Automatic Detection of the Topical Structure of the Ministerial Posts on Social Networks

Alena Zaitseva[1][0000-0003-2041-4238] and Ivan Mamaev[2][0000-0003-3362-9131]

[1] School of Journalism and Mass Communications, Saint Petersburg State University,
Russia, 199004, Saint Petersburg, VO, 1 Line, 26,
[2] Saint Petersburg State University,
Russia, 199034, Saint Petersburg, Universitetskaya emb. 11,
az998@mail.ru, mamaev_96@mail.ru

Abstract. The paper discusses the development of a corpus of Russian ministerial posts based on VKontakte social network. The study is aimed at revealing topical structure of ministerial communities. We performed a series of experiments which include LDA topic modeling and automatic topic labeling that help to improve the interpretability of topics. To implement the procedures, we used Python libraries for NLP. Experiments allowed us to find out pivotal topics that the government of Russia covers on social networks nowadays.

Keywords: Social Network, Ministerial Post, Corpus Linguistics, Russian, Topic Modeling, LDA, Automatic Topic Labeling

1 Introduction

In the era of digitalization of society, many areas of life are reflected on the Internet. Government agencies, services and ministries are no exception: using the available tools and applications of online platforms, departments create communities in which they publish posts about the current state of the country, as well as its regions.

There are a great number of ministries in Russia: The Ministry of Defense, The Ministry of Foreign Affairs, The Ministry of Culture, etc. At first glance, it seems that the names of the ministries are directly related to the topics of the problems and the posts that they publish on social networks. However, the topical structure of ministerial groups is much more complicated. For example, the coronavirus pandemic has affected many areas of the Apparatus of the Government of Russia, some common issues appearing among a lot of ministries. The most obvious one is related to the problems of online education and the ways to overcome them, they are dealt with by both the Ministry of Education and the Ministry of Digital Development, Communications and Mass Media.

This study is aimed at detecting main topical areas in ministerial posts on VKontakte social network. This social network is very popular among residents of Russia[1],

[1] https://gs.statcounter.com/social-media-stats/all/russian-federation/2019

therefore ministries and other governmental departments are eager to create their communities there. We will use LDA and topic labeling algorithms to reveal main topics of ministerial posts. These procedures have never been carried out on the corpus of ministerial posts.

2 Related works

Large text collections, known as corpora, are becoming more and more popular among linguists, political scientists, sociologists and a number of other scientists. They try to use corpora to carry out their practical researches.

In English papers [3, 5, 13, 14], functioning of politically oriented posts (for example, Twitter or Facebook) and their impact on the public opinion of users are described. Most of the authors note that the integration of governmental apparatus and social networks is connected with a certain degree of risk.

As regards the Russian segment of studying political discourse, it should be noted that there is a small layer of works on this topic. The papers [8, 9] are based on diachronic description of political vocabulary and its frequency behavior. The experiments were conducted on the basis of Google Books Ngram Viewer[2]. The authors pay special attention to the names of political figures and important events in the history of Russia. The corpus, that was used by the scholars, is compiled on the basis of a large number of digitized versions of printed publications, it does not focus on texts that one can find on social networks. In our study, we try to describe some linguistic features of ministerial posts on online platforms, it will fill in the gaps in the Russian corpus linguistics and topic modeling researches.

3 Experiment

3.1 Corpus collection and preprocessing

To conduct further experiments, there is a need to collect a corpus. We used Python[3] and the beautifulsoup4[4] library to create a parser for scraping posts from 15 communities of ministries, agencies and services on VKontakte social network. We took 2019-2020 posts as they reflect the current situation in Russia in various spheres of life.

The size of the final corpus is 2 311 480 words. The procedure for preprocessing the corpus involves the following steps:

1. Removal of non-textual elements (emoticons, images, etc.);
2. Obtaining tokens using regular expressions;
3. Lemmatization (normalization) of the received tokens and resolution of morphological ambiguity using pymorphy2[5];

[2] https://books.google.com/ngrams

[3] https://www.python.org/

[4] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[5] https://pymorphy2.readthedocs.io/en/latest/

4. Removal of the words (pronouns, prepositions, etc.) if they are in the stop-list;
5. Adding bigrams and trigrams with the help of gensim[6];
6. The division of preprocessed posts according to ministries and departments;
7. Saving the preprocessed corpus in .txt format.

3.2 Topic modeling

Topic modeling consists in the representation of compressed topical descriptions of documents. The topic model of a text collection defines each topic as a discrete distribution over a set of terms, each document is defined as a discrete distribution over a set of topics. The texts are presented as a sequence of topics, which are randomly and independently selected from some distribution. A topic in topic models is a set of words characterized by co-occurrence within a document.

Nowadays topic modeling is used in a great variety of computational linguistics researches: one can use the algorithms for detecting hidden communities on social networks, determining main topics of users' posts on Twitter, or revealing topics of social media news [1, 6, 7].

To implement further procedures, we need to choose a suitable algorithm for the experiment. We decided to focus on one of the popular algorithms for topic modeling – Latent Dirichlet Allocation (LDA). The gensim library is used for LDA as it provides a lot of possibilities to work with other libraries (for visualization, etc.).

First of all, we need to compute the optimal number of topics for the corpus using the U-Mass measure. It reflects topic coherence value which is treated as a level of human interpretability of the model based on relatedness of words and documents within a topic. The formula is as follows:

$$core(v_i, v_j, \epsilon) = \log \frac{D(v_i, v_j) + \epsilon}{D(v_i)}, \tag{1}$$

where $D(v_i, v_j)$ is the number of documents that contains v_i and v_j words, $D(v_i)$ shows the number of documents containing v_i words. The highest value of $core(v_i, v_j, \epsilon)$ shows that the model includes the appropriate number of topics. We conducted a series of experiments with the following parameters: the minimum number of topics was 5, the maximum was 70, and the step was 5. We obtained graphic data (Figure 1).

[6] https://radimrehurek.com/gensim/

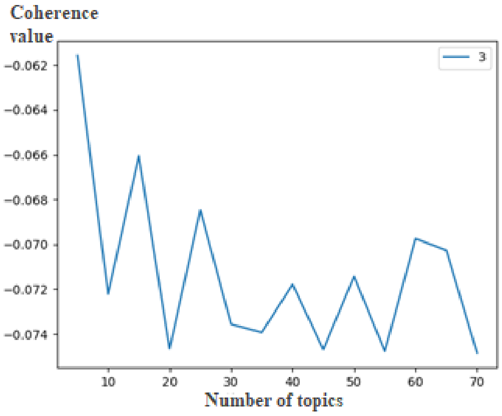

Fig. 1. Optimal number of topics for the ministerial corpus

Figure 1 shows that the coherence value is greatly decreased when the number of topics within the corpus is increased, so the optimal number of topics is 5.

The LDA algorithm is trained, the following parameters being used: the number of iterations is 10, the number of expected topics is 5, the number of topic words in the set is 10. The topics are represented as lists of lemmata.

Table 1. LDA topics within the ministerial corpus

Number of a topic	Topic
1	россия, российский, страна, дело, министр, международный, иностранный, федерация, вопрос, отношение (russia, russian, country, affair, minister, international, foreign, federation, issue, attitude)
2	россия, полиция, сотрудник, мвд, дело, российский, служба, полицейский, область, внутренний (russia, police, employee, ministry of internal affairs, affair, russian, service, police, region, internal)
3	военный, россия, российский, оборона, учение, сила, международный, конкурс, армия, флот (military, russia, russian, defence, exercises, force, international, competition, army, navy)
4	россия, российский, образование, школа, день, ребёнок, работа, культура, проект, мир (russia, russian, education, school, day, child, work, culture, project, world)
5	россия, российский, проект, развитие, министр, производство, страна, работа, область, участие (russia, russian, project, development, minister, production, country, work, region, participation)

As regards the graphic representation of topics in the corpus, the resulting sets can be visualized using pyldavis[7] (Figure 2).

[7] https://github.com/bmabey/pyLDAvis

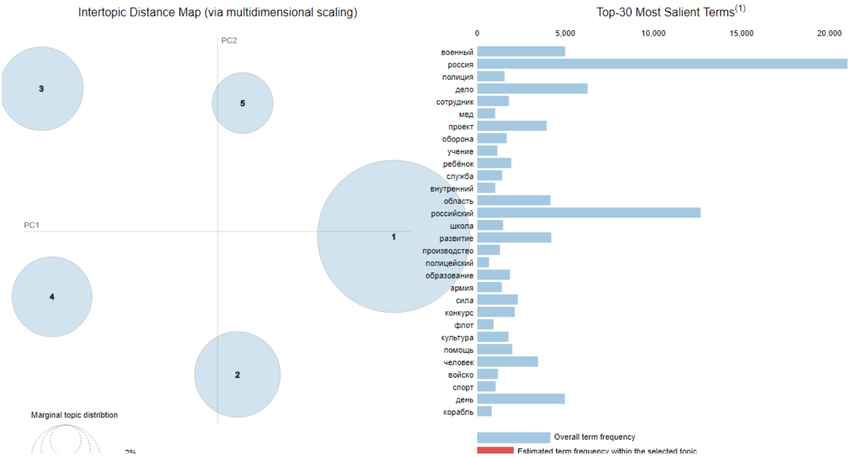

Fig. 2. Representation of topics of the ministerial corpus

It is clear that the first topic covers the biggest part of the text collection while there is much less information about the others.

It is also important to note that neither bigrams nor trigrams appeared in the topics. Their absence can be connected with low weights of the obtained n-grams.

The main disadvantage of LDA is the impossibility of automatic selection of topic labels: a user may have certain difficulties while interpreting the results of obtained topics. To find the most accurate label, scholars try to improve topic modeling algorithms and implement a method known as automatic topic labeling. The next section will discuss this procedure.

3.3 Automatic topic labeling

A label is a word or a sequence of words that covers the general content of given sets of words. Sometimes relevant labels are manually assigned to topics. However, the procedure of automatic topic labeling allows us to facilitate the interpretation of topics, as well as to save time and effort spent on manual assigning. Nowadays, it is being developed both for English and Russian corpora.

There are two main ways to get topic labels: external sources and internal sources [11]. External sources can refer to online thesauri (WordNet, etc.) or sites that help to obtain topical labels for a set of words (Wikipedia, etc.). As for internal sources, topical labels can be extracted using procedures of automatic summarization or creating word2vec models [4].

In our research we used some methods proposed in [7, 12]. First, we transformed the topical words into a query for Google, extracted first 5 site headlines, identified collocations with the help of pymorphy2 (ADJ+NOUN, NOUN+NOUN, etc.) and ranged them according to the number of occurrence in the Russian National Corpus. If there were no occurrences, we didn’t consider the label to be a candidate.

Then we created two word2vec CBOW models: the first one is based on the ministerial corpus itself, the second one is based on the corpus of users' posts of VKontakte social network, it is proposed in [7]. The parameters of the models: minimal frequency of occurrence is 5, the size of a vector is 100, the context window is 5. The results are ranged with the help of the cosine similarity.

Below we present a comparative table of candidates for topic labels.

Table 2. Results of automatic topic labeling

Topic	**Google and the Russian National Corpus**	**Word2Vec models**	
		Ministerial posts	**Users' posts on social networks**
россия, российский, страна, дело, министр, **международный**, **иностранный**, федерация, вопрос, отношение (russia, russian, country, affair, minister, **international**, **foreign**, federation, issue, attitude)	российская федерация, официальный представитель, **внешняя политика**, **министерство иностранных дел**, **международная жизнь** (russian federation, official representative, **foreign policy**, **ministry of foreign affairs**, **international affairs**)	президент, субъект, дело, правительство, председатель (president, subject, affair, government, chairman)	заявить, независимый, представитель, активист, российский (declare, independent, representative, activist, russian)
россия, **полиция**, сотрудник, **мвд**, дело, российский, служба, полицейский, область, **внутренний** (russia, **police**, employee, **ministry of internal affairs**, affair, russian, service, police, region, **internal**)	**мвд россии**, **правоохранительная система**, управление мвд, полиция россии, рубеж россии (**ministry of internal affairs of russia**, **law enforcement system**, department of the ministry of internal affairs, police of russia, border of russia)	павел, субъект, правительство, полковник, совместно (pavel, subject, government, colonel, mutually)	чиновник, активист, **следственный**, конституционный, антиконституционный, должность (official, activist, **investigative**, constitutional, anticonstitutional, occupation)
военный, россия, российский, **оборона**, учение, сила, международный, конкурс, армия, флот (**military**, russia, russian, **defence**, exercises, force, international, competition, army, navy)	российская федерация, **министерство обороны**, **военно-морской флот**, военная техника, военный округ (russian federation, **ministry of defence**, **navy**, military equipment, military district)	проходить, совместно, андрей, выступить, назначить (pass, mutually, andrew, come out, appoint)	социалистический, великобритания, гражданский, украинский, **оборона** (socialistic, great britain, civil, ukrainian, **defence**)

россия, российский, **образование**, школа, день, ребёнок, работа, культура, проект, мир (russia, russian, **education**, school, day, child, work, culture, project, world)	повышение квалификации, первая школа, государственный университет, **история образования**, главный портал (advanced training, first school, state university, **history of education**, **main portal**)	совместно, проходить, андрей, республика, инициатива (mutually, pass, andrew, republic, initiative)	молодёжь, **вуз**, инженер, отцовство, обучаться (youth, **university**, engineer, paternity, study)
россия, российский, **проект**, **развитие**, министр, производство, страна, работа, область, участие (russia, russian, **project**, **development**, minister, production, country, work, region, participation)	национальная программа, министерство науки, деловая программа, **управление экономики, цифровая экономика** (national program, ministry of science, business program, **department of economics**, **digital economy**)	генеральный, моисеев, дума, проходить, грамотность (general, moiseev, duma, pass, literacy)	**предпринимательство**, **стратегический**, экологический, региональный, федеральный (**enterprise**, **strategic**, ecological, regional, federal)

The easiest way of obtaining a topic label is to choose the first word in a set, but russia is the first word in all the sets, as it is one of the most frequent within the corpus. It cannot be a topical word, so there is a need to analyze the results of using internal and external sources. The idea is to find the same words or collocations in all the columns or try to find words which can describe the same semantic field.

Mind the word2vec model based on the ministerial corpus. The resultant labels are repeated sometimes; it may be connected with the size of the corpus. The size of the second corpus is almost four times as many as the size of the first one, so there are more relevant results. At the same time, a lot of proper names are represented as candidates for topic labels. The names denote people working in a particular ministry: for instance, Moiseev works in the Ministry of Finance. Although names are indirectly related to the topical sets of words, they cannot be candidates for labels.

4 Results and Evaluation

After running several experiments, we obtained the topical structure of ministerial posts on VKontakte social network. The main topics are related to:

1. foreign policy;
2. internal affairs;
3. country defense;
4. education;
5. country development.

Such topics as health, transport, protection of environment, etc. were not mentioned in the models. It may be connected with several issues.

1. A ministry doesn’t have a topical community on the social network or it doesn’t publish a lot of posts on social networks so users are unable to get all the information on the current state of a ministry. This fact is closely linked to the degree of the state “openness”[8].
2. An obtained topic itself contain more specific ones. For instance, the last set of words describes both economic and industrial processes in Russia.
3. The *health* topic, that is still acute because of the coronavirus pandemic, is scattered across various ministerial communities, each department assesses the impact of the coronavirus on its own area. This is why the topic of health is most likely absorbed by larger topics.

As regards the assessment of the procedures, we will discuss the main advantages and disadvantages. First of all, nowadays a lot of scholars have analyzed different Russian corpora of social networks with the help of LDA algorithm, a great number of papers proving it [1, 2, 12]. Moreover, we can reveal the inner structure which is described by certain syntagmatic and paradigmatic relations between words within each topic [10].

Table 3. Some syntagmatic and paradigmatic relations

Syntagmatic relations	Paradigmatic relations
Adjective-modifier relations: российский – федерация (russian - federation) международный – отношение (international - relation)	**Hypernymy and hyponymy:** страна – россия (country - russia) сотрудник – полицейский (employee – policeman)
Noun-modifiers: оборона – сила (defence – force)	**Derivational relations:** россия – российский (russia - russian) полиция – полицейский (police – policeman)
Adjective-modifier and noun-modifiers relations: министр – иностранный – дело (minister – foreign – affair)	**Meronymy and holonymy:** оборона – армия, флот (defence – army, navy) образование – школа, ребёнок (education – school, child) россия – область (russia – region) работа – производство (work – production)

Mind that verb-modifier relations were not mentioned in the table. The reason is that verbs in the corpus of ministerial posts can be either high-frequency or low-frequency: сказать, встретить, подписать, обсудить и т.д. (say, meet, sign, discuss, etc.) so they were not included in the resultant topic models.

As regards topic labels, we should note that the main advantage is using a double-ranking algorithm (Google PageRank and the number of occurrence in the Russian

[8] https://ach.gov.ru/news/gosudarstvo-sredney-zakrytosti-rezultaty-novogo-reytinga-otkrytosti-gosorganov

National Corpus), it helped to avoid candidates with low frequencies. The problem was to assess the quality of the obtained labels, as there is no gold standard for Russian topic labels. It was decided in involve 5 independent assessors, they were asked to assess each label on Google Forms using the following grades: 1 is for relevant labels, 0 is for irrelevant ones. Some results are presented below (Figure 3).

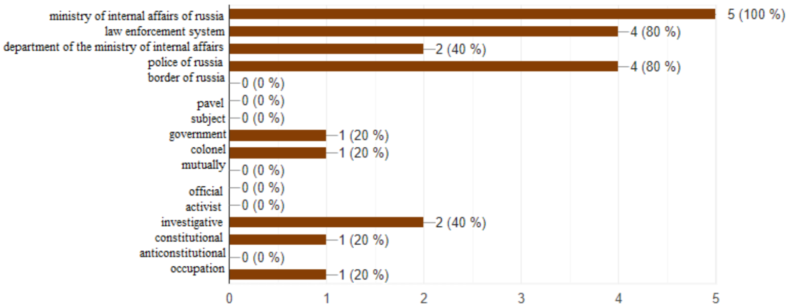

Fig. 3. Diagram of candidates for topic labels for the topic: russia, police, employee, ministry of internal affairs, affair, russian, service, police, region, internal

Then we transformed the chart into the table for better interpretation. Below there is a part of such a table (Table 4).

Table 4. Some examples of the results of an expert assessment of candidates for topic labels for the topic: *russia, police, employee, ministry of internal affairs, affair, russian, service, police, region, internal*

n-grams	*мвд россии (ministry of internal affairs of russia)*	*правоохранительная система (law enforcement system)*	*управление мвд (department of the ministry of internal affairs)*
Assessing results	5	4	2
n-grams	*полиция россии (police of russia)*	*рубеж россии (border of russia)*	
Assessing results	4	0	
Unigrams	*павел (pavel)*	*субъект (subject)*	*правительство (government)*
Assessing results	0	0	1
Unigrams	*полковник (colonel)*	*совместно (mutually)*	*чиновник (official)*
Assessing results	1	0	0

As far as you can see, the assessors agreed that the ministry of internal affairs of russia label is the best option for the topic, and all the unigram labels are the worst ones. It can be explained by the fact that n-grams better describe specific topics, and unigrams are often used to cover common topics [7].

5 Summary

Automatic ways of analyzing texts on social networks have become pivotal nowadays. In this paper, we created the corpus of ministerial posts, performed linguistic analysis of the data with the help of LDA topic modeling and topic labeling. We described the topical structure of the text collection and found out the main topics which are important for the Russian government. We also analyzed paradigmatic and syntagmatic relations between lexical units within each topic.

Results, that were obtained during the experiments, prove consistency of the statistical model and provide common knowledge on current issues of Russian ministries.

Further researches can be carried out in the following directions:

- the enlargement of the corpus by adding information from other social networks (Facebook, etc.) and comparison their topical structures;
- comparing several topic modeling algorithms (for instance, LDA and LSI);
- developing a gold standard for automatic topic labeling of Russian topic models;
- classifying and clustering texts of the ministerial posts within one corpus;
- automatic extraction of keywords from ministerial posts.

References

1. Bodrunova, S., Blekanov, I., Kukarkin, M.: Topic modeling for Twitter discussions: Model selection and quality assessment. In: Proceedings of the 6th SGEM International Multidisciplinary Scientific Conferences on SOCIAL SCIENCES and ARTS SGEM2018, Science and Humanities, 207–214. STEF92 Technology Ltd., Sofia, Bulgaria (2019).
2. Bodrunova, S., Blekanov, I., Kukarkin, M.: Topics in the Russian Twitter and relations between their interpretability and sentiment. In: Sixth International Conference on Social Networks Analysis, Management and Security, 549–554 (2019).
3. Bodrunova, S., Blekanov, I., Smoliarova, A., Litvinenko, A.: Beyond Left and Right: Real-World Political Polarization in Twitter Discussions on Inter-Ethnic Conflicts. Media and Communication, 7(3), 119–132 (2019).
4. Cano Basave A.E., He Y., Xu R.: Automatic Labelling of Topic Models Learned from Twitter by Summarisation, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Stroudsburg, PA, USA, Association for Computational Linguistics, 618–624 (2014).
5. Garrett, R. K.: Social media's contribution to political misperceptions in U.S. Presidential elections. PLOS ONE, 14(3), 1–16 (2019).
6. Koltsov, S., Pashakhin, S., Dokuka, S.: A full-cycle methodology for news topic modeling and user feedback research. In: Staab, S., Koltsova, O., Ignatov, D.I. (eds.) SocInfo 2018. LNCS, 11185, 308–321. Springer, Cham (2018).
7. Mamaev I., Mitrofanova O.: Automatic Detection of Hidden Communities in the Texts of Russian Social Network Corpus. In: Filchenkov A., Kauttonen J., Pivovarova L. (eds) Artificial Intelligence and Natural Language. AINL 2020. Communications in Computer and Information Science, 1292, 17–33. Springer, Cham (2020).
8. Masevich, A., Zakharov, V.: Corpus Linguistics Methods in Humanitarian Studies. Computational linguistics and computational ontologies, 24–43 (2016).

9. Masevich, A., Zakharov, V.: Variation in the representation of names of political functionaries in diachronic researches based on text corpora. Computational linguistics and computational ontologies, 56–73 (2018).
10. Mitrofanova, O.: Probabilistic Topic Modeling of the Russian Text Corpus on Musicology. In: LMAC 2015, CCIS 561, 69–76. Springer Nature (2015).
11. Mitrofanova, O., Mirzagitova, A.: Automatic assignment of labels in topic modeling for Russian corpora. In: Proceedings of the 7th Tutorial and Research Workshop on Experimental Linguistics, 115–118 (2016).
12. Mitrofanova, O., Sampetova, V., Mamaev, I., Moskvina, A., Sukharev, K.: Topic modelling of the Russian corpus of Pikabu posts: author-topic distribution and topic labelling. In: Proceedings of the International Conference « Internet and Modern Society» (IMS 2020), International Workshop «Computational Linguistics» (CompLing-2020) (2020, in press).
13. Mohammad, H. B.: Government's Presence on Social Media. A Study with Special Reference to Jordan. In: Research Journal of Applied Sciences, Engineering and Technology, 7, 4813–4816 (2014).
14. Stieglitz, S., Dang-Xuan, L.: Social media and political communication: a social media analytics framework. Soc. Netw. Anal. Min. 3, 1277–1291 (2013).