Topics of ethnic discussions in Russian social media

Authorship info has been withdrawn for anonymous reasons.

Abstract. The paper reveals the topic structure of ethnic discussions in the Russian-speaking social media and presents. The informational basis for the study is 2,659,849 texts from Russian-speaking social media published for two-year period from 2014 to 2015. Each text contained at least one ethnonym from constructed list. We conclude that the ethnic discussions are full of problematic topics. The most salient topics are the topics about Ukraine-Russia relations over the recent conflict between two countries. The study also shows racial bias in criminal topic towards peoples of the North Caucasus which are often mention in the context of crimes and terrorism.

Keywords: ethnicity, topic modelling, social media, Russia.

1 Introduction

The topic of ethnicity has been interested social scientists for long time. The reason is that ethnicity traditionally refers to a very small number of prescribed social statuses, that are acquired from birth, regardless of person's will and desire and cannot be changed. As such statuses are often are perceived as an integral and unchanging part of the human personality they experienced by people more intimately and have a high potential for conflicts. These circumstances make it important to study ethnic issues, especially in a such polyethnic country like Russia.

Public attitudes to various ethnic minorities are often far from being equal. In particular, in Russia, according to many polls (Bessudnov, 2016) Caucasian[1] have been found to arouse the most negative attitudes, followed by Central Asians. These attitudes may stream from the "offline" to the "online" space and easily disseminate to the large audiences, they become a factor influencing interethnic conflict (Gibson & Lando, 2001) and hate crime (Chan, Ghose, & Seamans, 2016) back in the offline world. Therefore, it is important to monitor ethnicity-related online content and to develop instruments for such monitoring. In this study we seek to reveal topic structure of ethnicity related discussion in the Russian language social media and compare public attention to different ethnic groups.

2 Related work

With the explosive development of the Internet and the accompanying phenomena, many works devoted to the study of ethnic discursions in the global network have appeared. An example is an article whose authors studied the differences in the self-

---

[1] In Russian language the term "Caucasians" denotes the peoples of Caucasus region.

presentation of various ethnic groups in the Facebook social network site (Grasmuck, Martin, & Zhao, 2009). The study of self-representation in a non-anonymous social network was presented as one of the distinctive features of the work. According to the authors, non-anonymous social networks sites encourage users to be more realistic and honest in their representations. The empirical material for the study was the student's profiles of 83 Facebook users of various nationalities supplemented by 63 interviews. As a result, the researchers found that ethnicity is very noticeable among Facebook users. African Americans, Latinos and Indians (Native Americans) create more elaborate personal pages that signal their ethnicity much more clearly than White or Vietnamese. Some users emphasize their ethnicity, opposing the tendency to conceal ethnic differences.

Ethnic studies on the Internet are not limited to just describing how ethnicity manifests itself in online space. Researchers from the United States have shown that this relationship operates in both directions (Chan et al., 2016). They found that Internet access is positively related to the number of crimes motivated by racial hatred: more people have Internet access have in certain area — higher level of racism is there.

It is also worth to mention a large amount of literature describing methods for identifying ethnic enmity in social media (Davidson, Warmsley, Macy, & Weber, 2017; Xiang, Fan, Wang, Hong, & Rose, 2012 and other works), but these works usually do not provide any substantive results, focusing on methodological part of the research.

There are also articles on ethnicity in Russian media. The two following articles are also study ethnic identity on the social networking site, but this time the site is VKontakte — the largest Russian social networking site. The first work, authored by Dilyara Suleymanova, is called "Tatar Groups in Vkontakte: The Interplay of Between Ethnic and Virtual Identities on Social Networking Sites" (Suleymanova, 2009). She studied communities formed around issues related to Tatars. The author claims that Vkontakte is a powerful tool for constructing ethnic networks, connecting Tatars from all over Russia, and also functioning as a tool for building ethnic identity and negotiation. On the one hand, the Tatar groups reproduce and mobilize traditional ideas about what it means to be a Tatar using topics of Tatar language and Islam. Moreover, some Tatar groups construct alternative versions of the Tatar identity.

The authors of the next work — researchers from the Higher School of Economics, Daniil Alexandrov, Aleksey Gorgadze and Ilya Musabirov — investigated Caucasian groups in the social network Vkontakte (Alexandrov, Gorgadze, & Musabirov, 2016). Using network analysis, they built a network of groups base on common membership. The authors identified different clusters of ethnic groups and determined the proximity of these clusters. Thus, it was revealed that clusters containing Armenian and Azerbaijani groups do not contain common groups, apparently due to the Nagorno-Karabakh conflict between Armenia and Azerbaijan. Also, the authors used topic modeling to reveal the nature of relationships between groups in clusters. They found that religion, being not so salient topic, facilitates the establishment of links between clusters.

Researchers are also examined how various ethnic groups are presented in media space. For example, according to some researchers, the image of Chechens (one of the ethnic groups living in the North Caucasus) "has been subjected to such demonization in the Russian media that virtually any publication relating to Chechens — even if it has nothing to do with the armed conflict in Chechnya - turns out imbued with the language of hostility' (Ahmetieva, 2007, p. 203). This opinion was confirmed in the course of research into the texts of Russian-language blogs to identify attitudes toward labor immigrants (Bodrunova, Koltsova, Koltcov, & Nikolenko, 2016). The researchers found that people from the North Caucasus are portrayed in the blogosphere as aggressive and hostile subjects. A deeper, qualitative analysis has shown that Daghestanians are portrayed as mighty barbarians, and Chechens as terrorists. Previous studies on the topic of attitudes toward migrants also show that it is those who come from the North Caucasus who cause the most negative attitude of Russians (Bessudnov, 2016; Foxall, 2014).

One of the most relevant studies on the role of ethnicity in the mass media was carried out by Fadeev in his work (Fadeev, 2017). The hypothesis of this study said that more integrated ethnic groups (Ukrainians and Jews) are more positively represented in the media of St. Petersburg compared to the relatively new ones (Tajiks and Chechens). This hypothesis was partially confirmed. Just as in our study, as a data collection tool, the author used commercial media aggregator, configured to collect texts from the media containing the necessary words, and, like us, in this study, an attempt was made to link the use of ethnonyms with the general context of the message. As the main method of analysis, hand coding of the tonality and subject of messages was selected with the subsequent analysis of the percentage distributions of these characteristics for each ethnicity.

The aforesaid literature review on the topic of ethnicity in the Internet shows that the authors writing the topic of ethnicity mainly consider separate diasporas (groups in social networking sites). Our work compares favorably with the fact that it is not limited to the study of any one ethnic diaspora but aims to describe ethnic discursions on the macro level.

3 Data

The empirical object of the research were texts on ethnicity from Russian-speaking social media. In addition, the data we needed were to satisfy two additional conditions: 1) they should be relevant to ethnic discussions; 2) they should be representative of all Runet texts on this topic.

To achieve the first of these conditions, we compile a comprehensive list of words used to search for texts related to ethnic discursions. This list included not only post-Soviet ethnonyms, but also other words related to them in one way or another. These words were 1) ethnonyms that denote ethnicity not belonging to the group of post-Soviet ethnicities but taking an active part in the life of the country (Jews, Gypsies), 2) racial slur, 3) words denoting a geographic location (Highlanders, Europeans, Asians), 5) obsolete words with ethnic meaning (Rusich, basurmanin). For each word we generated male and female forms and built a list of bigrams (more than 30 bigrams for each ethnic group) Some words were removed from the list to prevent homonymy (homonyms are the words which sound or spelled in a similar way but have different meanings). Thus, together with relevant bigrams the final list of words comprises over 4,000 units described 97 ethnic categories.

To ensure representativeness we collect all texts related to ethnic discussions from the period of time from January 2014 to December 2015. The text is considered as relevant to ethnic discussions in case it contained at least one word or bigram from the generated list. The texts were gathered using the social media monitoring service IQ-Buzz. IQBuzz monitor pages from thousands of websites looking for predefined words. Typically, such services are used by commercial companies for marketing purposes like to track effect of advertising campaigns or changes in business reputation, but we reoriented IQBuzz to solve research problems. The disadvantage of this solution is the uncertainty about how exactly the data was collected and how complete this sample. IQ-Buzz declares that it able to track “all the mentions in the Internet” but we cannot check this allegation. What about the advantages of data obtained with this service, we would like to note their saturation with additional information. Whenever possible IQBuzz provides the data about the author of the text, such as location, age and gender.

4 Distribution of ethnic texts by regions and domains

The 10 most representative regions can be seen in Table 1. It can be seen that 40% of texts are on ethnicity were generated by the users from two major Russian cities — Moscow and Saint Petersburg. Compared with them, each of the remaining Russian regions takes a much smaller share. This proportion reflects the fact that 1) these cities are the most populated subjects of Russia (the correlation between the share of the population in the region and the proportion of downloaded texts equals 0.8 which means more people live in the region more texts from that region we have), 2) the level of Internet penetration in these cities is higher, than in other regions (“The Development of the Internet in Russian regions,” 2015). However, even taking into account bigger number of Internet users in these cities, the share of these cities cannot be explained by the above factors. I suggest, that a large proportion of reports on ethnic issues from Moscow and Saint Petersburg may indicate an intensification of ethnic processes in these densely populated cities, which is consistent with the theory of urbanists, who pointed to density of population as a source of intense interactions between people in cities and greater complexity of these interactions (Wirth, 1938).

Table 1. Distribution of texts by region (ten most significant regions)

Region	**% of texts from that region in dataset**	**% of population from that region in Russian population**
Moscow	26,60%	8,43%
St. Petersburg	13,40%	3,60%
Perm Region	4,30%	1,79%
Krasnodar region	3,20%	3,79%
Sverdlovsk region	3,00%	2,95%
Rostov region	2,60%	2,88%

Samara Region	2,50%	2,18%
Tatarstan	2,30%	2,65%
Novosibirsk region	2,10%	1,89%
Chelyabinsk region	2,00%	2,39%

As for the distribution of texts by source, it is not surprising that vast majority of the collected texts (82%) are produced by the users of Russia's largest social network site VKontakte.

While interpreting the obtained results, however, it is worth to remember that the source of the revealed differences may be a limitation of the data collection tool, whose authors, while promising the widest possible coverage of sources, cannot guarantee this.

5 Topics of ethnical conversations online and their metrics

We analyzed the topic profile of ethnic discussions, obtained with LDA. Topic modeling has already been successfully used to reveal the characteristics of ethnic discussions (Alexandrov et al., 2016; Apishev, Koltcov, Koltsova, Nikolenko, & Vorontsov, 2016) and researchers from the Laboratory for Internet Studies created a modification of this algorithm called ISLDA, aimed, among other things, to facilitate extraction of ethnically relevant topics from texts (Nikolenko, Koltcov, & Koltsova, 2017).

As mentioned above, we built LDA topic model on 97 topics by the number of selected ethnic categories. To understand the meaning of the topics, we manually labeled each one. The labeling included reading words and texts, related to the given topic with highest probability. Some topics could not be easily interpreted, the interpretation of 22 topic raised difficulties so they were labeled as “uninterpretable”. The presence of such topics is a usual phenomenon in topic modeling. Moreover, as social media have short length and multimodal structure, topic models are often produce uninterpreted topics with this kind of data. The number of tricky topics we obtained can be considered as a satisfactory result.

Topic labeling results show that some topics are about ethnicity while the others don’t touch ethnic issues at all — there are topics about crime, politics, work, cinema, family, economy, navy, army, housing and so on. Comparing this set of topics with topics from random sample of Vkontakte texts, which represent a kind of “natural” topic structure, we see that difference lies in the fact that our set contains more socially significant topics rather than topic about everyday issues (Rykov, Nagornyy, & Koltsova, 2017) like, for instance, games, music, cooking, health and beauty. Users use the social network Vkontakte more like a place where they can save interesting culinary recipe, browse for new films and music and read gardening recommendations. This huge part of the posts is created with the instrumental aim to provide a quick access to potentially useful information. Other texts are automatically generated by numerous applications, most often for advertising purposes. Together with uninterpreted, these topics make up the bulk of the VKontakte. Although a small number of sociopolitical topics were identified in that study (“Christianity” and “Islam,” “Ukraine-Russia relations” and “City events”), their number is significantly less compared to the topics derived from our dataset.

Table 2. Most salient topics of ethnic discussions (25 most salient topics)

ID	Topic	Salience	Polarity
97	Relations of Russians with Other Nations	162170	-0,08
32	Ukrainian-Russian relations	130185	-0,28
44	Russian Society, Russia as a National State	83245	-0,23
27	The revolution in Ukraine	77735	-0,66
48	Conflict relations between Russia and the West	75529	-0,61
15	Children and family	72469	0,17
12	Porn	70656	-1,55
28	Uninterpreted topic	69375	0,88
67	Uninterpreted topic	65616	-0,04
33	The activities of the Russian authorities	62222	0,61
39	Peoples of the North Caucasus, Islam, terrorism	56334	0,23
23	Uninterpreted topic	53387	-0,59
57	Poetry	50895	0,15
6	Economy	48274	0,86
79	Psychology and children	47089	0,14
14	Crimes and murders	46394	-0,22
16	Movies, festivals, performances	43962	-0,13
19	WWII	43618	-0,5
82	Ancient Slavs	41839	-1,15
73	Tourism and rest	41558	0,36
10	Christianity and Orthodoxy	40175	0,81
4	Uninterpreted topic	40032	-0,3
49	Tatars and other Turkic peoples	35559	0,07
63	US-Russian relations, US condemnation	34971	-1,38
37	The Internet	34688	-0,29

With respect to a large number of topics it is clear that, despite their connection with ethnonyms, they cannot be considered as ethnic topics. As an example, there are sport topics (#26 Football, #41 Boxing), politics (#54 and others) and history (№69 and others). What about pure ethnic topics the most vivid examples are topics #29 (Armenian-Turkish relations, genocide of Armenians), #49 (Tatars and other Turkic peoples of Russia), #50 (Uzbeks), #71 (Jews and Judaism), #97 (Relations of Russians with other peoples) and #39 (Peoples of the North Caucasus, Islam, terrorism).

In addition to topics labeling we calculate two topic's metrics: salience and polarity. The first one shows how widely this topic is covered by users, the second indicates most probable topics for texts with positive/negative sentiment scores. To obtain an index of the topic's salience we sum probabilities of all texts in a given topic. It turned out that the most notable topics are topics # 97 (Relationships of Russians with Ukrainians and Caucasians, 6.11%), #32 (Ukrainian-Russian relations), #44 (Russian Society, Russia as a national state, 3.13%), #27 (Ukraine's Revolution of 2013-2014, 2.93%) and #48 (Conflict relations between Russia and the generalized West, 2.84%). From this list it can be seen that the users of Russian-speaking social media discuss events in the east of Ukraine most of all. Since Russia's conflictual relations with the West are often discussed in the context of sanctions and the Ukrainian conflict in general, it can be considered that four of the five most notable topics affect this issue. As for the least represented topics, most of them are difficult to interpret.

The polarity index of the topic was calculated as the sum of products of probabilities of this topic in text and polarity score divided by the overall salience of the topic. To get polarity score for each text we used SentiStrength software with LINIS Crowd sentiment lexicon (Koltcov, Koltsova, & Alexeeva, 2016). SentiStrength ascribes two scores to each text: negative and positive. We used the default version that calculates the two text scores as the maximum of the grades of all occurring words of the respective class (positive or negative). This approach has shown optimal results for short texts like tweets (Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010), that are close in length to posts from social media we seek to analyze. The overall sentiment score of each comment was calculated as the sum of the negative and the positive scores. Since the distribution of the polarity index is close to normal, we calculate z-scores to determine significance of the difference in polarity between the topics. The analysis shows three negative topics, two of them turned out to be about ethnicity (#29 and #59) and the last one is uninterpretable. The first of these two was formed around texts on Uzbek topic. The central ethnicities of the second topic are Armenians, Turks, Azerbaijanis, and the conflicts 1) between Turks and Armenians in the context of the event known as Armenian Genocide[2] and 2) between Armenians and Azerbaijanis over the territories of Nagorno-Karabakh is the central conflict.

The next step was to identify ethnicities, characteristic for topics. In order to know the degree of ethnicity presented in topic, we summarized the probabilities of the topic in all documents in which ethnicity occurs and divided this number by the amount of documents in which this ethnicity occurs and by the sum of the topic probabilities in all documents. The next step was to scale the values obtained on the z-scale. Bringing the results to z-values allows you to make well-founded conclusions about the significance of the expression of an ethnic category in the topic, which, in turn, can be used to test a wide variety of hypotheses.

Let us take, for example, the statements in various works that the media are a powerful source of the spread of racist discourse (Dijk, 1984), especially about the peoples of the North Caucasus (Ahmetieva, 2007; Karpenko, 2002; Kugai & Kovaleva, 2015; Shnirel'man, 2007). Given the topic of crime (#14) among the topics received, we can determine which ethnicities are more often mentioned in the context of communications on this topic, thus testing the hypothesis of bias in the media, especially among the representatives of the peoples of the North Caucasus.

The most probable words for the topic are: “murder”, “time”, “business”, “group”, “detain”, “crime”, “police”, “find”, “killed”, “employee”, “day”, “house”, “court”, “perish”, “district”, “place”, “crime”, “name”, “get”, “happen”, “city”, “dead”, “death”, “police”, “kill”, “February”, “camp”, “prison”, “chief”, “criminal”. It can be concluded that this topic clearly refers to crimes and their consequences. Among these words, however, we do not meet any one that belongs to any ethnic category, which so far does not allow us to conclude that there is bias in the media. Let's take the next step and use

[2] Turkey denies the word “genocide” is an accurate term for these events.

the metric we calculated to see which ethnic categories are significantly more characteristic of the topic than the others.

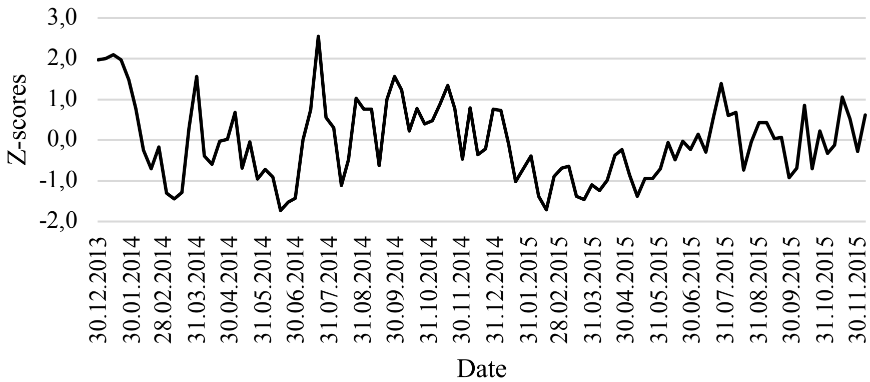
Table 3. Ethnicities in the topic "Crimes and Murders"

Ethnicity	**Z-score**
Ingush	2,0683
Mansi	1,9923
Kavkazec (Peoples of the Caucasus)	1,8843

Z-value equal to 1.8843 corresponds to a significance level of 0.06. So, we can argue, the topic "Crimes and Murders" is not an ethnically neutral topic. It is significantly more probable for texts in which three terms are mentioned: "Ingush", "Mansi" and "Caucasian", two of them — "Ingush" and "Caucasian" — belong to the peoples of the North Caucasus region. These results confirm the conclusions made in the above-mentioned works on the racial bias in the coverage of crimes, and are consistent with the results of the polls, which showed that most people, who believe into connection between crimes and ethnicity, consider peoples of North Caucasus to be responsible for these crimes ("Should we specify the nationality of criminals in mass media?," 2012).

Confirming the existence of a topic that speaks of ethnicities in the context of committed crimes, we trace the dynamics of its representation in the texts of our corps. So, we can identify significant outbursts of the theme in social media, which can show what kind of events excited their users. To do this, we divide all the texts into intervals of a long week and calculate the average probability of the topic in the texts at each of the intervals. Since the values obtained are normally distributed, it is possible to scale them according to the Z-scale.

Fig. 1. Topic "Crimes and Murders"



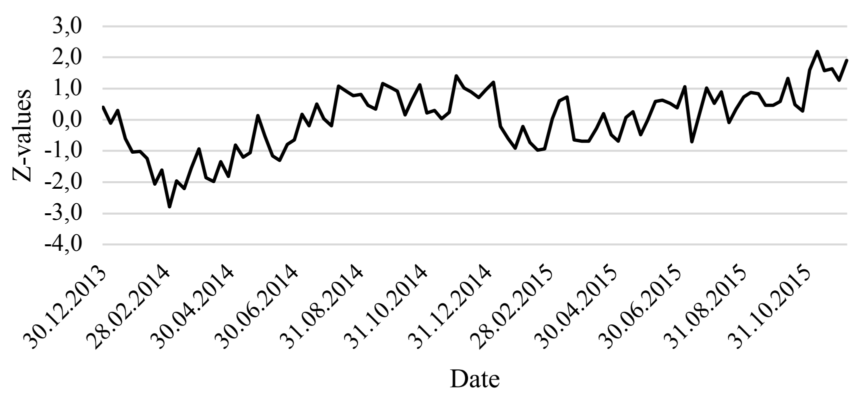
On the resulting graph (Figure 1), no trend can be traced, fluctuations are seen throughout the entire time interval, the interpretation is difficult. There are two intervals when salience of the topic #14 significantly increases. The first is at the beginning of 2014, and the second one is on the 20th of July 2014. Monitoring media during this period of time we have found several events that occurred in the second half of July and could cause a considerable discussion of the topic of crimes. The first event is a catastrophe in the Moscow metropolitan shadow on July 15, the terrorist nature of which, although not confirmed, was actively discussed in social media; the second is the airliner catastrophe over the Donetsk region, also discussed in terms of crime, especially Ukrainian media ("Bandits from DNR shot down the Boeing 777 in the Donetsk region! (conversation)," 2014).

Analogy of the theme "Peoples of the North Caucasus, Islam, terrorism" leads us to similar conclusions about the bias in social media towards the peoples of the North Caucasus. Let's look at the most likely words for her: "Chechnya", "Chechen", "Dagestan", "Muslim", "Kadyrov", "terrorist", "militant", "Caucasian", "Islam", "threatening", "Islamic", "republic", "Allah", "Muslim", "Ramzan", "Mansur", "Eivaz", "mosque", "brother", "war", "Ali", "imam", "IS", "Arab", "Russia", "Makhachkala".

There are three different topics actually mixed in one. The first one tells about the peoples and republics of the North Caucasus, the second is about terrorism and militants, and the third is about Islam. This neighborhood testifies that in Russian-language social media these three phenomena are closely related: both people from the Caucasus and the Islamic religion are perceived in the context of terrorism. In the eyes of the social media audience, terrorism acquires, therefore, ethnic and religious traits. The list of ethnic categories characteristic of the topic (Lezgins, Dargins, Avars, Kumyk, Chechens, Vainakhs, Dagestanians) only confirms these conclusions.

An analysis of topic salience dynamic shows one period, which falls to mid-November 2015, when the salience of the terrorist topic has increased significantly. This splash in the discussion on terrorism is easily explained, although the explanation is in no way connected with the peoples of the North Caucasus: on November 13, 2015, a series of major terrorist attacks took place in Paris, in which more than a hundred people were killed. Such a resonant event caused a noticeable discussion in the media, which was reflected in the graph (Figure 2).

Fig. 2. Topic #39 "Peoples of the North Caucasus, Islam, terrorism"



6 Conclusion

In this work we explore topic structure of ethnic discussions in Russian-language social media. Despite the considerable number of works on ethnicity, since most of them focus their analysis on separate ethnic diasporas it is difficult to find large-scale studies devoted to ethnic discussions in a certain language and establishing relations between all major ethnic groups living on the territory inhabited by native speakers of the given language. Due to the rich data set of 2,659,849 texts from major Russian social media sites each of which contains at least one of the 97 post-Soviet ethnic categories, this research claims to give comprehensive description of the ethnic discussions that was taking place in Russian-speaking social media from 2014 to 2015.

We have found that the topic structure of the ethnic discussions differs markedly from the topic structure of a typical social media site, which is the social network site Vkontakte. Texts in which ethnonyms are present contain more topics related to actual social and political issues, while topics of everyday life activities predominate in the "natural" topic profile built on randomly selected texts from Vkontakte social networking site. The revealed difference indicates that ethnic discussions is a problematic field, on which there is acute and often conflictual communication. It has the potential to grow into activities for constructing social problems as Spector and Kitsuse said (Spector & Kitsuse, 1977).

In the topic structure of social media texts, we also identified the most salient and most negative topics. It is noteworthy that the largest share in the ethnic discussions is occupied by topics about Ukraine-Russia relations over the recent conflict between two countries. In this case, it is difficult to separate ethnic topics from political ones since the conflict between states is projected into the sphere of inter-ethnic relations. As for the significantly negative topics, we managed to find two negative and interpreted topics, the first is formed around Uzbek ethnicity and the second is about the Turkish-Armenian relations in the contexts of Armenian Genocide.

Acknowledgements. Withdrawn for anonymous reasons.

References

1. Ahmetieva, V. (2007). People with dog heads: the image of Chechens in the Russian media. In A. Verkhovitsky (Ed.), *The lhate speech against society* (p. 259). Moscow: Center "Sova". (in Russian)
2. Alexandrov, D., Gorgadze, A., & Musabirov, I. (2016). Virtual Caucasus on VK Social Networking Site. In *Proceedings of the 8th ACM Conference on Web Science* (pp. 215–217). New York, NY, USA: ACM. https://doi.org/10.1145/2908131.2908205
3. Apishev, M., Koltcov, S., Koltsova, O., Nikolenko, S., & Vorontsov, K. (2016). Mining Ethnic Content Online with Additively Regularized Topic Models. *Computacion Y Sistemas*, *20*(3), 387–403. https://doi.org/10.13053/CyS-20-3-2473
4. Bandits from DNR shot down the Boeing 777 in the Donetsk region! (conversation). (2014, July 17). [https://www.youtube.com]. Retrieved October 5, 2017, from https://www.youtube.com/watch?v=-LyD6FgHE8I (in Russian)

5. Bessudnov, A. (2016). Ethnic Hierarchy and Public Attitudes towards Immigrants in Russia. *European Sociological Review*, jcw002. https://doi.org/10.1093/esr/jcw002
6. Bodrunova, S., Koltsova, O., Koltcov, S., & Nikolenko, S. (2016). Are "Migrants" All the Same? Mapping Attitudes to the Resettlers From Post-Soviet South in the Russian Blogosphere. *The European Consortium for Political Research*. Retrieved from https://ecpr.eu/Events/PaperDetails.aspx?PaperID=30392&EventID=95
7. Chan, J., Ghose, A., & Seamans, R. (2016). The internet and racial hate crime: Offline spillovers from online access. *MIS Quarterly: Management Information Systems*, *40*(2), 381–403.
8. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *ArXiv:1703.04009 [Cs]*. Retrieved from http://arxiv.org/abs/1703.04009
9. Dijk, T. A. van. (1984). *Prejudice in Discourse: An analysis of ethnic prejudice in cognition and conversation*. John Benjamins Publishing.
10. Fadeev, P. (2017). Ethnic groups of St. Petersburg in the media. *The World of Russia*, (1). (in Russian)
11. Foxall, A. (2014). *Ethnic Relations in Post-Soviet Russia: Russians and Non-Russians in the North Caucasus* (1st ed.). New York: Routledge.
12. Gibson, S., & Lando, A. L. (2001). *Impact of Communication and the Media on Ethnic Conflict*. IGI Global. Retrieved from https://www.igi-global.com/book/impact-communication-media-ethnic-conflict/139301
13. Grasmuck, S., Martin, J., & Zhao, S. (2009). Ethno-Racial Identity Displays on Facebook. *Journal of Computer-Mediated Communication*, *15*(1), 158–188. https://doi.org/10.1111/j.1083-6101.2009.01498.x
14. Karpenko, O. (2002). Language games with "guests from the south": "Caucasians" in the Russian demo-press 1997-1999. In *Multiculturalism and Transformation of Post-Soviet Societies* (pp. 183–188). Moscow. (in Russian)
15. Koltcov, S. N., Koltsova, O. Y., & Alexeeva, S. B. (2016). An Opinion Word Lexicon and a Training Dataset for Russian Sentiment Analysis of Social Media. In *Computational Linguistics and Intellectual Technologies* (pp. 277–287). Moscow.
16. Kugai, A., & Kovaleva, A. (2015). Racist discourse of electronic media as a factor in constructing interethnic conflicts. *Management Consulting*, (1 (73)). (in Russian)
17. Nikolenko, S. I., Koltcov, S., & Koltsova, O. (2017). Topic modelling for qualitative studies. *Journal of Information Science*, *43*(1), 88–102. https://doi.org/10.1177/0165551515617393
18. Rykov, Y., Nagornyy, O., & Koltsova, O. (2017). Digital Inequality in Russia through the Use of Social Network Site: A Cross-Regional Comparison. Presented at the Digital Transformations & Global Society 2017, Saint-Petersburg.
19. Shnirel'man, V. (2007). Media, "ethnic crime" and migrantophobia. In A. Verkhovitsky (Ed.), *The hate speech against society* (p. 259). Moscow: Center "Sova." (in Russian)
20. Should we specify the nationality of criminals in mass media? (2012, August 10). Retrieved May 12, 2017, from http://fom.ru/Bezopasnost-i-pravo/10586
21. Spector, M., & Kitsuse, J. (1977). *Constructing Social Problems*. Menlo Park: CA: Cummings.
22. Suleymanova, D. (2009). Tatar Groups in Vkontakte: The Interplay between Ethnic and Virtual Identities on Social Networking Sites. *Digital Icons: Studies in Russian, Eurasian and Central European New Media*, *1*(2), 37–55.
23. The Development of the Internet in Russian regions. (2015). Retrieved May 10, 2017, from https://yandex.ru/company/researches/2015/ya_internet_regions_2015

24. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment in Short Strength Detection Informal Text. *J. Am. Soc. Inf. Sci. Technol.*, *61*(12), 2544–2558. https://doi.org/10.1002/asi.v61:12
25. Wirth, L. (1938). Urbanism as a Way of Life. *American Journal of Sociology*, *44*(1), 1–24.
26. Xiang, G., Fan, B., Wang, L., Hong, J., & Rose, C. (2012). Detecting Offensive Tweets via Topical Feature Discovery over a Large Scale Twitter Corpus. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (pp. 1980–1984). New York, NY, USA: ACM. https://doi.org/10.1145/2396761.2398556