Measuring Prejudice and Ethnic Tensions in User-Generated Content

Olessia KOLTSOVA[a,*], Svetlana ALEXEEVA [a,b], Sergey NIKOLENKO [a,c], Maxim KOLTSOV [a]

National Research University Higher School of Economics, Russia [b] *St. Petersburg State University, Russia*
Steklov Mathematical Institute at St. Petersburg, Russia

Abstract. With the spread of social media, ethnic prejudice is becoming publicly available to widening audiences and may have serious offline consequences. This creates demand to detect prejudice and other signs of ethnic tension in user-generated texts, and this task is absolutely different from measuring prejudice with surveys – an approach traditionally developed in psychology. In this work we use a hand coding instrument based on psychological definitions of prejudice and sociological methods of questionnaire construction. Compared to our previous research, we double our hand-coded collection that reaches 14,998 unique user texts retrieved from the Russian language social media. We then train computer classification algorithms to "guess" prejudice as detected by human coders and show significant improvement in quality compared to our earlier results. Still, as not all aspects of prejudice get detected sufficiently well, we analyze potential causes of low quality and outline directions for further improvement.

Keywords. Ethnicity, prejudice detection, user content, machine learning.

Introduction.

The role of prejudice for interethnic anxiety, contact and conflict has been a long-studied topic in ethnic psychology and political science [1-2]. Recent explosive growth of social media has allowed prejudiced views to spread to large audiences with potentially increased risks of offline spill-over [3], which has led to new research tasks of measuring and monitoring online ethnic prejudice and/or tolerance. Unlike polls, user texts contain only what users choose to share and are often ambiguous. Ultimately, their influence is limited to the meanings that readers manage to extract from them, therefore, the key to prejudice detection in texts is interpretation by ordinary people. Accurate detection begins with a set of well-elaborated questions to human coders designed to overcome text ambiguity and human subjectivity as much as possible. However, with millions of texts online the next step is to teach computer to automatically "see" what humans see

* Corresponding Author: ekoltsova@hse.ru

in texts. Development of automatic methods of prejudice detection is the main goal of this research.

Related work.

Although there are many approaches to prejudice in psychology and social science, it has been noted by Quillian [4] that most of them rely on early Allport's definition which views prejudice as "an antipathy based on faulty and inflexible generalization" [5], while the positive counterpart of prejudice is usually referred to as positive stereotype. With some exceptions [6] prejudice is thus usually seen as a type of attitude; therefore, methods to reveal it are usually survey-based and include a large variety of scales [7] most often focusing on ethnic prejudice. To the best of our knowledge, psychology has never sought to detect (ethnic) prejudice expressed in "natural" texts – that is, those produced by the objects of study for the purposes other than prejudice research.

This line of inquiry has been developing in linguistics, communication and media studies. The most elaborated approach by van Dijk [8-9] is, however, aimed at revealing deep and complex structures of prejudiced text with sophisticated discourse analysis. A stricter measurement instrument is offered by Ponarin and colleagues [10], still neither approach scales for large amounts of texts that are increasingly available online.

Contemporary computer science and computational linguistics offers a number of methods to automatically detect various text features, however, prejudice-related methods are being in their cradle. Most often, researchers attempt to detect hate speech that is vaguely defined and ideologically burdened [11-12]. Very little attention is paid to the development of instruments for manual text mark-up that are later used as the ground truth for testing automatic instruments.

To our knowledge, there are no works aiming at detecting presence of ethnic conflict or other inter-ethnic tension in texts although some works seek to predict generalized offline conflict with social media data [13-14], including Russian-language content.

Research goals and data.

This research elaborates on our earlier work [15] in which we offered a pilot instrument for prejudice and conflict detection based by Ponarin and colleagues [10] and other sources. In that research we obtained a decent level of quality that, nevertheless, has room for development. In this work we improve our earlier results and
increase the number of prejudice aspects that we are able to detect. We examine when and how the quality of the instrument can be augmented and analyze in-depth when and why automatic methods work poorly. We outline directions of inquiry that may lead to further improvement.

We used data from the same collection of 2,660,222 user texts mentioning at least one of 115 post-Soviet ethnic groups. This collection covers a two-year period and embraces all Russian language social networking sites. It is fully described by Koltsova and colleagues [15]. However, here we increase our sample for hand coding from 7,181 to 14,998 unique texts.

Prejudice and conflict are detected along with their positive counterparts to make results more sound. Overall prejudice and positive stereotyping are detected with a combination of two core questions: (a) whether the text is devoted to the topic of

ethnicity, i.e. whether the ethnic status of the character(s) matters for the meaning of the text; and (b) what is the overall attitude of the text author to the character described with an ethnonym – positive, negative or neutral. The second question is given to coders only if the topic of ethnicity is present, because only if the ethnic status of the character is important, the attitude to this character may be based on it and, therefore, may be prejudiced.

To detect more nuances of prejudice we ask if the ethnic characters are seen as superior/inferior, aggressors/victims, dangerous/non-dangerous, and if the call for violence against them is present. We also ask if the mentioned character is an individual or the generalized group as we hypothesize that stereotyping is related to generalizations. We, indeed, find that in around 88% of interpretable instances ethnic characters are generalized mentions of ethnic groups.

For conflict detection, we use questions asking if inter-ethnic conflict or positive inter-ethnic interaction are mentioned in the text. For tension detection, we use the questions about presence of general negative and positive sentiment; we think that on the aggregate level, prevalence of negative sentiment in texts about a certain ethnic group can indicate tensions related to it. Finally, we use a number of control and filtering questions.

We get each text coded by at least three independent persons, and as one text may contain multiple ethnonyms we obtain 32,701 unique instances of ethnonyms in texts detected, albeit not all of them detected by multiple coders.

As we have already experimented with feature selection and feature weighting (unigrams against bigrams+unigrams, and tf-idf weights against raw frequencies) and found they have a modest effect on the quality of our instrument, our two major assumptions are about the effects of collection size and inter-rater agreement. The latter in our case may depend both on coding quality and on the ambiguity of the issues being coded; unfortunately, it is difficult to differentiate between them because there seems to be no ground truth other than the opinion of coders.

Results.

In this research we replicate our classification procedure from Koltsova and colleagues on the enlarged collection. We train logistic regression with scikit-learn library with the best parameters from the previous research thus "teaching" the computer to guess what humans think about the texts. For that purpose, we average their scores and round them to obtain distribution of our texts over the "true" (human-driven) classes. We then check the work of our algorithm against these classes with traditional quality metrics: recall (the ability of the algorithm to find all texts of a given true class) and precision (the ability to find only those texts that are truly of a given class). Not all classes are of equal interest for us, but those that are not interesting (e.g. texts without any attitude) prevail in numbers. As prevailing classes usually contribute most to the quality of classification, and are simultaneously better predicted, we focus our attention on precision and recall for our target classes – e.g. for texts with conflict vs texts without conflict. Therefore, it is for these classes that we report the gain over the baseline algorithm: a classifier that randomly assigns classes to texts / instances, albeit keeping the true class proportion.

We start from the core question about attitude towards ethnic characters (see Table 1). We see that both quality metrics have improved significantly compared to the previous research [15]. The only exception is recall for positive attitude detection.

However, we do not see much improvement for the variables related to inter-group interaction and general sentiment (Table 2). With our collection being doubled, the quality has increased only by a few percent and it has even dropped for recall in positive sentiment. Although in any case we exceed the baseline, this situation is surprising as usually an increase in collection size is thought to be important for quality.

Table 1. Attitude to ethnic characters expressed in user texts: quality of prediction.

	Precision	Recall	F1-score
Class -1 (negative attitude)	0,61	0,36	0,45
Class 0 (neutral attitude)	0,73	0,93	0,82
Class +1 (positive attitude)	0,67	0,36	0,47
Average	0,67	0,55	0,58
Gain over random baseline for class -1	0,44	0,19	0,28
Gain over random baseline for class +1	0,46	0,14	0,25
Gain over previous research for class -1	0,18	0,15	0,17
Gain over previous research for class +1	0,18	0,03	0,08

The major difference of the *attitude* variable from the rest is that from a text-level feature it has become an instance-level feature which means that the increase in collection size has been dramatic – from 7 to 32 thousand entries. In the earlier research we analyzed data for only one ethnonym for each of 7,181 texts, while here we used the full data. The quality has increased despite some 11 thousand of instances were coded only by one coder and other 12 thousand got graded by two coders who diverged in 2,824 cases. Also, 9,925 texts contain multiple ethnonyms, who received different average grades on attitude in 2,375 cases and opposite grades in 640 cases. Although we did not filter out those cases either, the quality is still good.

Table 2. Presence of positive and negative sentiment, inter-ethnic conflict and positive inter-ethnic interaction in user texts: quality of prediction.

		Precision	Recall	F1-measure
Negative sentiment	Class 1 (has neg. sent.)	0,81	0,87	0,84
	Average	0,75	0,73	0,74
	Gain over prev. research, class 1	*+0,05*	*+0,09*	*+0.07*
Positive sentiment	Class 1 (has pos. sent.)	0,66	0,36	0,46
	Average	0,72	0,64	0,65
	Gain over prev. research, class 1	*+0,03*	*-0,08*	*-0,06*
Inter-ethnic conflict	Class 1 (has conflict)	0,70	0,64	0,67
	Average	0,71	0,71	0,71
	Gain over prev. research, class 1	*+0,03*	*+0,07*	*+0,06*
Positive interaction	Class 1 (has pos. interaction)	0,61	0,30	0,41
	Average	0,70	0,62	0,63
	Gain over prev. research, class 1	*+0,03*	*+0,03*	*+0,04*

We then assume that the quality of classification may depend on inter-rater agreement: the classes that humans find difficult to discern between may indeed possess fewer lexical differences and thus be harder to detect automatically. We compute several inter-rater agreement metrics on subsets of cases that got three grades each. We then compare our variables by the following metrics: Krippendorf's alpha, Fleiss' kappa, average pairwise Cohen's kappa, average pairwise agreement, the share of cases with three coinciding grades, the share of cases with two coinciding grades, and the ratio between these two shares both in general and for target classes specifically. The

differences between these metrics for different variables are in most cases not too high, and we do not observe any clear relation between them and the quality of classification. Therefore we do not repost the values here.

Instead, we have found in literature [16] that positive classes are usually harder to predict than negative which is what we observe in our case. We, nevertheless, exceed the quality reported in [16]. We also observe that using tf-idf weights instead of raw frequencies highly prioritizes precision over recall, and although the overall F1-measure is usually higher with tf-idf weights in developing alert systems for early detection of ethnic tensions, recall might be more important.

Next, we examine class sizes for all variables, including those that could not be modeled before due to their scarcity. We define class size as the number of cases that were coded by at least three coders of whom at least two thirds agreed on a given class plus the number of cases coded by two coders both of whom agreed. We find that call for violence is represented by only 106 instances and presentation of ethnic groups as dangerous – by 601 instances. Most other class sizes start from 1000 although *aggressor* is smaller, still we choose to develop a pilot model for both victim/aggressor and inferior/superior variables (see table 3). These models differ from the others in one respect: only cases representing the target classes have been taken into the analysis with the dominant neutral class being excluded. We obtain good quality for two of the four categories. Although *aggressor* is the smallest class, it is *victim* and *inferior* that are predicted worse, and generally we do not find any correlation between target class size and absolute quality or quality gain over the baseline.

Table 3. Treatment of ethnic characters as victims, aggressors, inferior or superior groups in user texts: quality of prediction.

	Quality			Gain over random classifier		
	Precision	Recall	F1	Precision	Recall	F1
Class 0: victim	0,54	0,44	0,48	0,07	0,15	0,11
Class 1: aggressor	0,70	0,77	0,73	0,16	0,06	0,11
Average	0,62	0,61	0,61			
Class 0: inferior	0,65	0,64	0,64	0,19	0,18	0,18
Class 1: superior	0,70	0,70	0,70	0,13	0,16	0,16
Average	0,67	0,67	0,67			

Conclusion and future research.

In this work we improve our instrument that now shows higher quality in detection of more aspects of prejudice and inter-ethnic tensions in online user texts than before. In particular the system is good at predicting when an ethnic character or a group is presented as aggressor, which allows tracking cases of group-level blame attribution. It is equally good in finding cases when a group is treated as superior over others which allows detecting ethno-centric biases. General negative sentiment is also well predicted; coupled with analysis of ethnonyms mentioned in respective texts it may help analysts determine which ethnic groups are related to this negativism and, therefore, might arouse ethnic tensions.

Some other aspects are predicted with reasonable quality. Among them are the cases in which ethnic characters and groups are treated as inferior which is one of the most direct indications of prejudice. Presence of inter-ethnic conflict is also fairly well predicted being one of the most important elements for constructing early alert systems that can help prevent ethnic conflict both online and offline. Predictions of negative and positive attitude to ethnic characters, as well general positive sentiment and positive inter-ethnic interaction stands relatively high in terms of precision, but quite low in terms

of recall. To date the algorithm fails to find about two thirds of such cases. Shifting from tf-idf term weights to raw word frequencies increases recall up to 45% but results in the drop of precision to as low as 50%.

We also examine dependence of quality of prediction on collection size, target class size and various measures of inter-rater agreement, but we do not find any sound trend. However, so far for all variables we have used as many cases as possible, including those coded by a single coder in order to increase target class size. As the latter does not appear to be crucial, one of the directions for future research is to set stricter inter-rater agreement thresholds for including cases into target classes. With a large proportion of ambiguous texts it may make sense to try fuzzy classification. Finally, there is a lot to try in terms of models other than logistic regression, as well as in terms of enrichment of the data with external information, such as distributed word representations.

Acknowledgements

This work was supported by the Russian Science Foundation (grant 15-18-00091).

References

L.R.Tropp, The psychological impact of prejudice: implication for intergroup contact, *Group Processes and Intergroup Relations* **6:2** (2003), 131-149.

D.P.Green, L.R.Seher, What role does prejudice play in ethnic conflict? *Annual Review of Political Science* **6** (2003), 509-531.

M.Hsueh, K.Yogeeswaran, S.Malinen, Leave your comment below: can biased online comments influence our own prejudicial attitudes and behaviors? *Human Communication Research* **41:4** (2015), 557-576.

L. Quillian, New approaches to understanding prejudice and discrimination. *Annual Review of Sociology 32* (2006), 299–338.

G.W. Allport, *The Nature of Prejudice*, Addison, New York, 1954.

E.R. Smith, Social identity and social emotions: toward new conceptualizations of prejudice. In: D.M. Mackie, D.L. Hamilton (eds*) Affect, Cognition and Stereotyping: Interactive Processes in Group Perception*, Academic Press, INC, San Diego, 1993, 297-315.

T. D. Nelson, *Handbook of Prejudice, Stereotyping, and Discrimination*, Psychology Press, New York, 2009.

T.A. van Dijk, *Prejudice in Discourse.* Benjamins, Amsterdam, 1984.

T.A. van Dijk, *Racism and the Press*, Routledge, London, 1991.

E. Ponarin, D. Dubrovsky, A.N. Tolkacheva, R. Akifiev,. Index of press (in)tolerance. In: A. Verkhovsky (ed) *Hate Speech against the Society*, Sova, Moscow, 2007, 72-106 (in Russian).

N.D. Gitari, Z. Zuping, H. Damien, J. Long, A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering* **10**, (2015), 215–230.

W. Warner, J. Hirschberg, Detecting hate speech on the world wide web. *Proceedings of the Second Workshop on Language in Social Media (ACL-2012)*, (2012), 19–26.

M. Rosell, C. Mårtenson, F. Johansson, P. Hörling, M. Malm, S. Truvé, J. Brynielsson, Detecting emergent conflicts through web mining and visualization, *European Intelligence and Security Informatics Conference*, (2011), 346-353.

T. Delavallade, L. Mouillet, B. Bouchon-Meunier, Monitoring event flows and modeling scenarios for crisis prediction: application to ethnic conflicts forecasting. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **15**:supp01, (2007), 83-110.

O. Koltsova, S. Nikolenko, S. Alexeeva, O. Nagorny, S. Koltcov, Detecting interethnic relations with the data from social media, *Proceedings of the Second International Conference "Digital Transformation & Global Society"*, (2017) (forthcoming).

M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, A. Kappas, Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology* **61:12**, (2010), 2544–2558.