Methodological Challenges for Detecting Interethnic Hostility on Social Media

Olessia Koltsova(✉)

Laboratory for Internet Studies, National Research University Higher School of Economics, Room 117, 55/2 Sedova Street, Saint-Petersburg, Russia
ekoltsova@hse.ru

Abstract. Detection of ethnic hate speech and other types of ethnicity representation is user texts is an important goal both for social and computer science, as well as for public policy making. To date, quite a few algorithms have been trained to detect hate speech, however, what policy makers and social scientists need are complete pipelines, from definition of ethnicity to a user-friendly monitoring system able to aggregate results of large-scale social media analysis. In this essay, the author summarizes the experience of development of such a system in a series of projects under the author's leadership. All steps of the offered methodology are described and critically reviewed, and a special attention is paid to the strengths and the limitations of different approaches that were and can be applied along the developed pipeline. All conclusions are based on prior experiments with several large datasets from Russian language social media, including 15 000 marked up texts extracted from a representative one-year collection of 2.7 million user messages containing ethnonyms.

Keywords: Ethnic representations · Machine learning · Social media

1 Introduction

Studies of interethnic relations and related issues have a long tradition and well-developed methodologies based on approaches of social science. They allow using the results for practical purposes, including ethnic conflict forecasting and prevention, monitoring ethnic fractionalization and inter-ethnic hostility, development of ethnic tolerance, among others. Nevertheless, in the last two decades two new factors have emerged allowing a new research optics being applied to monitoring of interethnic relations. First, rapid development of the Internet has made it a repository of attitudes of the growing "online population" and a space where socially important discussions and conflicts evolve. Even now user generated content can be regarded as an important source of public opinion, or at least its reasonable proxy, which can supplement and sometimes substitute opinion polls. Second, development of data mining, especially related to large text collections, enables researchers to automatically detect trends in such collections, e.g. to reveal topical structure of discussions, topic salience, sentiment prevailing in texts and topics, etc. This has been shown in many works, including those by our research group [1, 2].

S. S. Bodrunova et al. (Eds.): INSCI 2018 Workshops, LNCS 11551, pp. 7–18, 2019.
https://doi.org/10.1007/978-3-030-17705-8_1

Thus, the analysis of internet data with such methods can be used for monitoring of interethnic hostility, its strength, context, geolocation and a range of other parameters. However, until recently data mining research was developing mainly within mathematics and/or computer science and was aimed at development of algorithms and mathematical models rather than at social modeling based on such tools. This methodological gap has been shown by our research group. Thus, one of the main algorithms for studying large text collections – topic modeling [3–5] – is widely used in scientometrics, but only in the last few years the first attempts have been made to apply it for the analysis of social problems based on texts of internet users [6, 7]. Supervised machine learning has been more widely used for a wide range of goals [8–10] including hate speech [11, 12], still its application to the ethnicity-related issues is in its cradle [13, 14].

The first attempts to apply machine learning to user content for social science goals, including those of detecting ethnicity related opinions have revealed a whole range of methodological problems. The nature of these problems is absolutely new for social scientists. These problems are related not only for technical obstacles and lack of mathematical expertise among social scientists, but mainly to the absence of verified approaches to connecting new mathematical models and technical solutions with social science tasks. An incomplete list of such problems includes: formulation of tasks for data driven research that are nevertheless relevant to important social science problems, operationalization of concepts in a way applicable for machine learning, making meaningful samples from big text collections, procedures for interpretation of machine learning results, hypothesis testing on big data where standard statistics does not work, and more. In this paper we address some of those problems focusing on a task of detecting human opinions about ethnicity expressed in social media.

The rest of this methodological essay is structured as follows. In the next section the author describes the data that were used in methodological experiments in different projects lead by the author. These experiments serve the bases for methodological conclusions made in the subsequent sections. Section 3 addresses approaches to defining ethnicity as the object to be detected in user texts. Section 4 reviews advantages and limitations of different methods to detecting representations of ethnicity that were tested in the mentioned above experiments. Finally, Sect. 5 describes solutions implemented in the online system developed under supervision of the author. It shows how some of the mentioned problems may be solved in practice.

2 Data

Methodological reflections presented in this paper are based on the following data.

1. A range of samples created from the LiveJournal data collected for different projects. These samples include all posts by top 2000 Russian speaking bloggers for the periods: (1) 12 months from mid-2013 to mid-2014 – 1.58 million messages; (2) for one month of 2013 – 103,000; (3) for 3.5 months 2013 – 364,000; (4) for 4 months 2013 – 235,000. These different collections were created for different experiments.

2. Collections of random VKotakte users that include all their posts and comments to them, for two years 2013 and 2014. VKontakte users were selected randomly from each of 87 Russian regions; approximately 800 users from each region. This produced a collection of 74,000 users, approximately 9 million posts and 1 million comments to them. From this collection, a smaller sample was made that included only three regions (Tatarstan, Buryat republic and Tver oblast) and 222,545 messages.
3. A collection of messages from all Russian-language social networking sites monitored by IQBuzz commercial aggregator. This collection contains all messages containing at least one of the words or bigrams from a specially constructed vocabulary of Post-Soviet ethnonyms for two years – 2013 and 2014. After cleaning from duplicates the collection comprise around 2.7 million messages of which 80% are produced by VKontakte. Of them, 60% come from group pages and 40% - from individual accounts.
4. A marked-up collection of 15,000 messages selected randomly from collections 2 and 3 so as to represent each of 115 Post-Soviet ethnic groups. Each text of this collection contains from 10 to 90 words and has been hand coded by at least three assessors answering a long list of questions.

Collections 2–4 were collected specifically for the study of ethnic relations online. Numerous experiments performed on them are described in the respective papers [1, 2, 15, 16]. Here, we do not give all details of these experiments focusing on reflecting on our methodological experience instead.

3 Defining Ethnicity and Texts on Ethnicity

The concept of ethnicity has been much theorized about, while public opinions about ethnicity have been a constant object of empirical research. However, this stream of literature turns to be irrelevant when it comes to automatic detection of opinions about ethnicity online.

The concept of ethnicity has been most often discussed together with the related concepts of nation and nationality [17, 18]. One typical opposition in this context is described with primordial vs constructivist approach opposition [19: pp. 39–46, 17: pp. 20–45]. Primordialist approach claims that ethnicity is determined by ancestry, that the ethnic status is ascribed at birth and that the ethnic boundaries are fixed. Constructionist approach claims that ethnicity is nothing more than self-perception and perception of an individual by others, i.e. it claims that ethnic status and ethnic boundaries are collectively constructed, negotiated and challenged. That makes ethnicity a purely social category as opposed to “biological”. Within constructionism one can discern culture-centered approaches and polity-centered approaches, the latter making ethnicity close to the concept of nation. However, defining ethnicity according to any of these approaches gives no cues for mining attitudes to ethnicity in lay texts, because lay persons do not usually have knowledge of theoretical concepts, neither they follow them in everyday talk even if they do know social theory. Ethnicity in user texts is thus the usage of ethnic markers by text authors.

A researcher's task at this point may seem to be reduced to compiling a list of ethnic status markers (names of ethnic groups, or ethnonyms) and retrieving texts containing these words. However, as shown in our experiments, this approach works only partially. First, there is no formal criterion to discern between ethnonyms and nation names: formally, we can distinguish stateless ethnic groups (Roma, Kurds), non-ethnic nations (Egyptians, Indians) and those that coincide or nearly coincide (French, Norwegians). By including French and excluding Indians one makes an overall voluntaristic choice. When all ethnic groups and nations are included, many texts that are yielded with such keyword search in fact deal with what obviously looks like international relations and war. Plural form of ethnonyms may be used to denote governments or generally states and countries. On the other hand, texts with ethnonyms denoting individuals may deal with issues other than ethnicity, while ethnicity is in fact used as one of identifiers along with gender, age, profession, personal name and other markers that may be used in non-discriminative and non-problematic manner. Such texts include those that present results of international contests, including sports and culture, in case they do not politicize or "ethnicize" those issues. Finally, ethnicity may be discussed without mentioning any ethnonyms if more general concepts are applied or if a text is referential (e.g. a reply to a post where ethnonyms are mentioned).

This ambiguity is explained with the fact that lay text authors use ethnonyms for different purposes and do not have a goal of differentiating between ethnically related and ethnically irrelevant texts. Drawing the boundary between such texts thus turns to be the task of a researcher. Returning to the start of this section we can see that for studying opinions about ethnicity in lay texts researcher's role shifts from defining ethnicity per se, as an abstract concept, to defining what a text about ethnicity is. In our works we suggest the following solution [2]. Texts about ethnicity are defined as texts:

1. where the major actors are private persons of a given ethnicity or ethnic groups, and not states or their official representatives (e.g. "Turks have broken a recent international agreement" is not about ethnicity, while "Chinese are not good at European languages" is);
2. where ethnicity is important for the outcomes or is used as an explanation (e.g. "The singer was a real Yakut, so we've heard an authentic throat singing" is about ethnicity; "We've just returned from an exhibition of a French photographer, and it's too late to go anywhere else" is not).

This is not the only possible solution. In fact, solutions should depend on research goals, but this example shows that, furthermore, sampling relevant texts turns into a separate machine learning problem. It can be broken into two parts: first, building an approach and an instrument for automatic differentiating between relevant and irrelevant texts, and second, finding such texts among millions of other texts without substantial losses in precision, recall (completeness), time, memory and computational resources. Although these two tasks seem unrelated, our experiments show that differentiating between relevant and irrelevant texts in a small high-relevant collection and in a collection of millions of noisy texts are two very different tasks. Therefore, a single algorithm is needed that can simultaneously look for texts based on preliminary criteria and then classify them in real time.

To the best of our knowledge, no research deals with this problem. Studies on ethnic hate speech detection and related topics usually focus on text classification once the texts are already available, and the proportion of the target class is high [11, 20, 21]. To our knowledge, only the newest research in China tackles this problem of "finding a needle in a haystack", although mostly in relation to monitoring crises and broader risky events [22]. Research on algorithms of real-time classification of streaming online data does exist, but it somehow attempts to develop domain-independent or at least ethnicity unrelated classification methods [23, 24]. Our system introduced in the last section so far gives only a partial solution for this problem.

4 Strengths and Weaknesses of Different Approaches to the Detection of Ethnic Representations in User Texts

Different methods of automatic text analysis possess different advantages and limitations, and the social science community is just beginning to get acquainted with those. Comprehensive overviews of automated approaches to text analysis for social scientists are provided in [25] who focus on political science tasks and [26] interested in journalism and media studies tasks. The former authors illustrate an explanation of different methods with a variety of research goals, such as classification of political texts into topical categories, known or unknown beforehand, extraction of political slants from texts and placement of text characters in political spectra. Broadly speaking, it is possible to extract unknown categories with unsupervised machine learning (UML) techniques, akin to cluster analysis, while those that are known beforehand are better to be searched for with supervised machine learning approaches (SML) that demand algorithm tuning and validation on a collection of texts whose categories are already marked up.

Many of our experiments with ethnic texts have been based on topic modeling, a group of UML algorithms that allow to co-cluster both texts and words into topics returning lists of most probable words and texts for each cluster. In our experiments, it is represented by three main algorithms: variational LDA [5], LDA with Gibbs sampling [4] and BigARTM [27]. This approach is very useful when topics – or in our case contexts in which different ethnic groups are discussed – are unknown beforehand. It is also good when topics are changing and any mark-up gets outdated fast. Finally, it is good when mark-up is too costly. However, our experience has revealed a number of severe limitations of this approach that has made it hardly usable for our goals.

First, it turned out that topic modeling is unable to extract relevant topics from large collections when the proportion of texts devoted to the topics of interest is very low (well below 1%). This is the case of ethnicity discussion online: a vast majority of user texts is everyday talk; only a small minority is related to social or political issues, and of them ethnicity is again a minority. It is important to note that for some reason TM is sensitive to the proportion, not the absolute number of relevant texts: when a few thousands of relevant texts are extracted from millions with alternative methods, and then mixed with a few thousand of irrelevant texts, TM works well. However, for such enrichment some mark-up is usually needed which partially makes UML senseless.

Second, although it is known that TM works poorly on short texts, our experience shows that it does not work at all if the texts, apart from being short, also represent everyday talk. Around 90% of topics yielded from random VKontakte collections turn to be uninterpretable, independently of the collection size or the number of topics. We have also observed that when a collection of short texts contains some proportion of longer ones TM interpretability improves a lot. Thus, mean length of texts in our IQBuzz sample is 332 words against 16.5 in random VKontakte collection; although both have power-law distribution, the first has yielded much more interpretable results than the second. Third, most topics that were yielded on non-enriched collections were rather about international relations than ethnicity. The effect when smaller topics get overshadowed by related larger topics has been also observed on other tasks. Likewise, less frequently mentioned ethnic groups get overshadowed by more frequently mentioned that tend to form larger and more interpretable topics. Among Post-Soviet ethnic groups, the most frequent are Ukrainians, Jews and Chechens. However, all of them, in turn, get overshadowed by nations of the global influence, first of all Americans and Germans. Finally, our experiments on news data [7, 28] show that TM results produced on them are way better than those obtained on collections of social media texts. This makes us think that topic modeling, at least in its current form, is not really suitable for short user texts. It is thus difficult to apply it for studying user opinions, even most broadly understood.

Supervised machine learning has been more widely applied to opinion detection. The most relevant literature in the field is mostly aimed at automatic detecting of hate speech in user-generated content [29] 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 [30]. Others attempt to capture broader types of hate speech, e.g. related to race, ethnicity/nationality and religion simultaneously [13, 21], or even generalized hate speech [12] and abusive language [31]. 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 [32].

In such works, a large variety of techniques is being offered and developed, including lexicon-based approaches [21], classical classification algorithms [20] and a large number of extensions for quality improvement, such as learning distributed lowdimensional representations of texts [33], using extra-linguistic features of texts [11] and others. Some draw attention to the role of human annotators and the procedure of annotations for classification results [34, 35].

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 judgment 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” [36] “unwanted”, “intimidating”, “frightening” [37] which can be summarized as actually, bad. The related concepts of prejudice are somewhat more precisely defined. As it has been noted by Quillian [38] that most of them rely on early Allport’s definition which views prejudice as “an antipathy based on faulty and inflexible generalization” [39] while the positive counterpart of prejudice is usually referred to as positive stereotype.

All the mentioned definitions 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.

We tend to think that for opinion detection on political sensitive issues it is important to elaborate concrete questions for human coders that would allow them to annotate texts independently of their political views or cultural values. In our research, we employed a range of questions that include both text-level and instance level aspects of opinions. Text-level aspects are: (1) general problematization of the topic in the text (does the text contain negative/positive sentiment); (2) conflict presence (does the text mention inter-ethnic conflict or positive inter-ethnic interaction?); (3) text topics (a choice from among 14 social and political topics, including ethnicity and “other”). Instance-level aspects are: (4) general attitude (What is the general attitude of the text author to a given ethnic group? Negative/positive/neutral); (5) perception of ethnic hierarchy (Does the author treat a given ethnic group as superior/inferior?); (6) danger perception (Does the author perceive a given ethnic group as dangerous?); (7) blame attribution (In case of conflict, does the author present a given ethnic group as a victim/an aggressor?); (8) call for violence (Does the author call for violence against a given ethnic group?). This is, again, not the only way to approach ethnicity-related opinion detection, however, it has produced reasonably good quality in prediction of some of the aspects.

We specially instructed coders that they are not expected to make moral judgements of text authors for the opinions they express. All coders were also trained to recognize attitudes in texts. Still we find a lot of divergence among coders. At meetings, they posed many questions and expressed difficulties in classifying different types of texts. Their overall judgement was that most categories were vague and prone to subjectivity. We must point at this as one of the major limitations of classification of social issues in texts. A machine cannot be expected to classify texts better than humans, and while humans widely diverge, the machine learning result will stay low. Even if some group of coders can be trained to think unanimously, their judgement will reflect the result of training and not the way in which a broader society thinks. A promising way to overcome this problem is not in seeking for consensus among a narrow group of trained humans, but accommodating for the lack of consensus within the procedure of machine learning through fuzzy logic. Texts should be assigned to classes with weights corresponding to the level of inter-coder agreement which will yield sets of core and

periphery texts for each class. An algorithm then might be trained to guess the level of consensus and to differentiate between the most typical and less typical texts in each class.

Another problem we have encountered is a trade-off between working with text-level and instance-level items. We find that around a half of ethnicity-relevant texts contain more than one ethnic group; of them 15% contain a combination of neutral and emotional attitudes to different ethnic groups, and 6% are opposite attitudes. In such situation, while predicting instance-level items, that is aspects of attitudes to specific ethnic groups, with the entire set of text features sometimes leads to prediction of different opinions with the same data. This misleads the algorithm and decreases its quality. This situation does not occur with the text-level items, but they are less informative in sociological sense.

One of the potential solutions for this problem is to try to detect attitudes to specific ethnic groups at the sentence level. So far, it has been seldom done. Most studies use only unigrams or bigrams as features, as we have already done [8, 11–14, 22, 31, 40–42]. Syntactic features have been used in [13, 14, 21, 31]. The problem with user texts is that their syntax is often flawed. However, sentences, according to our experience, are relatively well delimited either with dots or emoticons. Thus one could apply an approach based on windows of fixed length and additionally limited by end-of-sentence markers, without stricter syntactic parsing. Another problem is, however, that with such ambiguous issues as ethnicity opinion is most often expressed indirectly and dispersed across multiple sentences while co-reference resolution is difficult due to flawed syntactic structure. This problem to our knowledge has no solution so far. Below, we describe the solutions to some of the listed above problems implemented in our system.

5 Solutions Implemented in TopicMiner

Our system is devised to monitor ethnic relations on the Post-Soviet space. Its main goal is to trace, in a semi-automatic way, distribution of discussions about ethnicity in the Russian-language social media over time and space. The primary task of this tracing is early prevention of emerging inter-ethnic conflicts through a sequence of methodological steps. Those steps been translated into a system of concrete methods and algorithms, and they in turn have been implemented in a user-friendly software available online.

Online system is available at: https://topicminer.hse.ru/. It contains the following functionality and components.

First, the methodology takes into account that a user may have access only to noisy, unfiltered data with a low proportion of texts about ethnicity (e.g. raw dumps of social media messages). Our system does not collect data, but it contains a number of instruments for text preprocessing, whose core is a methodology that filters texts non-relevant to the topic of ethnicity. As mentioned above, our experiments have shown that detection of any ethnicity-related trends in large collections of texts is impossible without pre-filtering. Therefore, the methodology consists of two components: text selection based on a lexicon of ethnonyms containing 3680 individual words and 12670

bigrams (precision up to 74%) and a machine learning based selection (precision and recall around 74%). We recommend to combine these two approaches to increase recall.

Second, based on such collection enrichment, our system allows to extract topics, or contexts in which ethnical issues are discussed and which are not known to researchers beforehand. For this, we offer a number of improvements for topic modeling algorithms whose quality has been tested both manually and with a specially developed quality metric – tf-idf coherence. Our experiments have shown that a basic pLSA algorithm with our lexicon of ethnonyms yields the best results among all BigARTM algorithms. It is best suited for revealing the entire range of ethnicity related topics existing in a given collection, for comparison of those topics by their volume, and for detection of topics devoted simultaneously to several ethnic groups. To extract contexts related to a single pre-defined ethnic group, a better option is our other algorithm with a more aggressive partial supervision – ISLDA which also exceeds basic LDA both by the proportion of ethnically relevant topics and by their tf-idf coherence.

Introduced algorithms were tested on different collections listed above. Good results were achieved with collections containing a certain proportion of relevant and long texts. The main contribution into quality of the tested models came from our lexicon of ethnonyms. The overall conclusion from the experiments is that although topic modeling cannot be used for extraction of relevant texts from collections with a low proportion of such texts (and this task was solved via supervised classification), topic modeling nevertheless works well for detection of contexts in which ethnicity is discussed. All listed above algorithms are implemented in our system which also has functionality of tipping on ethnically relevant topics based on comparison of topics' top words with our lexicon of ethnonyms.

Third, the system is able to yield distributions of ethnically relevant topics over time and space and visualize them on a time scale or on the map of Russia, respectively. Besides simply summing the probabilities of a given topic over all texts of a given region or time period, our methodology includes specially tuned multimodal algorithms of topic modeling where timestamps and geolocation tags are made a separate modality. Our experiments have shown that this approach works better than simple summing for revealing topics concentrated in time, although it penalizes topics evenly distributed over time. For obtaining a more precise distribution by the Russian regions we have also calculated a set of correction coefficients accounting for uneven penetration of social networks across Russian subjects of Federation.

Fourth, our methodology allows revealing the listed above aspects of attitudes to the problems of ethnicity. This part of methodology is based on algorithms trained with a marked-up collection containing 15,000 messages about 115 postSoviet ethnic groups. For such aspects as danger and call for violence, there was no sufficient data to train a classifier. Other instance-level aspects have produced mixed results, of them the best quality was obtained for classes "superior" and "aggressor". At the text level, negative aspects – conflict presence and negative sentiment – are predicted better than positive ones; algorithms trained to predict these two aspects have been integrated into our online system. Besides this, the system was equipped with a function of sentiment analysis of topics based on comparison of topics' topwords with our sentiment lexicon.

Our experiments have also shown that doubling the size of the marked-up collection, although it improves quality of classification, does not solve the problem

radically; furthermore, the quality seems to be unrelated to the level of inter-coder agreement. This suggests that ways of further improvement of classification of attitudes to ethnic issues should be searched for via extracting specific grammatical constructions.

It should be noted that as direct calls for violence against any ethnic groups occur in less than 1% of ethnically relevant texts, negative attitudes are mostly expressed more indirectly or vague. Beyond LiveJournal positive aspects of attitude prevail over negative ones, although this may be explained with over-representation of small nationalities in the marked-up sample. Simultaneously, these marked-up texts are more characterized by generalized vision of ethnic groups, negative sentiment and conflict mentioning than by positive sentiment, mentioning of positive inter-ethnic interaction and of concrete persons of a given ethnicity. In other words, users problematize the topic of ethnicity in general more often than they express a direct negative attitude to certain ethnic groups or persons.

6 Conclusion

This methodological essay, based on a whole series of our projects, did not aim at presenting ready solutions that we report elsewhere. Instead, we have tried to attract the attention of the research community to important methodological problems that are seldom discussed in published academic papers because the latter tend to focus on successful results, not on the difficulties a researcher encounters on the way to them. We have shown that the existing methods to automatically detect ethnic hostility online, as well as other social categories, are still under development and should not be used as black boxes. At the same time, it makes little sense to wait until they ripen because efficient development of methods may occur only in collaboration of those who develop methods (computer scientists) and those who set the goals for them (social scientists).

Acknowledgements. This paper is mainly based on the experience from the research project "Development of concept and methodology for multi-level monitoring of the state of interethnic relations with the data from social media" RSF grant No 15-18-00091, 2015–2017, as well as the ongoing research implemented in the Laboratory for Internet Studies in the framework of the Basic Research Program of National Research University Higher School of Economics. The author is thankful to all project participants: Sergei Koltcov, Konstantin Vorontsov, Sergey Nikolenko, Svetlana Bodrunova, Murat Apishev, Svetlana Alexeeva, and Oleg Nagornyy.

References

1. Koltsova, O., Alexeeva, S., Nikolenko, S., Koltsov, M.: Measuring prejudice and ethnic tensions in user-generated content. Ann. Rev. CyberTherapy Telemed. (2017)
2. Koltsova, O., Nikolenko, S., Alexeeva, S., Nagornyy, O., Koltcov, S.: Detecting interethnic relations with the data from social media. In: Alexandrov, D.A., Boukhanovsky, A.V., Chugunov, A.V., Kabanov, Y., Koltsova, O. (eds.) DTGS 2017. CCIS, vol. 745, pp. 16–30. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69784-0_2

3. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. **42**(1), 177–196 (2011)
4. Griffiths, T., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. **101**, 5228–5235 (2004)
5. Blei, D.M.: Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012)
6. Flaounas, I., et al.: Research methods in the age of the digital journalism: massive-scale automated analysis of news content. Digit. Journal. **1**(1), 102–116 (2013)
7. Nagornyy, O., Koltsova, O.: Mining media topics perceived as social problems by online audiences: use of a data mining approach in sociology. NRU Higher School of Economics, (WP BRP 74/SOC/2017)
8. Chen, Y., Zhou, Y., Zhu, S., Xu, H.: Detecting offensive language in social media to protect adolescent online safety. In: Privacy, Security, Risk and Trust (PASSAT), International Conference on Social Computing (SocialCom), Amsterdam, Netherlands, pp. 71–80 (2012)
9. Scharkow, M.: Thematic content analysis using supervised machine learning: an empirical evaluation using German online news. Qual. Quant. **47**(2), 761–773 (2013)
10. Burscher, B., Odijk, D., Vliegenthart, R., de Rijke, M., de Vreese, C.H.: Teaching the computer to code frames in news: comparing two supervised machine learning approaches to frame analysis. Commun. Methods Meas. **8**(3), 190–206 (2014)
11. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In: SRW@ HLT-NAACL, pp. 88–93 (2016)
12. Warner, W., Hirschberg, J.: Detecting hate speech on the world wide web. In: Proceedings of the Second Workshop on Language in Social Media, Stroudsburg, PA, USA, pp. 19–26. Association for Computational Linguistics (2012)
13. Burnap, P., Williams, M.: Cyber hate speech on Twitter: an application of machine classification and statistical modeling for policy and decision making. Policy Internet **7**(2), 223–242 (2015)
14. Burnap, P., Williams, M.: Us and them: identifying cyber hate on Twitter across multiple protected characteristics. EPJ Data Sci. **5**(1), 1–15 (2016)
15. Apishev, M., Koltsov, S., Koltsova, O., Nikolenko, S., Vorontsov, K.: Mining ethnic content online with additively regularized topic models. Computacion y Sistemas **20**(3), 387–403 (2016)
16. Nikolenko, S., Koltcov, S., Koltsova, O.: Topic modelling for qualitative studies. J. Inf. Sci. **1**, 1–15 (2017)
17. May, S.: Ethnicity, Nationalism and the Politics of Language. Taylor & Francis, Abingdon (2012)
18. Song, S.: The subject of multiculturalism: culture, religion, language, ethnicity, nationality, and race? In: Bruin, B., et al. (eds.) New Waves in Political Philosophy. Palgrave McMillan, London (2009). https://doi.org/10.1057/9780230234994_10
19. Yang, P.Q.: Ethnic Studies: Issues and Approaches. State University of New York Press, New York (2000)
20. Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, D., Daelemans, W.A: Dictionary-based approach to racism detection in Dutch social media. In: First Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS2016), pp. 11–16 (2016)
21. Gitari, N.D., Zuping, Z., Hanyurwimfura, D., Long, J.: A lexicon-based approach for hate speech detection. Int. J. Multimed. Ubiquit. Eng. **10**(4), 215–230 (2015)
22. Xu, Z., Liu, Y., Mei, L., Luo, X., Wei, X., Hu, C.: Crowdsourcing based description of urban emergency events using social media big data. IEEE Trans. Cloud Comput. **99** (2016)
23. Zubiaga, A., Spina, D., Martínez, R., Fresno, V.: Real-time classification of Twitter trends. J. Assoc. Inf. Sci. Technol. **66**(3), 462–473 (2015)

24. Yar, E., Delibalta, I., Baruh, L., Kozat, S.S.: Online text classification for real life tweet analysis. In: 24th Signal Processing and Communication Application Conference (2016)
25. Grimmer, J., Stewart, B.M.: Text as data: the promise and pitfalls of automatic content analysis methods for political texts. Polit. Anal. **21**(3), 267–297 (2013)
26. Günther, E., Quandt, T.: Word counts and topic models: automated text analysis methods for digital journalism research. Digit. Journal. **4**(1), 75–88 (2016)
27. Vorontsov, K., Frei, O., Apishev, M., Romov, P., Dudarenko, M.: BigARTM: open source library for regularized multimodal topic modeling of large collections. In: Khachay, M.Y., Konstantinova, N., Panchenko, A., Ignatov, D.I., Labunets, V.G. (eds.) AIST 2015. CCIS, vol. 542, pp. 370–381. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-26123-2_36
28. Koltsova O., Pashakhin S.: Agenda divergence in a developing conflict: a quantitative evidence from a Ukrainian and a Russian TV newsfeeds. Sociology, WP BRP 79/SOC/2017
29. Bartlett, J., Reffin, J., Rumball, N., Williamson, S.: Anti-social media. Demos, 1–51 (2014)
30. Kwok, I., Wang, Y.: Locate the hate: detecting tweets against blacks. In: des Jardins, M., Littman, M.L. (eds.) AAAI, Bellevue, Washington, USA, pp. 1621–1622. AAAI Press (2013)
31. 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)
32. 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)
33. 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)
34. 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)
35. Waseem Z.: Are you a racist or am i seeing things? Annotator influence on hate speech detection on Twitter. In: Proceedings of 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science, pp. 138–142. ACL, Austin (2016)
36. Gagliardone, I., Patel, A., Pohjonen, M.: Mapping and Analysing Hate Speech Online: Opportunities and Challenges for Ethiopia. University of Oxford, Oxford (2014)
37. Faris, R., Ashar, A., Gasser, U., Joo, D.: Understanding Harmful Speech Online. Berkman Klein Center Research Publication No. 2016-21 (2016)
38. Quillian, L.: New approaches to understanding prejudice and discrimination. Ann. Rev. Sociol. **32**, 299–338 (2009)
39. Allport, G.W.: The Nature of Prejudice. Addison, New York (1954)
40. Sood, S.O., Churchill, E.F., Antin, J.: Automatic identification of personal insults on social news sites. J. Am. Soc. Inf. Sci. Technol. **63**(2), 270–285 (2012)
41. Van Hee C., et al.: Detection and fine-grained classification of cyberbullying events. In: Proceedings of Recent Advances in Natural Language Processing, Proceedings, Hissar, Bulgaria, pp. 672–680 (2015)
42. Hosseinmardi, H., Mattson, S.A., Rafiq R.I., Han, R., Lv, Q., Mishra, S.: Detection of cyberbullying incidents on the Instagram social network. CoRR, abs/1503.03909 (2015)