Detecting Interethnic Relations with the Data from Social Media

Olessia Koltsova[1(✉)], Sergey Nikolenko[1,2], Svetlana Alexeeva[1,3], Oleg Nagornyy[1], and Sergei Koltcov[1]

[1] National Research University Higher School of Economics, Moscow, Russia
{ekoltsova, snikolenko, salexeeva, onagorny, skoltsov}@hse.ru
[2] Steklov Mathematical Institute, St. Petersburg, Russia
[3] St. Petersburg State University, St. Petersburg, Russia

Abstract. The ability of social media to rapidly disseminate judgements on ethnicity and to influence offline ethnic relations creates demand for the methods of automatic monitoring of ethnicity related online content. In this study we seek to measure the overall volume of ethnicity related discussion in the Russian language social media and to develop an approach that would automatically detect various aspects of attitudes to those ethnic groups. We develop a comprehensive list of ethnonyms and related bigrams that embrace 97 Post-Soviet ethnic groups and obtain all messages containing one of those words from a two-year period from all Russian language social media (N = 2,660,222 texts). We hand-code 7,181 messages where rare ethnicities are overrepresented and train a number of classifiers to recognize different aspects of authors' attitudes and other text features. After calculating a number of standard quality metrics, we find that we reach good quality in detecting intergroup conflict, positive intergroup contact, and overall negative and positive sentiment. Relevance to the topic of ethnicity and general attitude to an ethnic group are least well predicted, while some aspects such as calls for violence against an ethnic group are not sufficiently present in the data to be predicted.

Keywords: Interethnic relations · Ethnic attitudes · Mapping · Social media · Classification · Lexicon

1 Introduction

Social media have become a space where individuals and groups can both cooperate and engage in conflict being highly visible. In particular, judgments about ethnic groups or ethnicity issues in general may be of different valence (as it happens offline), but their dissemination can potentially proceed faster and reach wider audiences than before the internet era. The consequences and thus the importance of this large-scale public visibility of ethnicity related attitudes are still to be fully understood. While the effect of positive online interethnic communication on offline conflict reduction finds a limited proof, there is a large evidence of the impact of negative online content on offline interethnic conflict [15] and hate crime [9]. This creates demand for the methods

D.A. Alexandrov et al. (Eds.): DTGS 2017, CCIS 745, pp. 16–30, 2017.
https://doi.org/10.1007/978-3-319-69784-0_2

of monitoring of ethnicity related online content, in particular for instruments of its automatic mining from large data collections [8].

In this context, Russia, a country with a large number of both "home born" and migrant ethnic groups, has received relatively little attention from researchers [5, 6, 17, 18]. In this paper we seek to measure the overall volume of ethnicity related discussion in the Russian language social media, to compare public attention to different ethnic groups, and to develop an approach that would automatically detect various aspects of attitudes to those ethnic groups.

2 Related Work

Literature on ethnicity and the Internet comes from a large number of disciplines and seems vast, however, there are surprisingly few works that directly address the issue studied here. Most works are devoted to the formation of ethnic identity and boundaries [17, 19, 27, 29], its relation to online intergroup contact or communication [36], as well as ethnic online communities [18, 24] and influence of ethnicity on online connectedness [21] or online access. That said, such studies are most often centered on minority-generated discourse about respective ingroups rather than on their perceptions by the outgroups. There is very little literature on representation of non-dominant ethnic groups on the Internet, notably in the user-generated content [31], although it is a well developed topic in traditional media studies which examine both representations per se [41], and perceptions of those representations by the minorities themselves [33], as well as the relation between representations and public attitudes [34]. Another feature of this stream of studies is that ethnic minorities are usually migrant/diasporic groups [29], or, if they are "indigenous", they are often conceptualized rather as racial than ethnic groups [19]. Most of such research is done either from European or American perspectives [10]. In Europe, for some reason, most research is centered around migrants although usually only migrants of "other" ethnicity are meant. Surprisingly, indigenous ethnic minorities like Catalan in Spain or Irish within the UK and their relation to the Internet have been largely ignored. In the US, on the contrary, all groups except Native Americans can be seen as migrant, but the most problematic division seems to be between races [19] or groups perceived as races rather than as ethnicities [31].

Neither of these approaches fully applies to Russia that comprises a mixture of both migrant and indigenous ethnic groups who clash dependently of their race. Both Russian majority and some other groups may be confused about who is a migrant and who is a local on each given territory, and these notions are constantly being challenged in the public space. Thus, multidirectional interethnic interactions take place in the Russian language internet, and not all of them bare the signs of conflict.

Ethnic communication and conflict online has received its separate attention from researchers. Some studies have sought to compare offline and online interethnic communication [25, 36] and to understand motivations for online interethnic interaction [11]. As mentioned above, quite a number of works attempts to find relation

between online interethnic communication and offline conflict prevention [15] or dissemination [9], as well as offline minority mobilization [29]. This research usually does not map, monitor or systematize online attitudes towards various ethnic groups.

When it comes to methods of such mapping, an absolutely separate body of literature exists. This research mostly comes not from social science, but from computer science and computational linguistics, and it overwhelmingly English language oriented, with few exceptions [35]. It is mostly aimed at automatic detecting of hate speech in user-generated content [3], not always specific to the ethnicity issue, while the "positive" side of ethnic representations online misses researchers' attention at all. Hate speech is broadly understood as hostility based on features attributed to a group as a whole, e.g. based on race, ethnicity, religion, gender and similar features.

This research is very different in breadth and scope: some studies seek to perform race- or ethnicity-specific tasks, for instance aim to detect hate speech against Blacks only [23]. Others attempt to capture broader types of hate speech, e.g. related to race, ethnicity/nationality and religion simultaneously [8, 16], or even generalized hate speech [38] and abusive language [28]. Most studies acknowledge that hate speech is domain specific although some features may be shared by all types of hate speech, therefore some try to catalogue common targets of hate speech online [30].

In such works, a large variety of techniques is being offered and developed, including lexicon-based approaches [16], classical classification algorithms [35] and a large number of extensions for quality improvement, such as learning distributed low-dimensional representations of texts [12], using extralinguistic features of texts [40] and others. Some draw attention to the role of human annotators and the procedure of annotations for classification results [2, 39].

This latter topic leads to the problem of definition of hate speech needed to help annotators understand their job. Computer science papers seldom or never address this problem relying on human judgement as the ultimate truth, and when they do address it they mostly focus on making annotators capture the existing definitions of hate speech, not on critically assessing them or developing new ones. Meanwhile, most existing definitions we know are ethically non-neutral which makes them a difficult object for automatic detection. From the overviews we learn that hate speech, or harmful speech is usually defined via such attributes as "bias-motivated", "hostile" "malicious", dangerous" [14], "unwanted", "intimidating", "frightening" [13], which can be summarized as… actually, bad. All the mentioned attributes, as well as the titles of the concepts themselves mark the concepts they seek to define as ethically unacceptable. If so, to correctly detect them, human annotators have to share common values with the researchers, otherwise they would not be able to recognize hate speech in texts. Since not every derogation, disapproval or condemnation is ethically unacceptable (e.g. condemnation of genocide is considered absolutely desirable), language features of disapproval or derogation per se do not necessarily point at what the Western liberal discourse usually means by hate speech, and this makes it especially elusive when applied beyond the Western world. Further below, we elaborate on this problem by offering a number of relatively objective questions that can be answered by annotators of any ethical or political views. We also show how not only negative aspects of ethnicity related speech can be captured.

3 A Note on Ethnicity in Post-Soviet Space

Not only interpretation of results, but even mere cataloguing of Post-Soviet ethnic groups turns to be impossible without prior sociological knowledge. Here we give a brief introduction into the field to make further understanding of our work easier.

Ethnic landscape of Russia is a patchwork inherited from the Soviet Union and earlier Russian Empire that tried to grant each "nationality" a certain degree of autonomy depending on its bargaining capacity. As a result of dissolution of the USSR, Russia has found itself surrounded by 14 countries that under the Soviet rule had been granted the status of "first-order" republics with the right of secession. They include such diverse cultures as the Baltic that are now a part of the EU (Protestant, German or Uralic language group), predominantly Orthodox Slavic Ukraine and Belorussia and Romanian Moldova, as well as Southern Caucasus that includes Christian Georgia and Armenia and Muslim Azerbaijan. The latter together with five Central Asian countries, also Muslim, speaks a language of Turkic group. However, Central Asian countries range from highly secularized predominantly Russian speaking and economically developed Kazakhstan to the underdeveloped Tajikistan whose economy has been subverted by a number of armed conflicts. Former Soviet Union (FSU) countries serve as the major immigration sources for Russia.

Inside Russia the diversity is even higher. Of its 85 administrative units, 22 contain a titular ethnicity in their titles and are termed republics; a few more ethnic-titled units are included inside republics. All the rest, including the "core" Russian provinces, are termed regions and are named after their capital cities or other geographical names. Some ethnic groups have no definite territory, while others that a virtually extinct, do, and one tiny ethnicity has even got two titular regions. Siberia and the European North are populated by small ethnicities of Uralic and some other language families most of whom are close to extinction. Prior to the conquest by the Russian Empire most of them practiced shamanism and hunting-gathering lifestyles. However, many Turkic groups mainly from European Russia, but also from Siberia are much more alive, with Tatars, Bashkirs and Chuvashs being three largest ethnic groups in the country after Russians (and along with Ukrainians). Most of such groups had passed to Islam and sedentary economy prior to the Russian conquest. Tatarstan had even had its own strong statehood, and now it presents a unique combination of strong ethnic identity, industrial development and the ability to integrate. Northern Caucasus, on the contrary, is often described as the powder keg of Russia. Heavy battles have been fought here both between Russian/Soviet forces and the locals and among the locals themselves who had never had any strong tradition of their own statehood. In the 19th century the region became predominantly Muslim which only complicated the situation. Up to date, Northern Caucasian groups have most strongly resisted assimilation, with Chechnya being the leader and the symbol of this strategy (and with Chechens being the next largest group after those aforementioned). Some North Caucasus republics stay beyond the Russian legal and cultural order and at the same time are heavily subsidized. Inhabitants of "core" Russian regions often do not differentiate between Northern and Southern Caucasian ethnicities.

4 Data, Sampling and Lexicon Development

In this study we seek to monitor public representations of ethnic groups available to all or most consumers of user-generated content (UGC) in Russia, and therefore we limit our research to the Russian language texts. This introduces some asymmetry: while Russians in this situation would be mostly speaking to themselves, ethnic minorities would be mostly speaking to the outgroups. However, as political, cultural and "numeric" positions of the dominant nation and of the minorities are fundamentally asymmetric, this only reflects the real communicative situation. In our previous research [1, 5, 26] we adopted a few strategies that we have to abandon here. First, earlier we had sampled either the most popular or random producers of UGC, and then searched for ethnicity related content with topic modeling – an approach close to fuzzy clustering [4]. We found out that it is not optimal for detecting relevant texts as the topic of ethnicity is too rare for the algorithm to work properly. Second, for semi-supervised topic modeling and frequency counts, we had used the most complete list of world ethnic groups and nations. We did not exclude any or differentiate between ethnicities and nations since the boundary between them is thin. What we found out was much more related not to the in-Russia or FSU ethnicities, but to the nations that had major global or regional political influence, first of all – Americans, Germans, Ukrainians and Jews, but also to many European nations. The texts found with our approach in fact were devoted much more to international relations than to ethnicity. We then came to define the texts devoted to the topic of ethnicity as:

1. texts where the major actors were private persons of a given ethnicity or ethnic groups, and not states or their official representatives (e.g. "Russians blocked a UN resolution" is not about ethnicity, while "Russians made a fight at a football match" is);
2. ethnicity is important for the outcomes or is used as an explanation (e.g. "The cook was Tadjik, he knew everything about pilav" is about ethnicity; "We've just returned from an exhibition of a Tadjik painter, let's have lunch" is not).

Based on described above experience, we have come to the conclusion that we, first, should limit our research to FSU ethnicities in order to avoid global politics. Although this does not guarantee avoiding international relations within FSU, but it strongly mitigates this risk because FSU ethnicities are very often present in Russia as private persons or groups, unlike "far-abroad" nations. Second, we have concluded that we should preselect ethnicity related texts using keyword lists. The major goal at this stage able to influence the quality of the subsequent research was developing a comprehensive list of relevant keywords. Keyword search, compared to fuzzy clustering with subsequent labeling, has an obvious limitation: it cannot yield texts that discuss ethnicity without mentioning key words. That is, keyword search can give high precision, but low recall. However, when fuzzy clustering or topic modeling do not work, we consider keyword search as a first step towards elaborating a more sophisticated supervised classification approach that we address further below.

Luckily, in the Russian language all ethnic groups (except Russians) have different words for nouns and for adjectives (e.g. "Turk" vs "Turkish"). By taking nouns only

(and generating their female and plural forms), we have been able to increase the relevance of the texts being found. We have also automatically generated bigrams consisting of an ethnic adjective and a noun referring to a person or the nation (e.g. Tatar woman, Chechen people). The most difficult task was to limit our lexicon of ethnonyms – nouns referring to representatives of ethnic groups. Our list includes the following categories:

- Names of all ethnic groups indigenous to the Post-Soviet countries. Some groups are hierarchical: e.g. Mordva includes Moksha and Erzya; some are synonymous, e.g. Alans and Ossetians. Here we mostly used the data from the Russian Census 2010 and the lists of world nations from UN and other international organizations.
- Names of some other significant minorities. Here we had to exclude those groups that were too visible internationally and most often led to the texts on international relations or politics within respective countries (e.g. Germans). We succeeded in including Jews as we had found that while talking about respective international politic another word (Israeli) would be most often used. We included Gypsies as the word "Roma" is virtually unknown in Russian. Here we relied on the data about quantities of different ethnic groups in Russia from its Census 2010.
- Ethnophaulisms: abusive words denoting representatives of ethnic groups. Here we used literature [20, 22, 32, 37] and the lists of top words in ethnicity related topics found in prior topic modeling. A dozen FSU ethnic groups has precise pejorative equivalents (e.g. Jid for Jew); other ethnophaulisms have more blurred meanings: Khach for any Caucasian, or Churka for any Central Asian, but sometimes for any "Eastern" non-Russian. This ambiguity was the major reason why we chose to treat ethnophaulisms and their derivatives as "separate ethnic groups".
- Meta-Ethnonyms. Certain words referring to language groups (Slav, Ugr) or to the regional identity often function as ethnonyms in the Russian language. Some of them are emotionally neutral (Baltic, European, Siberian), but others can sometimes obtain negative connotations, depending on the context (Asian, Central Asian, Caucasian, Highlander). Note that Caucasian in our context means merely a representative of ethnic groups from the Caucasus, not White race in general. Dagestani also belongs to meta-ethnonyms although most Russians do not know that this a regional, not an ethnic name. Lists of top words in ethnicity related topics found in prior topic modeling were used here.
- Cossacks. These were actually a social group in the Imperial Russia with its distinct culture, but no specific language or religion (like Samurai in Japan). Cossacks were free armed peasants who once leaved at Russia's Southern and Eastern borders and were to resist the first blows of steppe nomads. They spoke either Russian or Ukrainian, depending on the region they lived in. As the Russian Empire grew, they lost their "jobs" and benefits and were finally exterminated under the Soviet rule. However, the recent recontructionist movement of Cossacks has demanded to proclaim them an ethnic group. Their half-militarized semilegal groups have been playing an important role in conflicts with ethnic minorities in the Southern Russia and have been often supported by the local authorities.
- Russians. As noted above, "Russian" in the Russian language has no noun. Adjective "Russian" leads to anything but ethnicity topic (e.g. Russian language

school tests and textbooks). Therefore, we included bigrams only, as well as various synonymous nouns (Ross, Rossiyanin etc.).

After forming this list, we checked if the listed items occurred in our Random Sample of user-generated posts. This sample included 74,303 users of VKontakte, the most popular Russian social networking site akin to Facebook. The users had been selected randomly from each of 85 Russian regions, after which we downloaded all messages from each account for all time and obtained 9,168,353 posts and 933,516 comments to them. This sample is large enough to judge which ethnic words are uncommon for the Russian language UGC, and these happen to be only very exotic ethnicities. We thus obtained a list of 4,063 words and bigrams covering 115 ethnic groups.

We then submitted this list to a commercial aggregator of the Russian language social media content and downloaded all messages containing at least on keyword for the period from January 2014 to December 2015. We made small samples from each ethnic group for hand-checking and found out that some ethnonyms had much more frequently occurring homonyms. We had to exclude these words; however, their synonyms stayed in the sample – e.g. "Mariy" that usually led to texts with the female name "Maria" was excluded, but a russified word "Mariets" stayed. After all cleaning we obtained 2,660,222 texts about 97 ethnic groups; we further refer to this collection as Keyword Sample.

5 Ethnic Groups Descriptive Statistics: Frequencies and Co-occurrence

Given that Russian language social media produce several million messages daily, three million messages in a two-year period is a tiny fraction of the entire volume which clearly shows a low interest of the general public in the topic of ethnicity. Mean length of messages with ethnonyms is much higher than that of the VKontakte random sample (332 words compared to 16.7) and 54.6% of texts contain more than 100 words. This suggests that while the vast majority of messages in social media are everyday small talk, texts related to ethnicity are often elaborated discussion pieces. It makes them much more suitable for various machine learning techniques that would fail on random texts due to their shortness. Ten most frequent ethnic groups include Russians, Ukrainians, Jews, Slavs, Asians, Europeans, as well as two largest Muslim minorities in Russia – Tatars and Chechens. The distribution of frequencies is power law, which is not very good for classification tasks if the original proportion is kept. However, we find substantial regional differences. As mentioned above, some regions in Russia are national republics named after their "titular" ethnic groups; when ranked by the number of mentions in respective regions, such ethnic groups on average gain 45 positions compared to their positions in the general frequency list. Almost a half of them finds themselves in top three most frequent groups in the respective region. Smaller groups gain more positions ($r = -0.5$), although in the entire collections larger groups are mentioned more often. This means that discussions about some ethnic groups being hardly noticeable at the national level are in fact quite important at the regional level.

In total, 45% of messages contain more than one ethnic group, with maximum being 60. This may mean that ethnic groups get opposed in the same text which should lead it to have different sentiment patterns. Such mixture would inevitably complicate automatic sentiment analysis, however, interethnic communication, both "positive" and "negative" can be detected only in texts with several ethnic groups. Also, attitudes to ethnic groups in multiethnic texts might differ from those in mono-ethnic messages in an unpredictable way. We thus can not exclude multiethnic texts from the research. We then examined the cooccurrence patterns to decide whether we should sample multi- and mono-ethnic texts separately, or sample clusters of most commonly cooccuring ethnic groups.

We obtain cooccurrence matrix for all 97 ethnicities and calculate a number of distance metrics (cosine similarity, chi square and distributional). We find a large variation in the ethnicities' overall proximity to others – that is, some of them are predominantly mentioned with others while others aren't. We use several algorithms of community detection, that give similar results; Fig. 1 reports the most sound solution based on chi square distance and Infomap algorithm [7]. Nodes depicted as circles with different motives represent ethnic groups, and their sizes reflect the absolute frequencies of those groups in the collection. Distances and edge thicknesses approximate the level of dissimilarity between ethnic groups, while motives of circles denote clusters that were found by the algorithm.

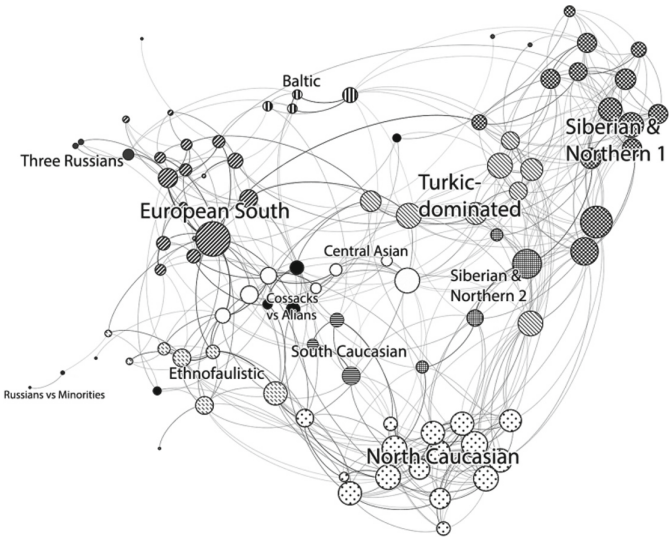

Fig. 1. Cooccurrence of ethnic groups in user posts.

We see that many clusters are formed based on regional and cultural similarity: the Baltic, North and Southern Caucasian, Central Asian (white cluster in the center spread horizontally). Two similar clusters of Northern and Siberian ethnicities and a cluster where in-Russia Turkic peoples prevail are in fact linked by quite a number of non-Turkic ethnicities that were assigned to the "Turkic" cluster. We thus see that

in-Russia indigenous ethnicities form a metacluster that is distinct from Central Asia and is very far from the Ethnophaulistic cluster. The latter includes pejoratives for Asian and Caucasian ethnic groups and is closely related to the respective clusters.

Along with Ethnophaulistic cluster, the largest grouping termed "European South" is also not unproblematic. It includes Russians and Ukrainians with their synonyms, including pejoratives, as well as Jews and Gipsies. The latter two groups historically have been connected to the European South, especially Ukraine contributing to its diversity and social cleavages. Russians and Ukrainians have been currently reconnected via a sharp conflict. There are also small clusters of Russians vs non-Russians. To summarize, ethnicities most probably form both problematic and non-problematic clusters, and we can not guarantee capturing conflict by sampling clusters.

We also see that clusters are quite interconnected, and some ethnic groups (e.g. Tatar, Buryat, Turkmen, Alan) demonstrate high betweenness. It is thus hard to sample from clusters both because of their interconnectedness and because they are of very different size; however, from both network and frequency analysis we see that we need small ethnicities to be overrepresented as their infrequency does not mean their unimportance. In fact, two clusters of small Northern and Siberian peoples together are comparable in size with the Central Asian cluster that consists of few large ethnicities. Also, since the distribution of mention frequencies is close to power law training classifiers on proportional sample would teach them to recognize texts related predominantly to the largest ethnic groups.

6 Hand-Coding: Sample and Procedure

We sampled 97 ethnic groups from the Keyword Sample and added the rest 18 from the Random Sample. Each ethnic group was represented by 75 texts except those that were fewer. The final collection for human assessment (Coding Sample) comprised 7,181 texts. We recruited and trained student assessors who first performed training coding that was checked a supervisor. After that unreliable assessors were excluded, and the rest we given more instruction. Each text finally was assessed by three persons who worked independently. Assessors were asked to answer the following questions: (a) is the text understandable? (yes/no/other language); (b) does it contain one, several or no ethnonyms?; (c) how strongly a general negative sentiment is expressed in it, if any? (no/weak/strong); (d) how strongly a general positive sentiment is expressed in it, if any? (no/weak/strong); (e) does the text mention interethnic conflict? (yes/no/unclear); (f) does the text mention positive interethnic interaction? (yes/no/unclear); (g) does the text contain the topic of ethnicity? (yes/no); (h) same question for ten more topics, including "other" and "unclear" (yes/no). Further, for each ethnonym mentioned the following questions were asked: (i) does the author refer to a concrete representative of an ethnic group or to the group as a whole? (former/latter/unclear); (j) what is the overall author's attitude to this group/person? (negative/neutral/positive); (k) does the author belong to the ethnic group s/he is writing about? (yes/no/not mentioned); (l) does the author call for violence against the mentioned ethnic group/person? (no/openly/latently); (m) is the ethnic group or person portrayed as superior or inferior compared to others? (former/latter/unclear/irrelevant); (n) is the ethnic group or person

portrayed as a victim or an aggressor in interethnic relations? (yes/no/unclear/irrelevant); (o) is the ethnic group or person portrayed as dangerous? (yes/no/unclear).

This list reflects our approach to capturing attitudes to ethnicity in UGC. Hate speech as ethnically non-neutral concept was never applied. Neither in written instructions nor during training sessions assessors were told whether it is ethically acceptable to portray ethnic groups as superior or inferior, call for violence against them or engage in conflict with them. The assessors could hold any positions on this matter and be of any ethnic group themselves; their task was to determine if a given feeling or position was present in the text. Moreover, they were explicitly told that making their own ethical judgement about text authors or text characters was not their job.

For this codesheet, a web interface was developed that works both from stationary computers and mobile devices. It has several questions per screen and the checkboxes with the most frequent answer checked by default. This makes the work of assessors much faster and minimizes errors. The interface also allows deleting the answers would an assessor find a mistake after submitting the work.

The results of coding present a "vortex". If an assessor found the text uninterpretable, no other questions were shown. We obtained 4,947 texts that were unanimously considered understandable, and 6,719 texts understood by at least one assessor (of them the vast majority was understood by two persons). Further, there were 6,383 texts in which at least one assessors found an ethnonyms. It means that the rest contained homonyms (e.g. reference to actor Alan Rickman instead of Alans). In particular, among mentions of eleven Siberian/Northern ethnic groups as much as 48-97% in fact were found to contain no ethnonym. For 86% of ethnicities, however, this rate of false-positives was below one-third. As a result, we obtained 4,121 texts that were found both understandable and ethnonym-containing unanimously. Only these texts got answers to questions c-f. Finally, of them only 2291 were unanimously considered to contain the topic of ethnicity. It means that the rest only mentioned an ethnonym – (e.g. "Ukrainian Ivan Petrenko won the gold medal in sprinting"). Therefore, only 2,291 texts got three answers on questions i-o.

It thus has been virtually impossible to work only with the most reliable data – that is, the texts that got three independent grades for each question. Although their reliability might have improved the quality of classification, their small quantity would have played against it. We therefore chose to work with texts that had received at least one grade. Even then, for some questions the "yes" answers were too rare (e.g. call for violence). Finally, we trained classifiers only for questions c, d, j (negative and positive sentiment and author's attitude); e, f (conflict and positive interaction), and g (presence of ethnicity as a topic). The latter was made to improve selection of relevant texts compared to simple keyword search that, as mentioned in the beginning, gave a relatively good precision but supposedly poor recall.

7 Classification and Results

The experiment procedure was the following. At first, for feature selection, a lexicon of bigrams was trained on the Keyword Sample with non-lemmatized texts with Gensim phrases function. The final lexicon included only words and bigrams that occurred not

less than 5 times in the collection (7,307 unigrams and 6,514 bigrams). After that we experimented with extracting words and bigrams from texts lemmatized with pymorpthy2. The final lexicon included 6,364 words and 7,039 bigrams.

The hand-coded texts were also lemmatized and transformed into the vector form using both absolute word frequencies and their tf-idf weights. As the target variables were the mean assessors' scores, they were often non-integer; therefore, they were either binarized or trinarized depending on the number of initial values that the corresponding variable could take. The thresholds for this procedure are given in Table 1.

Table 1. Quality of automatic classification of users' texts on ethnicity

Does the text contain:	Texts	Binarization/ trinarization	Avg precision	Avg recall	Avg F1	Avg accuracy & variance	Gain over base., target class, %
General negative sentiment	6,674	<0.3 = 0; ≥0.3 = 1	0.75	0.75	0.75	74.67 ± 1.50	20
General positive sentiment	6,688	<0.3 = 0; ≥0.3 = 1	0.71	0.66	0.68	75.1 ± 1.69	21
General attitude to an ethnic group	5,970	<1.3 = 0; [1.3; 2.35] = 1; >2.35 = 2	0.55	0.47	0.49	66.54 ± 1.74	21; 7
Interethnic conflict	6,701	<0.3 = 0; ≥0.3 = 1	0.72	0.71	0.71	75.22 ± 1.60	26
Positive interethnic interaction	6,711	<0.3 = 0; ≥0 = 10.80	0.71	0.61	0.63	82.80 ± 1.58	18
Topic of ethnicity	5,970	<0.8 = 0; ≥0.8 = 1	0.67	0.66	0.66	66.81 ± 1.82	16

The sample was 100 times randomly divided into a training set (90%) and a test set (10%). The classifier (logistic regression with scikit-learn library) was trained on the training sets and tested on the test sets. The results were averaged over all 100 runs.

The following quality metrics were calculated: (a) precision: the share of texts correctly assigned to a given class among all texts assigned to this class; (b) avg precision: mean of all values of precision over all classes; (c) recall: the share of texts assigned to a given class among all texts assigned to this class by assessors; (d) avg recall: mean of all values of recall over all classes; (e) F1-score (a variant of F-measure): F1 = 2 * (precision * recall)/(precision + recall); (f) avg F1: mean of all values of F1over all classes; (g) accuracy: the share of correctly predicted cases, %, that is the share of texts that were assigned to the same class as decided by the assessors; (h) avg accuracy & variance – mean of all values of accuracy over all classes and its variance; (i) gain over baseline for the target class, the baseline being the probability of assigning an item from a class of interest to its true class by a random classifier that keeps the true class proportion. That is, for each target class this probability equals its

share in the collection. The target class was each time the one that possessed the feature (e.g. texts that contain negative sentiment as opposed to those that do not). For the three-class task r there were two target classes: text with positive attitude and with negative attitude to ethnic groups as opposed to the neutral.

These metrics were calculated for classifiers trained on: lemmatized and non-lemmatized texts; texts vectorized with absolute word frequencies and tf-idf weights; with a word-only lexicon and with a lexicon that includes both words and bigrams. The best results were obtained on lemmatized texts with tf-idf weights and bigrams, and only they are reported below (see Table 1).

We find that general positive and negative sentiments, as well as interethnic conflict are reasonably well predicted. Positive sentiment detection, however, gains much more in terms of precision than recall compared to the baseline, while the prediction of the other two variables is more balanced. This is a common problem with the positive "end" of the sentiment spectrum: lexical features for it are much scanter than for the negative "end", and positive sentiment is much more often expressed indirectly. We can see the same tendency with detecting positive interethnic contact as compared to conflict. It, too, gains mostly in precision rather than in recall.

Prediction of the ethnicity topic yields modest results for a similar reason: the predicted feature is quite vague and hard to define. In fact, this was the question that aroused the largest share of doubts and disagreement among assessors. Finally, predicting attitude to an ethnic group faces the largest problems. We can see that while negative attitude get predicted fairly well with 20% gain over the baseline, detection of positive attitude gains 16% in precision and zero in recall. This most probably happens because nearly half of the texts contain more than one ethnic group. There are, however, several important arguments against excluding those texts for attitude prediction. First, we do not know how attitude formation in multiethnic texts substantially differs from that in mono-ethnic one: it may happen that different speech features are used in those two types of texts. Second, even if it is not the case, a classifier trained on "pure" mono-ethnic cases would inevitably "average" attitude scores for different ethnicities if they get different attitudes in the same text.

8 Conclusion and Future Work

In this work we find that although ethnicity related hate speech online is an important concern both for the public and the policy makers, ethnicity relevant discourse constitutes a tiny fraction of user-generated content. This makes a task of retrieving relevant content similar to finding a needle in a haystack. Furthermore, attitudes expressed in such short informal texts are, unsurprisingly, not easy to predict, as humans, too, diverge in their understanding of various aspects of those attitudes. We conclude that at present researchers of ethnicity representations in the Russian language social media can rely on our general negative sentiment and conflict detection classifiers to look for problematic texts, and (with more caution) on positive sentiment and positive interaction detection classifiers to find texts that can potentially contribute to interethnic understanding and peace. While these instruments can still be improved, their quality is already reasonably high. However, prediction of the other two variables should be improved.

Apart from obtaining more marked-up data (which is currently being done), several directions for improvement may be outlined. First, it is necessary to set stricter criteria for classifying a text as truly devoted to ethnicity. With more data is will become possible to select only texts on which the opinion of at least two persons coincided. Further, adding a more contrastive collection of non-relevant texts might improve the quality of topic classification. Next, predicting other variables based only on relevant texts might influence quality – in particular, excluding texts where ethnonyms were only mentioned, but the topic of ethnicity was absent, might bring forward lexical features used to express attitudes to ethnic groups.

In general, however, improvement of attitude detection should follow a different path. This task similar to opinion mining from consumer reviews with multiple entities, e.g. that compare several products. For such cases the existing approaches usually recommend to perform sentence-level analysis. A problem with social media texts is that often sentence division is not clear. Furthermore, ethnic groups, unlike most consumer goods, can interact, and this is usually expressed in the same sentence. It is thus necessary to do manual linguistic work to adjust the size of the window for left and right contexts of ethnonyms. It may be also useful to exclude ethnic groups that get rarely covered in highly opinionated pieces. However, as different groups may be described in different terms, to avoid a bias after such exclusion, it may make sense to further develop the approach of sampling from clusters – e.g. a cluster of small Siberian ethnicities might get represented by those that arouse more sentiment. Finally, it may worth trying to predict attitudes to each ethnic cluster separately (with conflict-driven clusters being divided).

We also find that attention to ethnic groups varies greatly by region and by group; we therefore expect that sentiment, attitudes and conflict levels also vary. It thus makes sense to detect not the absolute values aggregated by region, group or ethnicity cluster, but their change compared to the average level or to the previous period.

Acknowledgements. This work was done at the Laboratory for Internet Studies, National Research University Higher School of Economics (NRU HSE), Russia. It was supported by the Russian Research Foundation grant no. 15-18-00091.

References

1. Apishev, M., Koltsov, S., Koltsova, E.Y., Nikolenko, S., Vorontsov, K.: Mining ethnic content online with additively regularized topic models. Computacion y Sistemas **20**, 387–403 (2016). doi:10.13053/CyS-20-3-2473
2. Attenberg, J., Ipeirotis, P.G., Provost, F.J.: Beat the machine: challenging workers to find the unknown unknowns. In: Proceedings of 11th AAAI Conference on Human Computation, pp. 2–7 (2011)
3. Bartlett, J., et al.: Anti-Social Media. Demos, London (2014)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
5. Bodrunova, S., Koltsova, O., Nikolenko, S.: Are migranty all the same? Attitudes to re-settlers from post-soviet South in the Russian blogosphere (2016). Unpublished manuscript

6. Bodrunova, S.S., Litvinenko, A.A., Gavra, D.P., Yakunin, A.V.: Twitter-based discourse on migrants in Russia: the case of 2013 bashings in Biryulyovo. Int. Rev. Manag. Mark. **5**, 97–104 (2015)
7. Bohlin, L., Edler, D., Lancichinetti, A., Rosvall, M.: Community detection and visualization of networks with the map equation framework. In: Ding, Y., Rousseau, R., Wolfram, D. (eds.) Measuring Scholarly Impact, pp. 3–34. Springer, Cham (2014). doi:10.1007/978-3-319-10377-8_1
8. Burnap, P., Williams, M.L.: Cyber hate speech on Twitter: an application of machine classification and statistical modeling for policy and decision making. Policy Internet **7**, 223–242 (2015). doi:10.1002/poi3.85
9. Chan, J., et al.: The internet and racial hate crime: offline spillovers from online access. MIS Q.: Manag. Inf. Syst. **40**(2), 381–403 (2016)
10. Daniels, J.: Race and racism in Internet studies: a review and critique. New Media Soc. **15**, 695–719 (2013). doi:10.1177/1461444812462849
11. Dekker, R., Belabas, W., Scholten, P.: Interethnic contact online: contextualising the implications of social media use by second-generation migrant youth. J. Intercult. Stud. **36**, 450–467 (2015). doi:10.1080/07256868.2015.1049981
12. Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., Bhamidipati, N.: Hate speech detection with comment embeddings. In: Proceedings of the 24th International Conference on World Wide Web, pp. 29–30. ACM (2015). doi:10.1145/2740908.2742760
13. Faris, R., Ashar, A., Gasser, U., Joo, D.: Understanding harmful speech online. Berkman Klein Center Research Publication No. 2016-21 (2016). doi:10.2139/ssrn.2882824
14. Gagliardone, I.: Mapping and Analysing Hate Speech Online. Social Science Research Network, Rochester (2014)
15. Gibson, S., Lando, A.L.: Impact of Communication and the Media on Ethnic Conflict. IGI Global, Hershey (2015)
16. Gitari, N.D., Zuping, Z., Damien, H., Long, J.: A lexicon-based approach for hate speech detection. Int. J. Multimed. Ubiquit. Eng. **10**, 215–230 (2015). doi:10.14257/ijmue.2015.10.4.21
17. Gladkova, A.: Linguistic and cultural diversity in Russian cyberspace: examining four ethnic groups online. J. Multicult. Discourses **10**, 49–66 (2015). doi:10.1080/17447143.2015.1011657
18. Glukhov, A.P.: Construction of national identity through a social network: a case study of ethnic networks of immigrants to Russia from Central Asia. AI Soc. **32**, 101–108 (2017). doi:10.1007/s00146-016-0644-9
19. Grasmuck, S., Martin, J., Zhao, S.: Ethno-racial identity displays on Facebook. J. Comput.-Mediat. Commun. **15**, 158–188 (2009). doi:10.1111/j.1083-6101.2009.01498.x
20. Grishhenko, A.I., Nikolina, N.A.: Expressive ethnonyms as markers of hate speech [Jekspressivnye jetnonimy kak primety jazyka vrazhdy]. In: Hate Speech and Speech of Consent in the Socio-Cultural Context of Modern Society [Jazyk vrazhdy i jazyk soglasija v sociokul'turnom kontekste sovremennosti], pp. 175–187 (2006). (in Russian)
21. Kim, Y.-C., Jung, J.-Y., Ball-Rokeach, S.J.: Ethnicity, place, and communication technology: effects of ethnicity on multi-dimensional internet connectedness. Inf. Technol. People **20**, 282–303 (2007). doi:10.1108/09593840710822877
22. Korobkova, O.S.: Hate speech indicators in ethnic membership nominations: sociolinguistic aspect [Markery jazyka vrazhdy v nominacijah jetnicheskoj prinadlezhnosti: so-ciolingvisticheskij aspekt]. Izvestia: Herzen Univ. J. Humanit. Sci. [Izvestija Rossijskogo gosudarstvennogo pedagogicheskogo universiteta im. AI Gercena] 200–205 (2009). (in Russian)

23. Kwok, I., Wang, Y.: Locate the hate: detecting tweets against blacks. In: Proceedings of the 27th AAAI Conference on Artificial Intelligence, AAAI 2013, pp. 1621–1622 (2013)
24. McLaine, S.: Ethnic online communities. In: Cyberactivism: Online Activism in Theory and Practice, pp. 233–254 (2003)
25. Mustafa, H., Hamid, H.A., Ahmad, J., Siarap, K.: Intercultural relationship, prejudice and ethnocentrism in a computer-mediated communication (CMC): a time-series experiment. Asian Soc. Sci. **8**, 34–48 (2012). doi:10.5539/ass.v8n3p34
26. Nikolenko, S.I., et al.: Topic modelling for qualitative studies. J. Inf. Sci. **43**(1), 88–102 (2017)
27. Nakamura, L.: Cybertypes: Race, Ethnicity, and Identity on the Internet. Routledge, Abingdon (2013)
28. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.: Abusive language detection in online user content. In: Proceedings of the 25th International Conference on World Wide Web, pp. 145–153. International World Wide Web Conferences Steering Committee (2016). doi:10.1145/2872427.2883062
29. Parker, D., Song, M.: New ethnicities online: reflexive racialisation and the internet. Soc. Rev. **54**, 575–594 (2006). doi:10.1111/j.1467-954X.2006.00630.x
30. Silva, L., Mondal, M., Correa, D., Benevenuto, F., Weber, I.: Analyzing the targets of hate in online social media. In: Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016, pp. 687–690 (2016)
31. Steinfeldt, J.A., Foltz, B.D., Kaladow, J.K., Carlson, T.N., Pagano Jr., L.A., Benton, E., Steinfeldt, M.C.: Racism in the electronic age: role of online forums in expressing racial attitudes about American Indians. Cult. Divers. Ethnic Minor. Psychol. **16**, 362–371 (2010). doi:10.1037/a0018692
32. Sternin, I.A.: Politically incorrect national names in language consciousness of language's possessor [Nepolitkorrektnye naimenovanija lic v jazykovom soznanii nositelja jazyka]. Polit. linguist. [Politicheskaja lingvistika] **1**, 191–193 (2013)
33. Trebbe, J., Schoenhagen, P.: Ethnic minorities in the mass media: how migrants perceive their representation in Swiss public television. J. Int. Migr. Integr. **12**, 411–428 (2011). doi:10.1007/s12134-011-0175-7
34. Tukachinsky, R., Mastro, D., Yarchi, M.: Documenting portrayals of race/ethnicity on primetime television over a 20-year span and their association with national-level racial/ethnic attitudes. J. Soc. Issues **71**, 17–38 (2015). doi:10.1111/josi.12094
35. Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., Daelemans, W.: A dictionary-based approach to racism detection in Dutch social media. arXiv preprint arXiv:1608.08738 (2016)
36. Tynes, B.M., Giang, M.T., Thompson, G.N.: Ethnic identity, intergroup contact, and outgroup orientation among diverse groups of adolescents on the Internet. CyberPsychol. Behav. **11**, 459–465 (2008). doi:10.1089/cpb.2007.0085
37. Vepreva, I.T., Kupina, N.A.: The words of unrest in the world today: unofficial ethnonyms in real usage [Trevozhnaja leksika tekushhego vremeni: neoficial'nye jetnonimy v funkcii aktu-al'nyh slov]. Polit. linguist. [Politicheskaja lingvistika] 43–50 (2014). (in Russian)
38. Warner, W., Hirschberg, J.: Detecting hate speech on the world wide web. In: Proceedings of the Second Workshop on Language in Social Media, pp. 19–26. Association for Computational Linguistics (2012)
39. Waseem, Z.: Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In: Proceedings of the 1st Workshop on Natural Language Processing and Computational Social Science, pp. 138–142 (2016)
40. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In: Proceedings of NAACL-HLT 2016, pp. 88–93 (2016)
41. Zhu, Z.: Making the "invisible" a "visible problem"—the representation of Chinese illegal immigrants in US newspapers. J. Chin. Overseas **10**, 61–90 (2014). doi:10.1163/17932548-12341268