National Research University Higher School of Economics

as a manuscript

Koltsova Elena Yurievna

APPLYING AUTOMATIC LANGUAGE PROCESSING TO INVESTIGATE THE COVERAGE OF INTER-ETHNIC RELATIONS AND OTHER SOCIALLY PROBLEMATIC TOPICS IN LARGE COLLECTIONS OF USER-GENERATED TEXTS

Dissertation Summary
for the purpose of obtaining
academic degree
Doctor of Science in Philology and Linguistics

Moscow 2024

General information

Scientific problem and its importance

The proliferation of the Internet, and especially of services for communication between users who are not professional media producers, creates a large number of digital traces, including those containing texts. These traces present rich material for social scientists, as they bear valuable information about society, its elements and processes. The study of such traces for sociological purposes requires developing new linguistic methods, not well known to social scientists, but also not typical for linguistics. Given the large volume of such data, these methods rely on automatic natural language processing and the achievements of computer science, for which, however, solving sociological tasks is neither a focus. All this creates a deficit of relevant methodologies, and above all complex methodological approaches that contain all stages: formulation of a sociological problem in terms of computational linguistics, sampling text data in accordance with sociological criteria, automatic processing of texts and combining them with "manual" content analysis to obtain results relevant for social sciences.

This dissertation consists of eight articles published as an output of a single project and presents new original methodologies to analyze large volumes of Internet users' texts about interethnic relations. Interethnic relations is a traditional issue of sociological research, and until recently they were studied mainly through surveys. Recently, however, the attention of sociologists, political scientists, and psychologists has been drawn to the interethnic processes occurring on the Internet, especially to discussions on this topic which often turn into hate speech. Studies show that social media contribute to 'offline' interethnic conflicts and interethnic hate crimes (Williams et al 2020, Chang et al 2016). Some studies also show the positive impact of online interethnic friendships on offline interethnic cooperation and reduction of interethnic tensions both abroad (Žeželj et al 2017) and in Russia (Galyapina and Molodikova 2023).

This determines the importance of studying how ordinary social media users talk about representatives of different ethnic groups, including their own, in particular, what their explicitly expressed attitudes towards different groups are, and what contexts occurring in their utterances indicate implicit stereotypes (for example, "Caucasians" may appear most often in the context of crime). Answering these questions requires developing two types of methods for automatics analysis of user texts. Unsupervised machine learning (ML) methods, such as topic modeling, are useful for identifying unknown contexts of ethnic group representations, while supervised machine learning methods, such as neural network classification algorithms, solve the problem of automatically detecting pre-determined types of representations of ethnic groups and their relationships. These two groups of algorithms are complementary, and their combination can assist learning which types of representations of which ethnic groups predominate in which contexts, and thus can help to meaningfully interpret the nature of negative representations (e.g., the representation of some groups as dangerous may be associated with a religious context, while others may be associated with a context of unsanitary conditions and the risk of infections). Adapting these algorithms to such tasks and embedding them in sociological analytical toolkits in the form of ready-to-use methodological pipelines is of great practical importance. For example, early detection of outbursts of polarized discussions in social networks can help prevent conflicts offline or mitigate their consequences, and drawing attention to texts describing positive interethnic interaction can help establish interethnic dialogue.

Important terms

- Ethnic group - a social group identified by external and internal observers using some of the following attributes: language, culture, religion, area of residence, phenotypic traits, and "blood". In sociology, it is described as a social construct (a vague and possibly "unreal" concept actually used in society and with social consequences) rather than a term defined by social scientists themselves.
- Representation of an ethnic group - a way used to describe a group or its relations with other groups, creating a certain image or stereotype of this group
- Explicit representation of an ethnic group - expressing attitudes towards the group or its relationships with others that can be understood by the reader without reference to other texts; for example, representing an ethnic group as superior or inferior, dangerous or safe, or relationships as conflictual or peaceful.
- Implicit representation - selective mentioning of a group in certain contexts, which, as a rule, can be traced only by analyzing a large number of texts (e.g. mentioning Central Asian ethnic groups in the context of house construction, flee markets and infections, and the French in the context of haute couture, resorts and travel).
- Hate speech - in this paper: the negative extreme of a general attitude towards an ethnic group or an ethnic character in a text, where general attitude is a type of explicit representation.
- Text - one or more statements of the same author, posted by him/her as a separate publication (article, post, commentary)
- Ethnonym - a word or a group of words denoting an ethnic group or its representative in a text, including some quasi-ethnonyms (Caucasians) and ethnopholisms (ethnonyms with derogatory connotations).
- Ethnic Group as a Named Entity (EGNE) - the set of all ethnonyms describing the same ethnic group in that can be described by a single "root" ethnonym.

Research goal and objectives

The goal of this research is to propose a complex methodology consisting of newly developed or modified and tested methods of practical automatic language analysis for sociological tasks (focusing on the analysis of representations of ethnic groups and interethnic relations in social media texts).

Objectives

1. Formulate and operationalize a sociological definition of the concept of "ethnic group representation in speech" and its types.
2. Develop a comprehensive and verifiable methodology for manual markup of social media user texts, both for the pre-elected types of representations and for topic modeling results.
3. Create a series of corpora of social media texts, with and without markup, representing different groups of users.
4. Determine the applicability of topic modeling for extracting ethnic group representation contexts, for finding the most ethnorelevant texts for subsequent manual analysis, and for comparing the salience of different ethnorelevant topics within text collections.
5. Test different semi-supervised topic modeling extensions aimed to better extract narrow topics (in our case – ethno-relevant).
6. Develop and test a domain-specific sentiment lexicon for socio-political social media texts.

7. Test a wide range of classification models and approaches to text preprocessing, including the use of sentiment lexicon, for the task of identifying different types of explicit representations of ethnic groups.

Propositions to defend

1. The proposed typology of ethnic group representations in social media texts allows differentiating ethnic groups by the way they get stereotyped.
2. Classical topic modeling (LDA without modifications) extracts ethno-relevant topics (as a type of rare topics) from medium-sized collections pre-filtered by relevance, and does not solve this problem on random and very large text corpora. This prevents it from being used as a tool to measure the representation of rare topics in the general discourse of social media, but it does allow comparing the representation of different ethnorelevant topics with each other in smaller collections representing special subsets of users.
3. ISLDA algorithm in which groups of target ethnonyms are each assigned to their own range of topics allows to extract a larger number of ethno-relevant topics (especially in solutions with a large number of topics) that are additionally more concentrated and more coherent than those produced by classical LDA. ISLDA is suitable for studying topics about a limited number of ethnic groups on medium-sized collections with relatively long texts.
4. The BigARTM family of algorithms, regularised in a special way and using an ethnonym lexicon as a separate modality, allows finding ethnorelevant interpretable topics better than without a lexicon. This algorithm is suitable for very large collections of relatively long texts pre-selected by the ethnonym lexicon.
5. BigARTM with the extended lexicon allows to extract more topics, on average slightly more ethno-relevant and slightly less interpretable than BigARTM with the limited lexicon of ethnonyms. This allows recommending the use of an extended lexicon enriched by ethnic adjectives and country names for partial TM training.
6. The developed sentiment lexicon PolSentiLex outperforms RuSentiLex in identifying negative sentiment in ethnorelevant texts, positive and negative sentiment in socio-political blogs and does not differ in the quality of identifying positive sentiment in ethnorelevant texts.
7. Text-level representation classes are predicted with sufficient quality at the level of text as a unit of analysis (presence of inter-ethnic conflict, positive inter-ethnic interaction, overall negative and positive sentiment)
8. Representations of particular ethnic groups, including hate speech towards particular groups, are predicted at the level of EGNE, with the quality of the three-class classification (negative, or hate speech, positive and neutral representations) exceeding that of the two-class classification (presence or absence of hate speech).
9. Neural network algorithms significantly outperform classical classifiers in the task of hate speech detection, even though the size of the marked collection is relatively small.
10. Prediction of hate speech at the EGNE level using artificial neural networks demonstrates higher quality than text-level prediction; this becomes possible by adding an ethnonym denoting the target EGNE as a paired text to the text in which it occurs, when formulating the task as a pairwise classification before feeding it into the BERT model and further into the classifier.
11. Linguistic features improve the prediction quality when used together, but not separately, yielding the best results (a) among classical models – in the architecture of voting classifier, and (b) among ANNs – in the combination of Conversational RuBERT and the subsequent dense layer.

These propositions are further mentioned as "proposition i" in those places the Research Summary section that address the results corresponding to a certain proposition.

Novelty

The general scientific novelty of this interdisciplinary research is as follows

- For the first time a detailed definition of the concept of "representation of ethnicity in text" and its types tested in terms of their ability to differentiate ethnic groups has been formulated.
- For the first time, based on a large number of experiments, modifications and settings of topic modeling approach optimal for identifying interpretable ethnorelevant topics and, accordingly, implicit representations of ethnicity in large collections of social media texts. For this purpose, a comprehensive evaluation of the quality of the obtained topics and topic modeling solutions, combining classical and new automatic and manual methods, was implemented for the first time.
- For the first time, a highly accurate algorithm for detecting explicit general attitudes towards ethnic groups in texts, some of which contain more than one EGNE, is proposed.
- For the first time, a set of collections of social media texts in Russian has been created, some of which are large in size and represent broad groups of Internet users, while the other part contains unique multi-aspect markup, including that on representations of ethnicity.

In addition to general novelty, the work has significant novelty specific to the field of linguistics:

- A comprehensive method for assessing the quality of topic modeling has been proposed, including manual mark-up;
- A detailed methodology has been developed for marking texts according to a set of theoretically based types of attitudes towards ethnic groups;
- A lexcion of ethnonyms has been created, which can be used both to select ethno-relevant texts and to improve the quality of topic modeling when identifying ethno-relevant topics;
- A domain-specific lexicon for detecting sentiment of socio-political texts in social media has been proposed. Its positive role in the work of classifiers in identifying ethnic hate speech has been shown;
- The positive role of other linguistic features has been shown for improvement of the quality of hate speech detection algorithms;
- For the first time, a qualitative analysis of errors in the performance of neural network algorithms was carried out, thus identifying linguistic features of texts that either facilitate or set barriers for the detection of ethnicity-targeted hate speech;
- The developed text collections with markup are freely available and contribute to reducing the shortage of linguistic resources in Russian;
- All problems were solved for the Russian language for the first time.

Research articles for the defense

1. Bodrunova S., Koltsova O., Koltcov., S., Nikolenko. Who's Bad? Attitudes Toward Resettlers From the Post-Soviet South Versus Other Nations in the Russian Blogosphere // *International Journal of Communication*. 2017. Vol. 11. P. 3242-3264. (Scopus Q1, WoS Q3, List A)
2. Nikolenko S., Koltcov S., Koltsova, O. Topic modelling for qualitative studies // *Journal of Information Science*. 2017. Vol. 43. No. 1. P. 88-102. (WoS Q2, Scopus Q1, List A)

3. Apishev M., Koltsov S., Koltsova O. Mining ethnic content online with additively regularized topic models // *Computacion y Sistemas*. 2016. Vol. 20. No. 3. P. 387-403. (Scopus Q3, List C)
4. Koltsova O. Yu, Alexeeva S. V., Koltcov S. N. An Opinion Word Lexicon and a Training Dataset for Russian Sentiment Analysis of Social Media, in: Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва,1–4 июля 2016 г.) / Под общ. ред.: В. Селегей. Вып. 15. М. : Изд-во РГГУ, 2016. P. 277-287. Scopus Procs
5. Koltsova O., Alexeeva S., Pashakhin S., Koltsov S. PolSentiLex: Sentiment Detection in Socio-political Discussions on Russian Social Media, in: *Artificial Intelligence and Natural Language. AINL 2020. Communications in Computer and Information Science* Book 1292: Communications in Computer and Information Science. Cham: Springer, 2020. P. 1-16. Scopus Procs
6. Koltsova O., Nikolenko S. I., Alexeeva S. V., Nagornyy O. S., Koltsov S. Detecting interethnic relations with the data from social media, in: *Digital Transformation & Global Society: Second International Conference, DTGS 2017, St. Petersburg, Russia, June 21-23, 2017, Revised Selected Papers*. Springer, 2017. P. 16-30. Scopus Procs
7. Koltsova O., Alexeeva S., Nikolenko, S., Koltsov M. Measuring Prejudice and Ethnic Tensions in User-Generated Content // *Annual Review of CyberTherapy and Telemedicine*. 2017 (Scopus Q4, Список С)
8. Pronoza E., Panicheva P., Koltsova O., Rosso, P. Detecting ethnicity-targeted hate speech in Russian social media texts // *Information Processing and Management*. 2021. Vol. 58. No. 6. Article 102674. (WoS Q1, List A)

Other articles by the author on the topic of the study

9. Koltsova O., Pashakhin S. Agenda Divergence in A Developing Conflict: A Quantitative Evidence from A Ukrainian and A Russian TV Newsfeeds // *Media, War and Conflict*. 2020. Vol. 13. No. 3. P. 237-257. (Scopus Q2, List A)
10. Koltsova O., Koltsov S., Nikolenko S. I. Communities of co-commenting in the Russian LiveJournal and their topical coherence // Internet Research. 2016. Vol. 26. No. 3. P. 710-732. (WoS Q1, Scopus Q1, List A)
11. Nagornyy, O., & Koltsova, O. Y. (2019) Redefining media agendas: topic problematization in online reader comments. *Media and Communication*, Vo.7, No.3, P. 145–156. (Scopus Q2, List C)
12. Panicheva P., Mararitsa L., Sorokin S., Koltsova O., Rosso P. Predicting subjective well-being in a high-risk sample of Russian mental health app users // *EPJ Data Science*. 2022. Vol. 11. Article 21 (Q1 WoS & Scopus, List A)
13. Koltsova O., Scherbak A. N. 'LiveJournal Libra!': The political blogosphere and voting preferences in Russia in 2011–2012 // *New Media and Society*. 2015. Vol. 17. No. 10. P. 1715-1732. (WoS & Scopus Q1, List A)
14. Koltsova O., Koltcov S. Mapping the Public Agenda with Topic Modeling: The Case of the Russian LiveJournal // *Policy & Internet*. 2013. Vol. 5. No. 2. P. 207-227. (Scopus No Q in 2013, List A)
15. Koltcov S., Ignatenko V. and Koltsova O., Estimating Topic Modeling Performance with Sharma–Mittal Entropy, *Entropy* 2019, 21(7), 660. (WoS Q2, Scopus Q2, List C)
16. Koltsova O. Social media for joint experimental, survey and observational data collection: The case of VKontakte (VK). The SAGE Handbook of Social Media Research Methods, 2nd ed., 2022. (WoS-indexed handbook)

17. Koltsova Ye.Yu., Maslinksy K.A. Vyyavleniye tematicheskoy struktury rossiyskoy blogosfery: avtomaticheskiye metody analiza tekstov // Sotsiologia: metodologia, metody, matematicheskoye modelirovaniye, 2013. No 36. S. 113-139. (List D)

This numbering of articles is used hereinafter to refer to them when describing the content of the work. The numbering of articles 1-8 is given in the order of work execution (not in the order of publication).

Funding

The work was supported by the following grants:

- RSF Grant № 15-18-00091 "Development of the concept and methodology of multilevel monitoring of interethnic relations based on social media data".
- RHSF Grant No. 14-04-12031 "Development of a public database and crowdsourcing web resource for creating sentiment analysis tools".
- NRU HSE PFR grants: 83-2014, 68-2017, 67-2018, 61-2019 (partially).

Research Summary

Data

The study uses specially created collections of texts of two types: long blog texts, which are either original posts or reposts of professional media texts written in literary language, and short texts (posts and comments) of social media users written in informal language.

1. LJposts: the collection contains approximately 1.58 million texts over a period of one year (from mid-2013 to mid-2014) and includes all posts of the top 2000 bloggers on LiveJournal, the most popular platform for socio-political blogging in Russia at the time. The data were downloaded weekly according to the current ranking of bloggers. The collection is used in its entirety in paper 3 for topic modeling on big data and in papers 1, 2, 4 to form subsets 1.1-1.5.

1.1. LJposts11: a subset of LJposts comprising 11 weeks from 4 February to 19 May 2013 (363,579 posts). The period was selected using sociological criteria. Used in Article 1 to construct the LJposts11ethnic sample.

1.2. LJposts11ethnic: a marked-up subset of LJposts11 of 990 texts, which included top 30 texts from each of the 33 topics of the topic modeling solution recognized as ethnorelevant. The markup is based on an early version of the coding sheet and captures different types of ethnic group representations (including inferior / superior, ingroup / outgroup, etc.).

1.3. LJPosts12: subset of LJposts including the first 4 months of 2013 and 235,407 posts. Used in paper 2 to validate the ISLDA algorithm.

1.4. LJpostsRandom4: four collections representing 4 random months from 2012 and 2013, generated using the LJPosts principle. Used in paper 2 to validate the tf-idf-coherence quality metric.

1.5. LJpostsPol: a subset of LJposts of 70,710 posts, which included texts with a probability of at least 0.1 on 104 topics of a topic modeling solution recognized as socio-political. Used in Articles 4, 5 to form the first version of the sentiment lexicon.

2. LJcomments: All comments on posts in the LJposts collection, where all comments on a given post are aggregated into a single text. About 0.9 million texts. Used in article 5 to form the LJcommentsPol collection.

2.1. LJcommentsPol: a subset of LJcomments, formed similarly to the LJpostsPol collection. Includes 15,188 aggregated comment texts with a probability of at least 0.001 on 88 topics from a topic modeling solution recognised as socio-political. Used in Article 5 along with LJpostsPol to form the final version of the sentiment lexicon.

3. VKrandom: a cross-sectional collection of all texts from the walls of 74,303 VKontakte users, randomly selected from each federal subject of the Russian Federation in proportion to the regional VKontakte audience. Total 9,168,353 posts and 933,516 comments on them; collection time: summer 2015. Used in interim experiments for articles 3, 6.

4. RuEthnics: a collection of all posts (posts and comments) from all Russian-language social media for 1.5 years (January 2014 - December 2015) containing at least one of 115 post-Soviet EGNE from the author's lexicon of ethnonyms. Coverage of all social media and primary data collection was provided by IQBuzz. After deduplication and other preprocessing, the collection contains 2,660,222 texts about 97 ethnic groups. It was used in its entirety in paper

8 for pre-training the ConversRuBERT algorithm and in papers 6, 7, 8 for forming subsets 4.1-4.3.

4.1. RuEthnicsMarked1: a subset of RuEthnics of 7181 texts selected to represent all 97 ethnic groups and labelled according to the final version of the coding sheet capturing different types of representations of ethnic groups (including inferior / superior, aggressor / victim, dangerous / non-dangerous) and inter-ethnic relations (including the presence of ethnic conflict). Used in paper 6 for pilot training of an algorithm for recognizing different types of representations.

4.2. RuEthnicsMarked2: a subset of RuEthnics and an extension of RuEthnicsMarked1. Contains 14998 texts marked according to the final version of the coding sheet and some new data on added texts. Used in paper 7 to test the effect of collection size on classification quality.

4.3. RuEthnoHate: a subset of RuEthnics and an extension of RuEthnicsMarked2, which includes a procedure for selecting the highest quality marked-up texts from RuEthnicsMarked2 and a procedure for additional sampling and additional markup. The final collection contains 5,594 texts and is used in Paper 8 to train a neural network classifier to predict general attitudes towards an ethnic group (negative = hate speech, neutral and positive).

Datasets 1.2 and 4.1.-4.3 have a two-level structure, with text-level variables (e.g., date) and EGNE-level variables (relationship to EGNE). An observation in these datasets is a set of mentions of EGNE in a given text; thus, each text may account for a different number of observations in the dataset (equal to the number of mentioned ethnic groups). These datasets were labelled by three independent coders.

Rationale for corpora construction. LiveJournal was chosen as the most popular socio-political blog platform in Russia at the time of data collection. The study found that the most popular bloggers had more similarities to professional media producers than to ordinary Internet users, and the middle range of LJ ranking was polluted with bots. Therefore, later it was decided to construct a sample of VKontakte users, a social network that had become the most popular by the time of the next data collection. A random sample was used, stratified by subjects of the Russian Federation in proportion to the representation of regions in the VKontakte audience; inactive accounts were excluded. This sample fully satisfies the criteria of sociological representativeness (of VKontakte audience), however, mostly negative results were obtained using it. Therefore, it was decided to construct a similar sample, but enriched with ethno-relevant information. RuEthnics sample was formed as a full population of ethno-relevant messages for a certain period, which is also close to the sociological criteria of representativeness, except that it may not contain ethno-relevant texts that do not mention ethnonyms. In general, the RuEthnics group of corpora represents a reasonable compromise between sociological needs and the capabilities of automatic language processing methods.

Types of representations of ethnic groups in texts and a tool to measure them

A theoretically grounded and clearly operationalized definition of the phenomenon whose classes are to be identified by machine learning is the key to successful classification, especially for such fuzzy categories that sociologists have to work with. The theories of interethnic relations and intergroup interaction in sociology and psychology agree that the attitude to an ethnic group, both ingroup and outgroup, is a complex multifaceted phenomenon containing implicit and explicit components. Sociology has well-developed tools for measuring attitudes towards ethnic groups in opinion polls, and media research is widely used to identify biases in the representations of different social groups through manual analysis of samples of media texts, often qualitatively. However, no reliable tools have been proposed so far to capture how attitudes towards ethnic groups and its aspects

are represented in public texts of social media users, i.e. in situations where ordinary citizens do not respond to a structured questionnaire and where the analyzed text is not the result of professional media production.

Based on the analysis of different theories, this work proposes an instrument consisting of the following questions for specially trained text assessors (coders).

Text level questions

A. Does the text mention inter-ethnic conflict?
B. Does the text mention positive inter-ethnic interaction?

EGNE level questions:

C. Is the author talking about the ethnic group as a whole or about a specific character?
D. What is the author's general attitude toward the ethnic group or character?
E. Does the author belong to the ethnic group he/she is talking about?
F. Is the group or character described as inferior or superior to others?
G. Is the group or character described as a victim or aggressor in inter-ethnic relations?
H. Is the group or character described as dangerous?
I. Does the author call for violence against this group?

Theories of stereotyping indicate that people's opinions are characterized by overgeneralizations, where certain traits are attributed undifferentiated to the group as a whole, and this is reflected in question C. Theories of ingroup bias suggest that people tend to see their group in a more positive light, so that the representation of the group as one's own should be linked to its more positive image (question E). Hate speech theories draw attention to the need to differentiate between different degrees of hate speech, which is reflected in questions F, H and I. Finally, it should be noted that most theories focus on the negative pole of measuring attitudes towards ethnic groups, and this study aimed to identify the relationships of both polatiries. This is reflected in all the bi-polar questions (D,F,G,H) and in the pair of questions A and B. In addition, question A is aimed at identifying socially dangerous content even in texts where attitudes towards specific groups are difficult to identify, but conflict is traceable (for example, in texts where the author aggressively defends his ethnic group against the attacks of an opponent of unknown ethnicity).

Lexicon of ethnonyms

The lexicon of ethnonyms was used at different stages of the work and was created according to the following scheme: 1. Names of ethnic groups were collected from the UN data, the Russian census and some other sources, without separating ethnic ("Arab") and national ("Iraqi") groups. 2. A list of post-Soviet ethnic groups was formed from them, which was supplemented with meta-ethnonyms ("Slav"), some regionalisms ("Caucasian"), the quasi-ethnonym "Cossack" and ethnopholisms, the list of which was formed by expert judgement. The list was automatically extended with derivatives ("Armianka", "Armiashkal") and relevant bigrams ("Armenian girl", "Armenian people"). All derivatives, including ethnopholisms (e.g., "Yid" and its derivatives) were grouped under the name of the ethnic group ("Jew"), with the exception of ethnopholisms that do not point to any ethnic group unambiguously ("khach", "churka"), which were allocated to separate groups of derivatives ("khachikha", "khachonok"). The main list of post-Soviet ethnonyms contains 115 such groups.

General information on machine learning methods used in the paper

A. Unsupervised machine learning: topic modeling

Topic modeling (TM) is a group of methods that reduces the dimensionality of a word-document matrix where a row is a document (text), a column is a word, a cell is an absolute or weighted frequency of a word in the text. TM in its goal and in its end usage is similar to fuzzy co-clustering of columns and rows of such a matrix which results in each text and each word getting assigned to one or more groups from among a given number of groups. In TM these groups are understood as latent variables with the assigned meaning "topic", and the task is to recover the latent distributions of words and documents by topics from the observed data - the distribution of words by documents. Being a probabilistic algorithm, TM assigns each text and each word to each topic, but with different probabilities. The output data are two matrices - the probability matrix of words in topics and the probability matrix of topics in documents. A user can sort both the words and the texts in these matrices by their probability of belonging to the i_{th} topic; in high-quality TM solutions, the most probable words tend to give an indication of the content of the topic, and the most probable texts give an indication of the discourse characteristic of the topic. Thus, TM enables its user to quickly assess the topical structure of a large unreadable collection without knowing anything about the topics in advance, and to concentrate on reading only the documents that are most relevant to the user's task. The most common versions of TM are the Latent Dirichlet Allocation (LDA) group of algorithms and the earlier pLSA (probabilistic Latent Semantic Analysis).

TM has a number of unresolved problems, including the lack of reliable and generally accepted quality metrics and, consequently, criteria for selecting algorithm parameters. The problems encountered in this study are (1) lack of ability to detect topics if texts where these topics are strongly expressed constitute a very small proportion of the collection, (2) poor ability to scale from medium-sized collections (hundreds of thousands of documents) to very large collections (millions of documents), and (3) inability to handle short texts. For problems 1 and 2, this study proposes solutions, and for problem 3, it proposes ways to mitigate it indirectly. TM allows regularization, a procedure of adding new information to constrain the search for solutions and pushing it in a particular direction. This can be done by maximizing inter-topic differences (decorrelation) or by fixing certain words / groups of words within certain topics / groups of topics (semi-supervised approach, or partial training). These possibilities were used in this study (stages 1, 2).

B. Supervised machine learning: classifiers

Classification, designed to divide objects (e.g., texts) into predefined classes, involves the following stages. The algorithm obtains information about the observations, their features and class membership (the training collection), then iteratively selects coefficients either for each predictor feature or otherwise in a function that predicts the class as the dependent variable, so that the classes are predicted as correctly as possible according to a given quality measure. Such a trained model with fixed coefficients is then run to predict classes on a collection on which the algorithm has not been trained (the test collection). From this test run, quality measures of the algorithm are computed, among which the most common ones are precision, recall, F1-measure and accuracy. As a rule, the whole procedure is repeated several times with different partitioning of the available labelled data into test and training collections (cross-validation). An example of a simple classifier is logistic regression.

Artificial neural networks (ANNs) are complex sets of functions and algorithms whose general principle of operation is remotely similar to biological neural networks. Like simple classifiers, ANNs iteratively select feature weights in a class prediction model, starting with random or equal weights. These feature weights are fed for transformation simultaneously into multiple functions, called artificial neurons or nodes of the neural network, in the simple case one feature per node. A group of nodes that simultaneously receive a signal about all the weights is called a layer, and the entire neural network algorithm may contain one to several layers, where each successive layer receives the weight

information corrected at the previous layer. At the end, the prediction based on the feature weights optimized by the system is compared with the correct answer, and the process is iteratively repeated in one way or another.

As a rule, the basic structure of ANN is supplemented with a number of additional methods, such as: activation function (a threshold of feature weight, below which the node zeros this weight and sends a zero signal to the next layer), attention (an algorithm that differentiates features by importance), exclusion (an algorithm that randomly excludes trained neurons from further use in order to prevent overfitting of the model), concatenation (uniting groups of features obtained by different methods), etc., which were used in this paper. Neural network algorithms are also capable of generating features from raw data and accepting additional information that is not present in the data as input - for example, accepting vector representations of words (word embeddings) instead of words.

Word embeddings are the result of the work of a separate ANN algorithm, which was trained not on the collection under study, but on other data of larger volume. The simplest vector representation of a word is its representation as a sequence (vector) of numbers corresponding to the frequencies of its co-occurrence with all words in the collection. In practice, vectors of compressed dimensionality are used, where the dimensions are not words, but abstract latent variables, which are the result of reducing the dimensionality of the word space. Since word embeddings contain information about the co-occurrence of words in much larger arrays of texts than the texts under study, they allow ANNs to better "understand" texts, especially if there is little information in them - for example, if the texts are very short, and therefore the overlap between the vocabularies of even similar texts is minimal (which prevents them from being assigned to the same class). Note that if word embeddings are trained on texts that are not similar to the collection under study, they may introduce irrelevant information into the model, but for such cases, approaches to fine-tune word emdeddings on the collection under study have been developed. For each classification task, tuning of all the above aspects of the algorithm is required, which was done in step 4 of this study.

Step 1: Piloting the representation measurement tool

The following tasks were set at this stage: (a) to apply standard topic modeling to identify ethno-relevant texts and evaluate its effectiveness by human assessors; (b) to manually identify context types and types of explicit representations of ethnic groups in a subset of the most relevant texts; (c) to perform sociological analysis using standard statistical methods that would determine which types of representations and which contexts are typical for which ethnic groups, and whether the developed tool allows detecting the most relevant texts and differentiating between ethnic groups in terms of their representation.

Texts were selected from the LJposts11 collection, and the LJposts11Ethnic collection was subjected to manual markup. On these collections, the tasks were solved successfully (paper 1). After experimenting with different numbers of topics, an optimum was proposed (300-400 topics for collections containing 10^4 documents of this kind). In the topic modeling solution for 300 topics, 33 ethno-relevant and interpretable topics were identified, in some of which the contexts of ethnic group representation were well-pronounced. Examples of the 20 most likely words in such topics are: "reindeer, Khanty, holiday, chum, Surgut, north, reindeer herder, herd, Surgutian, tundra, horn, place, harness, Yakutia, Mansiysk, snowmobile, hunter"; "migrant, Chechnya, republic, Grozny, Chechen, country, Russian, citizen, Ramzan, Kadyr, Kadyrov, Chechen, citizenship, migration, Asia, Depardieu, regime". An expert reading of the most probable texts in these topics revealed that Khanty and Mansi are associated with and reduced to traditional craftsmanship, while Chechens are identified as foreign migrants with fewer rights to stay in the host territory than natives. Manual marking into contexts (political, economic, social, cultural) also showed statistically significant differentiation

between the aggregated groups of ethnicities (Tables 5, 6, Article 1). Thus, social and economic contexts prevailed in the coverage of Central Asian groups, while political contexts prevailed in the coverage of South and North Caucasian groups, with cultural context also significant for South Caucasians, and only cultural contexts co-occurred with other indigenous peoples of the Russian Federation.

Further it was shown that aggregated ethnic groups were well differentiated by scales D, F, G "general negative / positive attitude", "superior - inferior" and "victim - aggressor" - for example, North Caucasian ethnic groups were represented as significantly more aggressive than southern ones (Table 3, Article 1). The scale "ingroup-outgrup" also showed good differentiating power (Table 4 of Article 1), but the assessors found it difficult to classify EGNE by this criterium in large number of texts, so in the later version of the coding sheet it was replaced by a question about the ethnicity of the author, if it was obvious (question E). Questions asking whether the covered groups were assigned any actions or direct or indirect speech in the text showed little utility for differentiating between groups and were further eliminated from the representation measurement tool. Questions H, I collected very little data to draw conclusions about their usefulness, so they were kept in the tool for further study. Thus, an instrument for measuring the representations of ethnicity was formed, which has a statistically proven differentiating power (proposition 1 for the defense).

In addition, it was revealed that post-Soviet ethnic groups received dramatically less attention than nations of global or regional importance (Americans, Germans); however, the latter were represented not as ethnic groups, but as representatives of their countries and, accordingly, in the context of international rather than inter-ethnic relations. This led us to assume the need to differentiate between these contexts in order to identify inter-ethnic relations more effectively, which was tested at the next stage of the research.

Stage 2: Optimizing and testing topic models to identify implicit contexts of representation of interethnic relations

The first task of this stage was to test the idea of assigning a given set of words to a number, or an interval of topics, i.e., zeroing the probabilities of words from this set in all topics of the topic solution except those included in the given interval. This approach was called ISLDA - interval semi-supervised Latent Dirichlet Allocation - and was first tested in several experiments on the LJPosts11 dataset. Intervals of 3 topics were used for the 22 most frequent ethnic groups in the collection (i.e., a total of 66 target topics in a 200-topic solution) and a quality check was performed in the form of a word intrusion experiment and a similar topic intrusion experiment on assessors, evaluating topic cohesion and the algorithm's ability to assign texts to relevant topics, respectively. The experiments showed a higher human-assessed quality of target topics in ISLDA compared to non-target topics and compared to all topics in conventional LDA (Table 4 of Article 2). In addition, ISLDA target topics were better than similar topics in LDA as measured the tf-idf-coherence metric as well (Table 5 of Article 2), and experiments with the number of topics showed that the advantage of ISLDA over LDA in the ability to find large numbers of target topics on a given ethnicity appears in solutions with more topics. These results are reflected in proposition 3. The tf-idf coherence metric was also proposed in Article 2 and tested on the LJPosts12 collection, showing better results than traditional coherence.

Further, the idea of partial training was developed as follows: it was proposed to assign the entire ethnonym lexicon to a wide range of topics, which would constitute 30-50% of the total number of topics in the TM solution. This approach solves the problem of interval size selection in ISLDA and significantly expands the number of ethnonyms used. The approach was decided to be tested with a group of pLSA-based algorithms implemented in the BigARTM library. This library has a number of advantages, including simple embedding of several regularizers, as well as the functionality of

parallel computing based on the division of the collection into portions (batches). This allowed us to use it on a large collection of LJPosts containing 1.5 million posts. One of the subtasks was to test how dividing the collection into batches would affect the quality of the results. A total of 8 different models were trained on 400 topics; of these, the first two were baseline: (1) pLSA and (2) LDA in the BigARTM implementation. The second two models (3, 4) divided topics into background (250 topics) and target (150 topics) so that the occurrence of ethnonyms was penalized in background topics and encouraged in target topics,and a number of other regularisers were also used. Models 5, 6 used lexicons as a separate modality in addition to the previous techniques: model 5 used a lexicon of ethnonyms; model 6 enlarged it with ethnic adjectives and names of countries / provinces in case there were no ethnonyms in the collection (Turk -> Turkish, Turkey). This was done to test the hypothesis about the necessity of separating the topics addressing interethnic and international relations. Models 7, 8, in addition to all the previous settings, used different types of recursion: texts selected at the previous stages were fed to the input of the next stage of modeling.

The quality of the models was assessed both by coherence and tf-idf-coherence and by multiple assessors who were asked to evaluate (1) whether they understand why the given words were grouped in the topic, (2) whether they see a specific event or issue that the texts containing these words might cover, and (3) whether the topic is relevant to either inter-ethnic or international relations. Compared to word intrusion experiments, such an instrument allows distinguishing between different causes of poor topic quality. On all quality assessment metrics, models 5 and 6 clearly outperformed all others (Tables 2, 4, 5, 6 of Article 3), but did not differ significantly from each other on different aspects (Proposition 4). It is important to note that adding countries and adjectives to the lexicon caused the model to identify more topics on international relations, but did not undermine the quality and quantity of topics on interethnic relations, so model 6 was recommended as the first choice model to use (Proposition 6). In addition, the overall quality of the topics, despite the division of the collection into batches, was better than ISLDA by tf-idf coherence and not worse by expert judgement.

Further experiments with BigARTM were continued on the VKrandom collection - the best of all in terms of sociological representativeness. All of these experiments, including tests on texts aggregated by author, yielded negative results, failing to identify interpetable and/or relevant topics. The experiments were then moved to the RuEthnics collection, which is inferior to VKrandom in that it can miss ethnorelevant texts that do not contain ethnonyms. The range of model parameters to optimize, including the number of topics, was expanded in RuEthnics experiments, but the basic pLSA model still performed better than all other models. This can be explained by the fact that the role of partial training had been fulfilled by pre-selection of texts based on the presence of ethnonyms. The more successful modeling on RuEthnics compared to VKrandom can probably be explained by the presence of a certain proportion of long texts focused on one ethno-relevant topic in the RuEthnics collection; it is plausible that straightforward enlargement of VK texts by combining all posts of one author into one text did not facilitate the task of topic modeling because such quasi-texts did not acquire any dominant topic or a small group of dominant topics.

The following conclusions can be drawn from this series of experiments (thus summarizing the meaning of propositions 2-5). The task of identifying ethno-relevant topics from very large collections of user texts requires preliminary filtering of such collections. If the task of identifying non-obvious contexts of ethnicity representation - i.e. topics in texts that may not contain ethnonyms - is prioritized, the collection should not be filtered by keywords, but should be formed from relatively long texts and the bigARTM model should be used. If the priority is to identify topics in short texts, which are more likely to contain spontaneous user reactions, the collection should not be formed only from long texts, but filtered for the absence of ethnonyms. If the sample is representative of different social media (which is preferable), it is likely to contain some proportion of long texts, and this will allow topic modeling to be successful even when using models without partial training (but with parallelized computation).

When assessing the quality of the resulting topic models, attention should be paid to the following: (1) how different the topics are from each other (a large proportion of repeating topics is a sign of an algorithm failure or suboptimal settings); (2) how well the lists of the most probable words correspond to the lists of the most probable texts (inconsistency is a sign of poor quality of a topic); (3) the extent to which the topics are interpretable based on the most probable words; (4) the extent to which the topics are interpretable based on the most probable texts. Fast automatic assessment of the quality of topics based on words is possible using the tf-df coherence metric as the most consistent with human markup. In manual assessment, relying on direct questions, as opposed to relying on experiments, makes it possible to distinguish between topics that have no meaning, but possess a clear principle of formation (for instance, topics formed around obscene vocabulary or around the names of months), and topics that do have meaning , such as the above mentioned topics about the Khanty and the Chechens.

Step 3: Developing a sentimental lexicon

The sentiment lexicon, named PolSentiLex, has been developed for a number of tasks, including determining the overall sentiment of socio-political user texts and contributing to the feature space for algorithms trained to classify explicit representations of ethnic groups. The main idea in compiling the lexicon for such tasks was that sentiment weight should be assigned to words based on the meaning they have specifically in the context of socio-political user generated content. For this purpose, the following methodology was used (Article 4).

LJPosts collection was formed from top LiveJournal blogger texts as they were dominated by hot social and political topics at that time. Based on the TM for 200 topics, we, first, formed LJPostPol collection described earlier and, second, determined candidate words for the sentiment lexicon by selecting 200 most probable words from each of the topics labelled as socio-political. These words were supplemented with candidates from existing lexicons (RuSentiLex lexicon did not exist at that time), making the first version of the proto-lexicon of 9,539 words, 7,546 of which were found in the LJPostPol collection. Next, an interface was created in which assessors were presented with each word in a text (three different texts per word were included in the database) and asked to rate the overall sentiment of the word in the given context and the overall sentiment of the text on a scale from -2 to 2. A total of 32,437 word and text labels were obtained, with each text and word having been coded by at least three independent assessors whose scores were averaged and rounded. The quality of the lexicon was assessed by predicting the resulting sentiment scores of the texts based on the lexicon submitted to the SentiStrength software. The values of the precision and recall measures turned out to be above chance and at the level of existing predictions for Russian on similar tasks (Table 3 of Article 4), but not high enough to consider the work complete.

As the next step of this work, LJComments and LJCommentsPol collections were compiled, and the whole procedure described above was repeated on the texts of aggregated comments (paper 5). The second part of the proto-lexicon totalled 9,539 words, of which 6,860 occurred in the LJCommentsPol collection; 26,851 evaluations were obtained. The final version of the lexicon, obtained after combining the two parts, included 3,924 words recognized as non-neutral. By the time the work was completed, the RuSentiLex lexicon had been released, and the quality of PolSentiLex was evaluated in comparison with it. The performance of both lexicons was tested on the LJPostsPol collection, on which PolSentiLex had been trained, and on the RuEthnicsMarked2 collection, on which it had not been trained. The texts of the latter collection contained two separate labels for overall negative and overall positive sentiment on a three-point scale (not to be confused with overall attitudes towards an ethnic group), but it was overall sentiment that was predicted in the tests. The following combinations were compared: both lexicons embedded in SentiStrength (lexicon approach) and both

lexicons used as feature sets in the best solutions of three classical machine learning algorithms: SVM, KNN and NB. As a result, (a) the lexicon approach significantly outperformed machine learning on both collections; (b) PolsentiLex showed similar quality to RuSentiLex in models with machine learning and a significant advantage with the lexicon approach, as outlined in proposition 6; (c) negative class was predicted better than positive class (Figures 1, 2, 3 of Article 5).

To summarise this step, the advantage of PolSentiLex over RuSentiLex in this task is not surprising, because PolSentiLex is a domain-specific lexicon and RuSentiLex is a generic lexicon, and they do not replace each other. A more important result was the ability of the simple lexicon approach to give a quality above 60% on different metrics in the three-class task: this convinced us that our lexicon contains valuable information that could be used in the next generation MO algorithms (ANNs), so it was taken up for further work.

Step 4: Testing ML algorithms for predicting explicit representations of ethnic groups

The initial experiment of this stage predicted representations A, B (as a two-class task), D (as a three-class task - positive, negative, neutral) and the overall negative and positive sentiment of the text on the RuEthnicsMarked1 collection, which was created first (Article 6). The collection was formed so that ethnic groups were represented in the proportions in which they occured in the RuEthnics collection, with over-representation of some of the rarest groups. For the first experiment, all assessor scores were averaged, while the prediction was performed at the level of text, so that variable D (overall attitude towards the ethnic group) denoted the attitude dominant in the text and stayed undifferentiated by individual EGNE. Simple logistic regression was used as a classifier where the 6,364 most frequent words and 7,039 most frequent bigrams were used as features. The quality of the resulting predictions (Table 1 of Article 6) turned out to be noticeably higher than random, but still quite low for the main variable D and requiring further improvement. There was insufficient data for variables E,F,G,H and I in the dataset.

Therefore, the collection was further expanded into the larger RuEthnicsMarked2 dataset, formed according to the same principle. Text-level variables (A, B and sentiment) were again predicted with the entire text taken as a observation. For predicting variable D, EGNE-text pairs were used as observations (for text i, where ethnonyms x, y were mentioned, observations of type i-x, i-y were constructed). Since the body of the text for observations i-x and i-y would be the same, in order to distinguish between them during the formation of the feature matrix, information about the EGNE in relation to which the value of the variable D on a given observation should be predicted was added to the feature space of words and bigrams. The classification procedure was repeated. The results (Tables 1, 2 of Article 7) show an improvement in quality for all variables and classes except for the recall metric on overall positive sentiment. For recall by general negative attitude class and for precision by negative and positive general attitude classes, the improvement was significant (15-18%). These results reflect propositions 7 and 8.

Since the amount of data on variables E-I did not increase significantly, it was decided to focus on variable D, the main variable for our tasks, and to conduct extensive experiments with a large set of classical classifiers and neural networks (paper 8). Before training, a better RuEthnoHate collection was generated, with an enriched negative class and cleaned of texts with low inter-coder agreement. Prediction for variable D was performed on three classes (negative, positive, neutral) at the level of the EGNE-text pair (similar to article 7) and on two classes (negative, other) at the text level. For prediction at the EGNE level, information about the EGNE was added to the ANN in the form of a separate text in a pairwise prediction architecture (where the first layer of the ANN receives information from two inputs instead of one).

Classical classifiers (SVM, NB, LR, VC) and a number of feature sets were used as baseline models, including unigrams as a baseline set, negative words from PolSentiLex, and (for EGNE-level classification) words with a boosting factor from the context window around the target ethnonym (see Section 4.1. of Article 8 for a complete list). The feature sets, models and their parameters, including window size, were searched over the grid. Among these algorithms, at the EGNE level the best prediction quality was shown by VC (voting classifier) with all linguistic feature sets (including PolSentiLex), none of which improved the quality individually, but together with all others gave a significant increase (Table 4 of Article 8).

The input of the ANN was fed with word embeddings instead of words, including: Word2Vec CBOW model, Word2Vec models trained on National Russian Language Corpus and on RuEthnics, as well as Conversational RuBERT basic and pre-trained on RuEthnics. Hundreds of ANN models with different input data, architectures and settings of individual parameters were trained. LSTM+GRU models were taken as baseline models among ANNs; a schematic representation of the main tested architectures of this group of models is given in Fig. 1 of the article 8. Further, a number of more complex architectures were tested for Conversational RuBERT, presented in Fig. 2 of Article 8. An important feature of two of these three architectures is that the feature sets obtained from the output of BERT were then concatenated (combined) with the linguistic features described in Section 4.1. of Article 8. According to the results of all experiments (Table 6(2) of Article 8), Model 6, shown in Fig. 2, panel c, was the best model. It combines features from Conversational RuBERT with linguistic features that are jointly fed into a dense layer, but, unlike model 7, does not contain an LSTM layer (Proposition 11b). Its overall F1=0.892 and F1-measure on the target class (negative) F1hate=0.813. This is not only higher than all models, classical (proposition 9) and neural networks, performing three-class classification at the EGNE level, but even higher than algorithms performing two-class classification at the text level (where the best result is F1=0.864 and F1hate=0.760) (Proposition 10). Since the prediction of attitudes to ethnic groups at the EGNE level was performed for the first time not only for the Russian language, but in general, it could not be compared with analogues.

Overall, our experiments suggest the following. First, linguistic features are important for improving prediction of ethnicity representations. This points at the prospects for further work on input data formats for neural network algorithms. Second, classification at the EGNE level (analogous to aspects in ABSA) shows better results than classification at the text level. This confirms the correctness of our assumption that attitudes towards ethnic groups are not expressed at the level of texts and, moreover, may differ for different groups in the same text, which does not allow the algorithm to classify the text as a whole. Third, detecting the three types of attitudes is more effective than looking only for hate speech. This is paradoxical, because classification tasks involving two classes are, as a rule, easier to solve than three-class tasks. This demonstrates the correctness of our strategy of moving away from the binary concept of hate speech (yes/no) to the concept of a general attitude, which is characterized by three, rather than two states (negative, positive and neutral). This distinction allows to avoid collapsing neutral and positive classes, which leads to better results. Fourth, important conclusions were obtained by analyzing errors in the operation of neural network algorithms. Fourth, important findings were obtained during error analysis of neural network algorithm performance. It was found that all algorithms perform better if the attitude towards EGNE is expressed with lexicon falling inside the context window, and worse if it is expressed indirectly (irony, jokes, use of negative stereotypes) and with complex syntactic structures (multiple negations, questions, anaphora). Overcoming these problems, as well as predicting other types of representations of ethnicity not covered by this project, may constitute a subject for future research.