uses a Vector Space Model (VSM) - a mathematical model for representing textual data in a single vector space R$^n$. Each document is represented as a vector $(a_1, . . . , a_n)$, where $a_i$ is the weight $i$ of the lemma, reflecting its "importance" for the document. The weight $ai$ was determined using the weighting function tf-idf (term frequency or lemma frequency; inverse document frequency). So a document $d_i \in D$ in the form of a vector from the space R$^n$ has the form: $d_i = (tf_{i1}, . . . , tf_{in})$, where $tf_{ij}$ is a number with which lemma $t_j$ occurs in document $d_i$. Therefore, if a lemma occurs in a document $n$ times more often than another, then its value for analysis is $n$ times higher. Let's reduce this indicator:

$$wtf_{ij} = \begin{cases} 1 + \ln tf_{ij}, & if\ \ tf_{ij} > 0 \\ 0, & if\ \ tf_{ij} = 0 \end{cases} \tag{1}$$

However, the words that are constantly repeated in Russian-language texts do not reflect the specifics of the specific document under study. Words that are too often found in texts are usually not informative. To take this nuance into account, we find the inverse frequency of the document:

$$idf_j = ln\frac{|D|}{\left|\{d: t_j \in d \wedge d \in D\}\right|} \tag{2}$$

The numerator contains the number of documents in the collection, and the denominator contains the number of documents of the set $D$ in which the lemma $t_j$ occurs. One of the modifications of the formula is "softened" $idf$:

$$smooth_{idf\,j} = ln\frac{|D| + 1}{\left|\{d: t_j \in d \wedge d \in D\}\right| + 1} \tag{3}$$

The final formula for calculating the measure of "importance" of the lemma in the message:

$$\text{Tf–idf}_{ij} = \text{tf}_{ij} \cdot \text{idf}_{ij} \tag{4}$$

**Latent Dirichlet allocation method**. Topic modeling was carried out using the latent Dirichlet allocation (LDA) method implemented in the Gensim package together with the Mallet implementation. Latent Dirichlet allocation (LDA) is a mathematical method for identifying topics based on the assumption that the words that make up the text of a document are taken from a set of topic lemmas. The topic itself represents the probabilities with which each lemma from the dictionary belongs to this topic.

The LDA approach to topic modeling is a sequence of the following actions:

1. Each document is viewed as a set of topics in a certain proportion. Let us denote the set of topics by the letter $Z$. The number of topics is determinable. Let us set the task to determine $k$ topics $(|\ Z\ | = k)$ of a collection $D$.

2. Each topic is considered as a set of key lexemes in a certain proportion. The lemma will be denoted by $w$ and, accordingly, the set of all lemmas in the collection $W$.

3. Displaying the distribution of topics in documents and the distribution of keywords by topic. We represent this probability by the formula for multiplying probabilities: $p(d, w) = p(w|d) \cdot p(d)$. Consider the matrix $F = [p(w|d)]_{w \in W, d \in D}$. We represent this matrix as the product of the probability matrix of topics in documents $\Theta_{|D| \times k}$ with elements $\theta_{dz} = p(z|d)$ and the probability matrix of tokens in topics $\Phi_{k \times |W|}$ with elements $\varphi_{zw} = p(w|z)$. Let's get a probabilistic model:

$$p(d, w) = \sum_{z \in Z} \mathrm{p(w|z)} \cdot \mathrm{p(z|d)} \cdot \mathrm{p(d)} \tag{5}$$

4. Tuning hyperparameters: $\alpha \in \mathrm{R}^k$ and и $\beta \in \mathrm{R}^{|W|}$. The hyperparameter $\alpha$ is responsible for the expressiveness of topics in documents. The smaller $\alpha$, the sparser the distribution vector will be. The hyperparameter $\beta$ determines the sparsity of the vector describing the distribution of tokens in the topic. The recommended value is $\beta z$ = 0, 01. The process of generating documents in the LDA model is as follows: 1) for each document $d \in D$, a random vector $\theta d$ is selected that obeys the Dirichlet distribution law with coefficient $\alpha$; 2) the topic $zd_i$ is selected from the polynomial distribution with the parameter $\theta d$; 3) a lexeme $wd_i$ is selected from the distribution $\varphi zd_i$, which is the Dirichlet distribution with the coefficient $\beta$. The probability density function for the Dirichlet distribution of the random variable $x = (x_1, ..., x_k)$ with the parameter $\alpha = (\alpha_1, ..., \alpha_k)$ has the form:

$$f(x|\alpha) = \frac{\Gamma(a_1 + \cdots + a_k)}{\Gamma(\alpha 1) \ldots \Gamma(a_k)} \, x_k^{a_1 - 1} \ldots x_k^{a_k - 1} \tag{6}$$

To identify the parameters of the model, we use Gibbs sampling - an algorithm for obtaining a sample from the joint distribution of several random variables.

The quality of the models is determined based on the complexity (Perplexity) of the model and the consistency (Coherence) of the topic and is a convenient measure for assessing the quality of topic modeling. A topic is called coherent if the terms most frequent in this topic are not accidentally often found together in collection documents. In this work, a coherence measure was used, which is the log conditional probability (LCP), which estimates the probability of a less frequent word under the condition of more frequent and is calculated using the formula (7).

$$LCP(t) = \sum_{i=1}^{k-1} \sum_{j=i}^{k} \log \frac{N(w_i, w_j)}{N(w_i)} \tag{7}$$

where $\mathrm{w_i}$ is the i-th term in descending order, N (w) is the number of documents in which the term w occurs at least once, N (w, w′) is the number of documents in which the terms w, w′ occur side by side though once.

Perplexing is a measurement of how well a probabilistic model predicts a theme. This indicator is also used to compare probabilistic models. Low perplexing indicates that the probability distribution is well suited for sample prediction.

The approach to finding the optimal number of topics is to build a set of LDA models with different values of the number of topics (k) and select the one that gives the greatest coherence value. The choice of "k," indicating the end of the rapid growth in theme consistency, usually offers meaningful and interpretable topics. Choosing a higher value can sometimes yield more detailed subthemes. The rule must be followed: if the same keywords are repeated in several topics, this is a sign that "k" is too large.

With the help of experts, lexical groups were analyzed in terms of the presence of lexemes in them reflecting the identification, behavioral and emotional components of the volunteer movement in coordination with the social functions of volunteering and public expectations regarding this group. Experts were tasked with identifying the following lexemes in each topic, if any:

*1)* verbs and other parts of speech reflecting the joint activities of the participants in the event,

*2)* nouns or pronouns denoting group identity,

*3)* adverbs and other parts of speech characterizing the emotional assessment of events.

To solve the problem, experts evaluated the topics individually. According to the results, the token was considered allocated if five or more experts named it. In total, eight experts with academic degrees of candidate and Doctor of Sciences in sociological, psychological, and philological specialties participated in the examination.

### *C. Results*

**Topics highlighted by topic modeling**. As a result of automatic classification based on k-means clustering methods and topic modeling of the LDA body of texts of three VKontakte groups, the optimal number of topics was 23. In the volunteer group - 8 topics, in the volspbcenter group - 11 topics. For all topic models obtained, coherence is of high importance, which makes the models of high quality. The optimal number of topics in each group was determined based on the calculation of the coherence coefficient for the range of 2 to 40 topics. (Fig. 1).

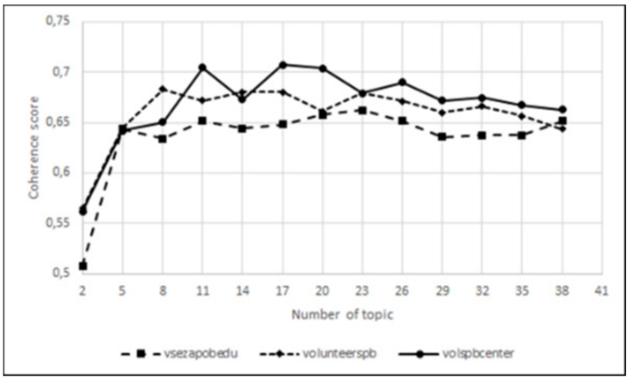


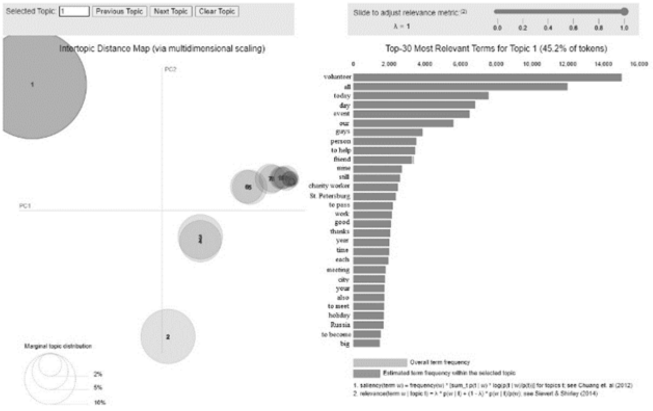
Fig. 1. Coherence score with different number of topics

The number of tokens in the topic was 16-28.

Example topic: 0.115 volunteer; 0.043 event; 0.024 to take place; May 0.017; 0.013 pass; 0.012 St. Petersburg; 0.012

required; 0.009 registration; 0.008 Palace Square; 0.008 to participate; 0.007 thanks; 0.007 competition; 0.006 collection; 0.006 start; 0.006 expensive; 0.006 participant.

Topics are grouped based on the core, that is, tokens with the highest coefficients at the top of the list. This is visually shown in Fig. 2.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig. 2. An example of the distribution of tokens in the topic "Volunteers at an event"

For the example above, these are lexemes volunteer; the whole, today, day, the event, united by the conventional name: "Volunteers at the event."

In total, the thematic model allowed us to identify three groups of topics for volunteers’ communications:

*1)* topics on coordination of volunteer activities (Search for volunteers. Organization of a meeting of volunteers. List of participants. Coordination of the meeting. Discussion of the past day. Preparation for the event).

*2)* topics of discussion of internal volunteer life (Competition of projects. We go to the theater).

*3)* topics of discussion of events with the participation of volunteers (Victory Day in the Great Patriotic War. Historical memory. Congratulations to veterans. Life of wonderful people. Volunteers at the festival. Volunteers at the city holiday. Volunteers at the event).

Lexemes are grouped together in themes by common distribution parameters. They have common contexts and related meanings. Therefore, it is necessary to analyze the part-of-speech composition of topics.

**Part-of-speech composition of topics**. The main feature of the proposed thematic model is the part-of-speech confusion of the composition of the themes. To solve the problem, we are interested in the parts of speech that denote 1) action, 2) identity and 3) emotionality. Here we are based on the provisions of social psychology about the group and group dynamics [27].

The following types are distinguished among action tokens:

*1)* lexemes of actions in a socially neutral sense (meet, wait, call, be, participate, apply, register, pass, etc.);

*2)* lexemes of denoting altruistic and voluntary actions (help, support, dedicate, be needed, etc.);

*3)* lexemes of action in the meaning of self-development, improvement (become, learn).

The lexemes of identity, reflecting identity, identification was expressed in parts of speech connected by syntactic connection with lemmas: "we", "you", "they", "our", "yours", "them". Identity lexemes reflected dominant identities:

*1)* group (volunteer, volunteer corps, desired volunteer, organizer, representative, participant, team, become a part, etc.);

*2)* national-civil (Russia, our country, Russian, volunteers_Russian_patriot, etc.);

*3)* territorial (volspb, volunteerspb, volunteers of St. Petersburg, Palace Square, Moscow, city, volspb2020_city volunteersspb, etc.);

*4)* symbolic (dress code, form No. 1, list).

The following two types dominate in the lexemes of emotionality:

1) emotions of heroization (gratitude, great holiday, the most, great, hero, feat, victory, etc.);

2) emotions of approval (good, necessary, thank you, dear friend, etc.).

## III. Results Discussions and Conclusion

The above analysis shows how topic modeling of texts is carried out based on the partial composition of topics.

The focus was on messages in volunteer communications of the social media VKontakte. Due to the adaptation of the topic modeling method, the topic of messages of volunteer communications was detailed and structured. The analysis of the partial composition of topics made it possible to reveal the orientation of volunteer activity, group identity and emotional background, as well as to identify the value content of volunteering - altruism, patriotism, citizenship, etc. The data received respond to the request of science and society, contributing to overcoming skepticism about volunteering and resolving disputes about the real motivation of volunteering and charity [28], [29], [30].

Summarizing, the method of topic modeling is useful for studying social movements, their values, goals, which are translated into society. Repeated research using the method allows it to be adapted, step by step to increase its effectiveness and in the future to develop a tool combining refined machine topic modeling with human interpretation. Topic modeling in the proposed adaptation plays the role of ambulance in solving controversial situations in the social sciences when interpreting a particular social phenomenon. Topic modelling helps to remove doubts caused by conflicting results from mass surveys and other cumbersome methods. But this positive side of it turns into several restrictions.

The difficulty is that topic modeling in the proposed adaptation is possible only after obtaining preliminary theoretical knowledge about the structure and properties of the empirical object. Therefore, topic modeling in this embodiment is more applicable to the analysis of social movements, knowledge of which contains a large amount of detailed information. While to analyze movements spontaneously formed in connection with a new event, additional developments are required.

The possibilities for such clarifications are limitless the method is constantly being improved. So, for example, Qing Deng adapts the method to analyze the discussion of anthropogenic disasters but says that the study of earthquake discussions has its own specifics [6]. In this regard, the question remains: is it possible and necessary to strive for the universality of the methodology in thematic modeling?

The results show the fundamental capabilities of technology to respond to several societal requests to increase the impact of empirical research in a transforming era and to find new approaches to data analysis and interpretation. In conclusion, we would like to draw attention to the following conclusions:

1. The developed topic model shows how manual work in text research can be greatly facilitated when working with a large amount of unstructured text data.

2. Analysis of the partial composition of topics is an effective tool for improving meaningful interpretation. In addition, the resulting discursive picture allows us to consider the partial composition of topics as a separate method.

3. The social significance of the results of topic modeling is that it makes it possible, by studying the topics of messages in the forums of various movements, to relate the dominant ideas, values, orientation of the activities of the "forums" to the expectations of society.

4. The prospects of thematic modeling are associated with improving the quality of pre-processing of texts, assessing the optimal parameters for building thematic models, expanding the text collection, working on interpretive analysis of materials, as well as identifying movements and groups of high social significance that affect the development of society and therefore is of particular interest as an object of thematic research.

## REFERENCES

[1] A. Korshunov and A. Gomzin, “Topic modeling in natural language texts,” *Proc. of the Inst. for System Programming of the RAS (Proc. of ISP RAS)*, vol. 23, 2012, pp. 215-244. Web: https://ispranproceedings.elpub.ru/jour/article/view/982. (RU)

[2] C. C. Aggarwal, and C. Zhai, Eds., *Mining Text Data*. New York: Springer-Verlag, 2012.

[3] D. M. Blei and J. D. Lafferty, “Dynamic topic models,” in *Proc. of the 23rd international conference on Machine learning*, New York, NY, USA, Jun. 2006, pp. 113–120.

[4] A. Triantafillidou, and P. Yannas, “Social media crisis communication in racially charged crises: Exploring the effects of social media and image restoration strategies,” *Comput. Hum. Behav.*, vol. 106, May 2020, 106. Web: 106269. 10.1016/j.chb.2020.106269.

[5] A. Y. K. Chua, and S. Banerjee, “The Topic of Terrorism on Yahoo! Answers: Questions, Answers and Users’ Anonymity,” *Aslib J. of Information Management,* vol. 72, Dec. 2019, pp. 1-16.

[6] Q. Deng, Y. Gao, C. Wang, and H. Zhang, “Detecting information requirements for crisis communication from social media data: An interactive topic modeling approach,” *Intern. J. of Disaster Risk Reduction*, vol. 50, 2020, Web: 10.1016/j.ijdrr.2020.101692.

[7] Y. Martín, S. L. Cutter, and Z. Li, “Bridging Twitter and Survey Data for Evacuation Assessment of Hurricane Matthew and Hurricane Irma,” *Natural Hazards Rev.*, vol. 21, Issue 2, May 2020, No. 04020003.

[8] L. Thomas, E. Orme, and F. Kerrigan, “Student Loneliness: The Role of Social Media Through Life Transitions,” *Computers & Education*, vol. 146, Mar. 2020, No. 103754.

[9] B. K. Bohrer, U. Foye, and T. Jewell, “Recovery as a process: Exploring definitions of recovery in the context of eating-disorder-related social media forums,” *Intern. J. of Eating Disorders*, vol. 53, no. 8, 2020, pp. 1219–1223.

[10] D. Ma, T. Osaragi, T. Oki, and B. Jiang, “Exploring the heterogeneity of human urban movements using geo-tagged tweets,” *Intern. J. of Geographical Information Science*, vol. 34, Issue 12, Jan. 2020, pp. 2475-2496. Web: 10.1080/13658816.2020.1718153.

[11] M. Paterson, and M.R. Glass, “Seeing, feeling, and showing ‘bodies-in-place’: exploring reflexivity and the multisensory body through videography “, *Social & Cultural Geography*, vol. 21, Issue 1, Jan. 2020, pp. 1-24, Web: https://www.tandfonline.com/doi/abs/10.1080/14649365.2018.1433866?scroll=top&needAccess=true&journalCode=rscg20.

[12] K. Xu, F. Wang, H. Wang, and B. Yang, “Detecting fake news over online social media via domain reputations and content understanding,” *Tsinghua Science and Technology*, vol. 25, Issue 1, Feb. 2020, pp. 20–27.

[13] J. Gray, L. Bounegru, and T. Venturini, “Fake news’ as infrastructural uncanny”, *New Media and Society*, vol.22, Issue 2, Feb. 2020, pp. 317-341, Web: https://journals.sagepub.com/doi/10.1177/146144481985691.

[14] M. del M. Gálvez-Rodríguez, J. Alonso-Cañadas, A. Haro-de-Rosario, and C. Caba-Pérez, “Exploring best practices for online engagement via Facebook with local destination management organisations (DMOs) in Europe: a longitudinal analysis”, *Tourism Management Perspectives*, vol. 34, 2020. Web: https://www.cabdirect.org/cabdirect/abstract/20203212145.

[15] D. Baishya, and S. Maheshwari, “WhatsApp Groups in Academic Context: Exploring the Academic Uses of WhatsApp Groups among the Students,” C*ontemp Educ Technol*, vol. 11, Issue 1, pp. 31–46, Nov. 2019.

[16] M. Y. Nejad, M. S. Delghandi, A. O. Bali, and M. Hosseinzadeh, “Using Twitter to raise the profile of childhood cancer awareness month”, *Netw Model Anal Health Inform Bioinforma*, vol. 9, Issue 1, Dec. 2020, pp. 1–5.

[17] W. Liu, X. Fan, R. Ji, and Y. Jiang, “Perceived Community Support, Users’ Interactions, and Value Co-Creation in Online Health Community: The Moderating Effect of Social Exclusion”, *International J. of Environmental Research and Public Health*, vol. 17, Issue 1, Dec. 2019, No. 204.

[18] J.-H. Kim, “An Exploratory Study of Health Inequality Discourse Using Korean Newspaper Articles: A Topic Modeling Approach”, *J. Prev. Med. Public Health*, vol. 52, Issue 6, Nov. 2019, pp. 384-392.

[19] B. Dahal, S. P. Kumar, and Z. Li, “Topic modeling and sentiment analysis of global climate change tweets,” *Social Network Analysis and Mining*, vol.9, Issue 1, Dec. 2019, article No. 24.

[20] A. P. Kirilenko and S. Stepchenkova, “Automated Topic Modeling of Negative Tourist Reviews,” *e-Review of Tourism Research*, vol. 17, Issue 4, 2020, Art. no. 4, Web: https://journals.tdl.org/ertr/index.php/ertr/article/view/538. (RU)

[21] T.-I. Yang, A. Torget, and R. Mihalcea, “Topic Modeling on Historical Newspapers,” in *Proc. of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Portland, OR, USA, Jun. 2011, pp. 96–104, Web: https://www.aclweb.org/anthology/W11-1513.

[22] T. Hofmann, “Probabilistic latent semantic indexing,” in Proc. of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, Aug. 1999, pp. 50–57.

[23] M. G. Kulikova, and S. V. Plevako, “The parts-of--speech paradigm of the Russian language: posing a problem”, *Scientific notes of the Transbaikal State Humanitarian Pedagogical University named after N.G. Chernyshevsky*, no. 3, 2009, pp. 217–220. (RU)

[24] M.I. Vasilkovskaya, “Institute of Youth Volunteering as a Socio-Cultural Phenomenon”, *The world of science. Sociology, philology, cultural studies*, vol. 2, Issue 9, 2018, p. 8, Web: https://sfk-mn.ru/en/08SCSK218.html. (RU)

[25] S. Shahin, “Facing up to Facebook: how digital activism, independent regulation, and mass media foiled a neoliberal threat to net

neutrality", *Information Communication and Society*, vol. 22, Issue 1, 2019, pp. 1-17, Web: https://www.tandfonline.com/doi/abs/10.1080/1369118X.2017.1340494.

[26] M. Chong, "Connective power of the twitter networks: Discovering the reverse agenda-setting effects of hashtag activism through topic modeling," in *Proceedings of the Association for Information Science and Technology*, vol. 56, pp. 629–630, Jan. 2019.

[27] D. Myers, *Social Psychology*. New York: McGraw-Hill Education, 2012.

[28] N.A. Zeynalova, "Motivation in the structure of intrapersonal conflicts among volunteers", in the collection: Psychological studio. Collection of articles by students, undergraduates, postgraduates, young researchers of the Department of Applied Psychology, 2020, pp. 96-98. (RU)

[29] A. A. Kuzminchuk, and D. F. Telepaeva, "Young Volunteers: Altruists or Egoists?", in *XXI Ural Sociological Readings. Social space and time of the region: problems of sustainable development,* 2018, pp. 424–429, Web: https://elar.urfu.ru/handle/10995/61438.

[30] P. S. Nedelko, "Main trends in changing of volunteers' motivation in Russia", in *Collection of articles based on the materials of the X International Scientific and Practical Conference, Ufa,* 2018, pp. 91–96, Web: https://www.elibrary.ru/item.asp?id=36622075.