Detecting ethnicity-targeted hate speech in Russian social media texts

Ekaterina Pronoza[1], Polina Panicheva[1], Olessia Koltsova[1], Paolo Rosso[2,1]

[1] Laboratory for Social and Cognitive Informatics, HSE University, Russia

{epronoza, ppanicheva, ekoltsova}@hse.ru

[2] PRHLT Research Center, Universitat Politècnica de València, Spain

prosso@dsic.upv.es

Highlights:

- We present a three-class instance-based approach to detect ethnicity-targeted hate speech in Russian social media texts;
- We show that ethnicity-targeted hate speech is more effectively addressed with the new three-class approach;
- In our task of instance-based ethnicity-targeted hate speech detection state-of-the-art deep learning models, while consistently outperforming classical machine learning models despite a relatively small dataset size, significantly benefit from a combination of linguistic and sentiment features with BERT pre-training and certain fine-tuning techniques;
- Deep learning models significantly benefit from specific ethnonym information added to text representation in instance-based ethnicity-targeted hate speech detection;
- We are making the **RuEthnoHate** dataset containing 5,5K social media texts, the first dataset annotated with ethnicity-targeted hate speech in Russian, available to the research community.

Abstract. Ethnicity-targeted hate speech has been widely shown to influence on-the-ground inter-ethnic conflict and violence, especially in such multi-ethnic societies as Russia. Therefore, ethnicity-targeted hate speech detection in user texts is becoming an important task. However, it faces a number of unresolved problems: difficulties of reliable mark-up, informal and indirect ways of expressing negativity in user texts (such as irony, false generalization and attribution of unfavored actions to targeted groups), users' inclination to express opposite attitudes to different ethnic groups in the same text and, finally, lack of research on languages other than English. In this work we address several of these problems in the task of ethnicity-targeted hate speech detection in Russian-language social media texts. This approach allows us to differentiate between attitudes towards different ethnic groups mentioned in the same text – a task that has never been addressed before. We use a dataset of over 2,6M user messages mentioning ethnic groups to construct a representative sample of 12K instances (ethnic group, text) that are further thoroughly annotated via a special procedure. In contrast to many previous collections that usually comprise extreme cases of toxic speech, representativity of our sample secures a realistic and, therefore, much higher proportion of subtle negativity which additionally complicates its automatic detection. We then experiment with four types of machine learning models, from traditional classifiers such as SVM to deep learning approaches, notably the recently introduced BERT architecture, and interpret their predictions in terms of various linguistic phenomena. In addition to hate speech detection with a text-level two-class approach (hate, no hate), we also justify and implement a unique instance-based three-class approach (positive, neutral, negative attitude, the latter implying hate speech). Our best results are achieved by using fine-tuned and pre-trained RuBERT combined with linguistic features, with F1-hate=0.760, F1-macro=0.833 on the text-level two-class problem comparable to previous studies, and F1-hate=0.813, F1-macro=0.824 on our unique instance-based three-class hate speech detection task. Finally, we perform error analysis, and it reveals that further improvement could be achieved by accounting

for complex and creative language issues more accurately, i.e., by detecting irony and unconventional forms of obscene lexicon.

1. Introduction

Rapid growth of social media has been contributing to proliferation of user content that contains judgements on groups or individuals based on their ethnicity. Speech expressing negative ethnicity-targeted judgements has been described in literature with a number of related concepts, such as hate speech, prejudiced or stereotypical speech, offensive or abusive language, uncivil or harmful language and others, and a variety of definitions of those have been proposed (for overviews, see [Niemann et al., 2019, Siegel, 2019, Haas, 2012]). Importantly, many forms of such speech have been shown to contribute to offline intergroup tensions and intergroup conflict [Williams et al., 2019, Müller & Schwartz, 2019], notably in such multi-ethic societies as Russia [Bursztyn et al., 2019] and in a broader Post-Soviet space torn apart by contradictions between more than a hundred ethnic groups. This explains the growing interest of researchers in the methods of detection and prevention of such speech forms [Warner & Hirschberg 2012, Gitari et al., 2015, Van Hee et al., 2015, Tulkens et al., 2016].

Of all the listed forms of negative speech, hate speech has been one of the major focuses in computational linguistics [Basile et al., 2019, Zampieri et al., 2019, Zampieri et al., 2020], although hate speech targeted specifically at ethnic groups has received only very modest attention [Gitari et al., 2015, Tulkens et al., 2016]. The concept itself is far from having a single definition [Fortuna & Nunes, 2018]. In computational linguistics, it is often defined as "any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics" [Nockleby, 2020]. This definition refers, first, to an overgeneralization based on group membership (stereotyping) and, second, to the treatment of specific group members as inferior. As observed by Fortuna & Nunes [2018] in their review of hate speech detection literature, practical definitions by different for-profit organizations or public bodies usually include the incitement of violence [Wigand &

Voin, 2017] and / or usage of language that attacks or diminishes groups such as ethnic minorities, as stated in the Facebook [Facebook, 2013] and Twitter [Twitter, 2017] practical definitions. Fortuna & Nunes propose to broaden these definitions by the inclusion of statements that have any negative bias against certain groups, even if they are expressed in subtle forms.

The following examples from our dataset illustrate these considerations. In Example (1), there is clearly hate speech present towards Baltic nations. However, Germans and Russians are also mentioned in the text with apparently no hateful attitude.

> *(1) Анатолий, Он сидит где нибудь в **Литве** и нагнетает.Ненавижу **прибалтов** нация обиженных тварей ни,как не могут смириться,что их еб@ли,как **немцы** так и **русские**.*
>
> *Anatoly, he is sitting somewhere in **Lithuania** and forcing the discussion. I hate **Balts**, they are a nation of resentful bitches, they still can't come to terms with the fact that they were f@cked, both by **Germans** and by **Russians**.*
>
> *(2) У меня мужчина **Азербайджанец**. Мы с ним уже давно вместе и знаете я ни сколько не жалею что всё таки он у меня есть. Полюбила не за внешность а за отношение к себе. Они действительно умеют любить. И ещё. В то время как **русские** бухают и работать не хотят. Люди другой национальности. Вертятся и добиваются многова*
>
> *My boyfriend is **Azerbaijani**. We have been together for a long time and I never regret that we are. I fell in love with him not because of his looks, but because of his attitude towards me. They can really love. And one more thing. While **Russians** booze and do not want to work. People of other nationalities. They keep trying and achieve a lot.*

Example (2) is, on the other hand, not so explicit in terms of hate speech. However, it contains a generalized statement of **Azerbaijani** boasting particular positive characteristics, in contrast with Russians, who are reported as "boozing and unwilling to work". This is a typical example of a generalized attitude towards

ethnicities in our dataset. In such cases, drawing a distinction between *explicit hate speech* and *negative attitude implying hate speech* is a difficult and often subjective task. As we find it important to include numerous examples like this in the hate speech class, we adopt the broader definition of hate speech.

In fact, our previous research on Russian-language social media [Koltsova et al., 2017a,b] supports the broader definition of hate speech above. For instance, outgroups are often treated as non-inferior, but hostile, dangerous, responsible for or causing certain problems or just guilty of being different. While in Russian-language social media inferiority is often ascribed to Central Asians, Caucasians (meaning those living in the Caucasus) are commonly described as both superior and aggressive [Bodrunova et al., 2017]. Furthermore, around a half of ethnicity-relevant social media texts mention more than one ethnic group [Koltsova et al., 2018], and those are often contrasted to each other as good to evil. In such cases presenting some ethnic groups as superiors or as victims implies others being seen as inferiors or aggressors without explicitly stating it. Thus, hate speech can also be present when there are only defensive statements or declaration of pride, rather than attacks directed towards a specific ethnic group [Warner & Hirschberg 2012]. In other cases, hate speech is expressed by indirect ways involving irony and sarcasm [Bosco et al., 2018]. We believe that these subtle forms of discrimination in online social media should be also considered. All this requires, first, broadening the definition of hate speech, and second, an approach that allows discriminating between judgements targeting different ethnic groups within the same text.

Recently, a few researchers have addressed the problem of abusive language detection in the Russian language [Andrusyak et al., 2018; Smetanin, 2020; Zueva et al., 2020]. However, these works are aimed at a general task of abusive speech classification. As a result, their authors do not analyze either the concept of abusive hate speech towards a specific target, or the respective annotation process. Moreover, their results are not generalizable to cases involving different targets within the same text.

In this work, we aim at detecting ethnicity-targeted hate speech in Russian-language social media using state-of-the-art deep learning models. We broaden the definition of hate speech following the above-

mentioned work of Fortuna & Nunes [2018] and their idea to account for subtle negative bias against certain groups. We thus define *ethnicity-targeted hate speech* as the speech expressing *negative attitude* towards an ethnic group or its individual based solely on their ethnic status. We classify attitude in ethnicity-targeted texts into three classes: (*negative, positive, neutral*), with the negative class implying the broader notion of hate speech.

Next, since we often have multiple targets of ethnicity-based speech in our texts, we adopt a different unit of analysis: the instance of ethnicity-targeted speech, represented by a pair *(ethnic group, text)*. *Ethnic group*, in turn, can be represented by one or more coreferent *ethnonyms*. For each ethnic group mentioned in the text, we solve the *three-class instance-based* classification task.

We construct a corpus with balanced proportions of different post-Soviet ethnic groups and with nearly real-life class distribution - that is, a corpus embracing not only extreme or pure cases, but also mixed, mild, subtle and contradictory types of speech which is especially difficult to predict. As to date there have been no attempts to solve ethnicity-targeted hate speech detection task for the Russian language, apart from a few works of our team [Koltsova et al., 2017a,b; Koltsova et al., 2018], this corpus is the first and so far the only marked-up collection for such task. We then detect ethnicity-targeted hate speech by classifying attitude towards ethnic groups with different machine learning approaches, ranging from traditional classifiers to deep learning models, including Long Short-Term Memory (LSTM) [Hochreiter & Schmidhuber, 1997] and state-of-the-art Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al., 2018], by first fine-tuning it on the target training texts as in [Mollas et al., 2020], then also pre-training it on in-domain texts as in [Wiedemann et al., 2020], and finally by combining its output with different sets of linguistic features.

In order to account for specific ethnic groups in the instance-based hate speech detection task, we leverage the natural language inference capabilities of BERT-based models [Sun et al., 2019, Hoang et al., 2019]: specifically, we construct an auxiliary sentence from ethnonyms denoting a single ethnic group in question, and treat the instance-based hate speech detection problem as a sentence-pair classification task.

By performing the experiments in *ethnicity-targeted hate speech detection*, we seek to answer the following research questions (RQs):

RQ1. Should ethnicity-targeted hate speech be addressed as a two-class (*hate/no hate*) or three-class (*negative/neutral/positive attitude*) problem? Specifically, is the underlying structure of ethnicity-targeted speech better described with two (*hate/no hate*) or three (*negative/neutral/positive attitude*) classes?

RQ2. Can instance-based hate speech detection benefit from a sentence-pair classification approach, namely, by adding specific ethnonym information as an auxiliary sentence into BERT?

RQ3. Can deep learning models benefit from linguistic features in hate speech detection?

The contributions of this study into the domain of hate speech detection are the following:

- To the best of our knowledge, our research is the first study of hate speech in the Russian language targeted at ethnic minorities;
- In contrast to all previous studies of hate speech which were designed as text-level two-class tasks, we show that ethnicity-targeted hate speech should be addressed with the instance-based three-class approach including negative, neutral and positive attitudes (**RQ1)**;
- We find that instance-based ethnicity-targeted hate speech detection performance significantly benefits from including ethnic information into the input text representation, namely, by adding specific ethnonym information as an auxiliary sentence into BERT (**RQ2**) which was never applied to this type of task;
- We provide detailed evidence demonstrating that in instance-based ethnicity-targeted hate speech detection, state-of-the-art deep learning models, while consistently outperforming classical machine learning models, significantly benefit from a combination of linguistic and sentiment features with BERT pre-training and an additional dense layer, but not from linguistics features separately (**RQ3**).
- Finally, we are making available to the research community the **RuEthnoHate** dataset containing 5,5K social media texts, the first dataset annotated with ethnicity-targeted hate speech in Russian.

The rest of the paper is structured as follows. In Section 2 we describe related work on hate speech detection, ethnic relations research in Russian, and relevant sentiment analysis techniques, and draw some conclusions situating our approach in terms of the related work. In Section 3 and 4 we present our dataset and methodology. The results of our experiments on ethnicity-targeted hate speech detection are illustrated in Section 5. In Section 6 we discuss the results and provide error analysis, and in Section 7 we conclude the study.

2. Related work

2.1 Hate speech detection

Hate speech may be divided into into a number of categories either by speech type, such as blackmail, insult, curse, defense, defamation and encouragement [Van Hee et al., 2015] or hate target: gender, race, national origin, disability, religion and sexual orientation [Mollas et al., 2020]. Hate speech detection problems typically include the following:

- Binary hate speech detection (hateful / non-hateful text)
- Classifying degrees of hate (strong hateful/weak hate/none)
- Classifying different categories of hate.

The problem of online hate speech detection has been widely studied in computational linguistics, and there exist a substantial number of hate speech corpora collected from both social media platforms (Twitter, Facebook, You Tube, Reddit, Formspring) and specific political websites and forums. The corpora were mostly constructed for English [Basile et al., 2019, Zampieri et al., 2019, Mollas et al., 2020, Rosenthal et al., 2020, Waseem & Hovy 2016, Davidson et al., 2017], but there also exist Spanish [Basile et al., 2019], German [Struß et al., 2019], Polish [Ptaszynski et al., 2019], Portuguese [Fortuna et al., 2019], Italian [Sanguinetti et al., 2018], Greek [Pitenis et al., 2020], Danish [Sigurbergsson et al., 2020], Dutch [Van Hee et al., 2015], Arabic [Mubarak et al., 2020] and Korean [Moon et al., 2020] hate speech datasets, to name

a few (for a review see [Poletto et al., 2020]). To the best of our knowledge, the only corpus close to hate speech in Russian is Russian Language Toxic Comments dataset [Belchikov, 2019].

Hate speech detection has been addressed in a number of recent shared tasks, mostly organized for English and other European languages [Basile et al., 2019, Zampieri et al., 2019, Struß et al., 2019, Ptaszynski et al., 2019, Aragón et al., 2019], and also Arabic, Hindi, and Turkish [Mandl et al., 2019, Zampieri et al., 2020, Mubarak et al., 2020]. Shared tasks on hate speech and offensive language detection typically consist of several subtasks where the teams have to 1) classify texts into hate speech/offensive or not, and 2) classify hate speech/offensive texts into targeted (i.e., hateful) and untargeted ones [Mandl et al., 2019, Zampieri et al., 2020, Mubarak et al., 2020], among other things. Sometimes hate speech has also to be classified into those targeting one person or a group of people [Ptaszynski et al., 2019], or a categorization into different types of hate targets is required [Zampieri et al., 2019, 2020, 42. Mandl et al., 2019]. A variety of techniques are used to solve these tasks. A different approach was adopted by SemEval'2019 Task 5 (HatEval) organizers [Basile et al., 2019] where the targets were already specified (women and immigrants), and the task was, firstly, to detect whether texts are hateful towards these targets or not, and, secondly, to classify hateful texts into aggressive and non/aggressive ones and into targeting one particular person or a group of people. Our approach is close to that of HatEval, in that potential targets of hate (ethnic groups) are already known, and the task is to identify hate speech towards them.

Existing hate speech detection methods usually involve either traditional machine learning, with the best results obtained by lexical features and elaborate feature engineering [Warner & Hirschberg, 2012, Dinakar et al., 2012, Gitari et al., 2015, Van Hee et al., 2015, Tulkens et al., 2016, 35, Davidson et al., 2017], or deep learning algorithms (CNNs, LSTMs and GRUs, and pre-trained Transformers [Mikolov et al., 2013, Mehdad and J. Tetreault 2016, Badjatiya et al., 2017, Del Vigna et al., 2017, Fortuna & Nunes 2018, Zhang et al., 2018, Wullach et al., 2020, Mollas et al., 2020, Moon et al., 2020]). Modern deep learning-based approaches typically use word embeddings as text representation (e.g., Word2vec [Mikolov et al., 2013], fasttext [Bojanowski et al., 2016, Corazza et al., 2020], ELMo [Peters et al., 2018], GloVe [Pennington et al., 2014], BERT [Devlin et al., 2018]). Transfer learning and multitask learning have been reported to

improve the overall quality of the models [Wiedemann et al., 2020, Abu-Farha & Magdy 2020, Elmadany et al., 2020].

A few works perform linguistic analysis and evaluate the effect of different linguistic phenomena on automatic hate speech detection [Cimino et al., 2018, Corazza et al., 2020]. They show that, given high-quality word-embedding representations in deep learning models, neither emotion lexica, nor special processing of specific word categories (e.g., emojis) make a significant contribution to hate speech detection performance.

Hate speech detection approaches are evaluated with traditional classification metrics (Recall, Precision and F1, accuracy and ROC AUC scores). However, most studies present F1 of "hateful" class as the focus metric [Warner & Hirschberg, 2012, Gitari et al., 2015, Van Hee et al., 2015, Tulkens et al., 2016, Basile et al., 2019, Zampieri et al., 2019, Zampieri et al., 2020, Waseem & Hovy, 2016, Mehdad & Tetreault, 2016, Badjatiya et al., 2017, Davidson et al., 2017], as the quality of the "hateful" class detection is most important.

A brief overview of the existing hate speech detection approaches is provided in Appendix 1. The results in terms of "hateful" F1 vary significantly between 0.46 and 0.95, and depend largely on the following issues:

- Problem formulation: hate speech towards a single specific target (race, gender, religion, nationality, sexual orientation) is typically identified better than hate speech towards multiple or unspecified targets;
- Methods: deep learning models usually yield higher scores than traditional ones;
- Size of dataset: models tend to perform better when trained on larger datasets;
- Class imbalance: better results are achieved on balanced datasets [Yuan et al., 2016] or by applying augmentation techniques in case of classes imbalance [Elmadany et al., 2020, Kapil et al., 2020];
- Topic bias of dataset [Poletto et al., 2020];
- Language.

2.2 Hate speech and ethnic attitude in Russian

Although in the past few years studies dedicated to the research on hate speech detection for languages other than English have emerged, studies on hate speech detection in Russian remain very scarce. We are aware of no research on ethnicity-targeted hate speech, except a few works by our team, and of at best four papers, to a varying degree related to other types of hate speech.

Smetanin [2020] offers a solution for a two-class toxic speech detection using a Russian-language dataset from Kaggle [Belchikov, 2019] and obtains F1=0.92 with a RuBERT-based model. The solution addresses the task of general toxicity detection, not involving any specific target. Moreover, in this dataset texts marked as abusive are usually heavily loaded with obscene words, which is not at all always the case in ethnicity-targeted hate speech. Andrusyak et al. [2018] address a task of general abusive language detection in mixed Russian-Ukrainian sociolects, which are, strictly speaking, not the Russian language. Zueva et al. [2020] propose to organize drop-out of the words denoting objects of hate to increase the performance of generally defined hate speech detection. However, they demonstrate a lower performance than that by Smetanin both on Belchikov's dataset, Ansrusyak's dataset, and the authors' artificial dataset (0.78-0.86 in terms of F1). Finally, an unpublished work on sexism detection described on GitHub[1] achieves its best result applying LSTM with pre-trained and fine-tuned ELMO embeddings (0.74 in terms of balanced accuracy score – this metric was used to mitigate the classes imbalance problem).

In our previous project, we explored a broader task: detecting different types of attitudes to a variety of ethnic groups in Russian social media [Koltsova et al., 2017a,b]. Although ethnicity-targeted hate speech is usually target-specific, the task was formulated so as to accommodate a multitude of ethnic groups due to specific traits of ethic conflicts in the Post-Soviet space. This space is populated by a very large number of ethnic groups many of whom are very small and relatively rarely discussed, but taken together messages devoted to them contribute a lot to hate speech, which is often multi-directional and contains more than two parties. As a part of this research, a large corpus of text potentially mentioning ethnic groups from the

[1] https://ansable.github.io/sexism_detection_in_russian/

Russian social media was collected, and a smaller part of it (7K) was annotated with such concepts as interethnic conflict, call for violence, superiority/inferiority of an ethnic group, and some others. General attitude towards ethnic groups (positive, negative, and neutral/contradictory) was also a part of the annotation, and a simple Logistic Regression with tf-idf weighted unigrams and bigrams was trained to classify it. Attitude classification results were quite poor (F1-macro = 0.58).

Having analyzed our previous results, we can state the following specific traits of the task of ethnicity-targeted hate speech identification in Russian:

1. *Multi-target and contrastive character of texts.* In Russian social media 50% of messages mention more than one ethnic group, and 21% contain opposite attitudes towards these [Koltsova 2018]. Moreover, claims against different groups may vary in content and the respective wordings.
2. *Non-binarity.* Binary ("hate - no hate") approach leads to information loss: openly positive attitudes, including declaration of pride and exclusiveness of certain nations, play an important role in downplaying the significance of other ethnic groups, especially when explicit comparisons are contained in the same text. Thus a three-class approach appears to better reflect the structure of ethnicity-targeted speech.
3. *Unnecessary presence of obscene lexicon.* While most datasets for hate speech detection include the most extreme and unequivocal cases for classification purposes, the real-life data seen in the Russian social media varies a lot from subtle to intermediate to explicit attitude towards ethnic groups.

2.3 Aspect-based sentiment analysis techniques

As our approach involves instance-based techniques for hate speech detection, it is related to aspect-based sentiment analysis (ABSA). In these, sentiment towards various aspects of the entity is taken into account when classifying sentiment towards an entity as a whole. ABSA is particularly relevant to our task, as it typically involves contradicting sentiment towards different aspects of an entity in a single text or even a single sentence. Thus, ABSA requires fine-grained instance-based analysis, which text-level classification

is incapable of. The same is true for our ethnicity-targeted data, where numerous ethnicities can be characterized by contrasting attitudes in a single text or sentence.

The high granularity in ABSA is obtained by either intensive contextual feature engineering [Wagner et al., 2014, Kiritchenko et al., 2014, Saeidi et al., 2016], or modifying deep learning methods to take aspect information into account. This is done with memory networks [Tang et al., 2016b], interactive attention networks [Ma et al., 2017] or LSTM with attention mechanism [Wang & Lu, 2018, Gu et al., 2018]. In [Sun et al., 2019] ABSA was treated as a natural language inference task: an auxiliary sentence was constructed from an aspect and ABSA was converted to a sentence-pair classification task. Then a pre-trained BERT model was fine-tuned for this task. In our work, we also adopt this technique for our hate speech detection task.

In Russian, similar techniques to ABSA have been developed in the following works: in [Karpov et al., 2016], dependency parsing output was integrated into a convolutional network, and Word2vec word embeddings pre-trained on a large in-domain corpus were used as input to the network. The best results in [Karpov et al., 2016] were achieved by a hybrid approach combining CNN-based method with the rule-based one (F1-macro = 0.538 for banks domain and F1-macro = 0.527). In [Arkhipenko et al., 2016], the problem of detecting sentiment towards different aspects of banks and telecommunication systems in Russian tweets was solved using a LSTM/GRU network with Word2vec CBOW embeddings as an input layer, demonstrating the best results (F1-macro = 0.552 for banks domain and F1-macro = 0.559 for telecommunications domain) in SentiRuEval'2016 shared task [Loukachevitch & Rubtsova, 2016]. These results were improved in [Golubev & Loukachevitch, 2020] by a BERT-based classification approach, where Russian pre-trained Conversational BERT was applied to the same problem (F1-macro = 0.795 for banks and F1-macro = 0.684 for the telecommunications dataset).

These results indicate that:

- ABSA problems can be effectively solved as natural language inference tasks with BERT models;

- Conversational RuBERT overcomes Common RuBERT in Russian ABSA, and by far overcomes other standard deep learning approaches (LSTM, CNN, BiLSTM) and traditional classifiers (SVM).

2.4 Our approach

In the current work, we build on the aforementioned considerations and focus on obtaining high quality in detecting ethnicity-targeted hate speech.

- First, we focus on a broad definition of hate speech as a negative attitude towards a group or an individual. The hate speech detection problem is solved by **detecting attitude in ethnicity-targeted speech** in Russian language texts, including **positive**, **negative** and **neutral** attitude;
- Our corpus is constructed by annotating 12K ethnic group instances in the messages from Russian social media and is representative of real-life data, in that we do not specifically avoid subtle, vague or intermediate cases;
- We focus on **instance-based hate speech detection**, whereas we identify hate speech towards every ethnic group mentioned in the text, and **compare it with a binary approach** classifying the **presence of hate speech towards** at least one ethnic group in the text;
- We analyze the contribution of the ABSA approach to instance-based hate speech detection by adding specific ethnonym information to LSTM-based models, and as an auxiliary sentence in BERT-based sentence-pair classification;
- We evaluate the results of **instance-based ethnicity-targeted hate speech detection** with F1 for the negative attitude class, F1-macro and average weighted F1;
- We experiment with **pre-trained deep learning models** (LSTM, Conversational RuBERT) with state-of-the-art word embeddings; we also examine the impact of careful **linguistic feature engineering** on the quality of instance-based hate speech detection.

3. Data

Our dataset has been formed in several steps, some of which were performed in our previous research.

1. We formed a list of ethnonyms based on the Russian Census [2010] and other sources. It represents a nested array of 115 Russian and post-Soviet ethnic groups, where each group is represented by a list of unigrams and bigrams (e.g. “Jew”, “Jewish girl”, “Jewish nation” etc), including ethnophaulisms (ethnic slurs) and pseudo-ethnicities (Caucasian, Asian).
2. We obtained a collection of all messages containing at least one ethnonym from our list ever posted on all Russian language social media during two years (from January 2014 to December 2015). Having been purchased from a commercial social media aggregator, IQBuzz[2], this collection turns out to be composed mainly (by 80%) of messages from Vkontakte, a replica of Facebook and the most popular social network in Russia. After filtering out duplicates the collection numbered 2,660,222 messages; hereafter this dataset is referred to as the **RuEthnics** dataset.
3. Next, we formed our first collection for annotation which was substantially smaller than RuEthnics. As the distribution of ethnic groups in the dataset was very unbalanced, we over-represented infrequent groups based on manually-derived balancing quotas adopted for each ethnic group. We also limited message length to the range [20; 90] words. In all other respects, the sampling was random. This ensured realistic class distribution in the collection.

   Each text was annotated by at least three independent specially trained annotators who were asked to select answers for a list of questions, including the filtering questions about text interpretability and the presence of ethnonyms. Among other things, this resulted in adding instances of ethnonyms that had not occurred in our initial list.

   As their main task, the annotators were asked to annotate the overall attitude of the text author to the ethnic group or one individual (negative/neutral/positive) making special emphasis on negative

[2] https://iqbuzz.pro/

ones that implied hate speech. The main question sounded as follows: "what is the overall attitude of the text author to the ethnic group or its representative?" (negative/neutral/positive).

This initial annotated collection comprised 14,998 texts as described in more detail in [Koltsova et al., 2017a]. Previously it was used for text-level attitude prediction [Koltsova et al., 2017b]. However, the reported quality of attitude classification towards ethnic groups was modest (F1-macro = 0.58). Therefore, we have substantially modified this collection for the current research.

4. For this, out of 14,998 texts we obtained 27,165 attitude instances (ethnic group, text) and then selected 11,067 instances on which at least two annotators agreed. Our negative class comprised 12% of the sample with 1,365 instances.
5. To increase the quality of the dataset, we enriched our sample by the following steps (see [Hernandez et al., 2013] for discussion): We trained a set of simple classifiers on 11,067 instances and obtained the best precision by Gradient Boosting (GB) with n-gram and linguistic features. It was then run on the full RuEthnics dataset. From the negative instances identified by GB we randomly selected 985 instances, according to the balancing quota approach. For these, we repeated the entire procedure of the annotation. The instances were added to our dataset, the annotators' labels being used as ground truth. The proportion of the negative class has increased by 5%, although the annotators disagreed with the classifier in 31% of cases. Statistics of the initial, enriched and final datasets are presented in Table 1.

Our final dataset contains 5,594 texts and 12,052 instances (ethnic group, text), of which 2,040 instances are negative (representing the hate speech class), 8,697 are neutral and 1,315 are positive. We will refer to this dataset as **RuEthnoHate** dataset[3].

The distribution of the frequencies of all ethnic groups, including those by class, is also available at our project webpage[4] and is power-law in both RuEthnic and RuEthnoHate collections, despite the

[3] **RuEthnoHate** is available at https://scila.hse.ru/data/2021/05/25/1438275158/RuEthnoHate.zip. Extended version of **RuEthnoHate** including annotators' disagreements is available at https://scila.hse.ru/data/2021/05/25/1438273746/RuEthnoHateExtended.zip.

[4] https://scila.hse.ru/data/2021/03/05/1398220409/Ethnic-stats.xlsx.

overrepresentation of small ethnic groups in the latter. While RuEthnic distribution naturally results from the activity of social media users, RuEthnoHate's distribution has become more uneven after assessors have marked up all ethnic groups, including those that had not been used for the formation of RuEthnoHate collection. As a result, the most frequent ethnic groups gained more mentions, while the tail of the distribution received a large number of rarely occurring ethnicities from outside the Post-Soviet space.

The frequency of mentions of an ethnic group does not correlate with the share of the instances of hate speech towards it. The largest shares of hate speech are predictably observed towards groups denoted with ethnophaulisms (mind that the boundaries of ethnophaulisms' meaning often do not correspond to specific ethnic groups). Apart from collective ethnonyms (Asians, Caucasians) and the infrequent ethnonyms that are likely to be statistical outliers, the five "true" ethnic groups with the highest shares of hate speech, in the descending order, are Ukrainians, Americans (=USA Americans), Jews, Gypsies, and Azerbaijanis. In the absolute numbers this list of leaders includes Russians and Chechens instead of Americans and Gypsies.

Table 1. Datasets

Dataset	№ instances in classes						
	Negative		Neutral		Positive		Total
Initial	1,365	12%	8,480	77%	1,222	11%	11,067
Enrichment	675	69%	217	22%	93	9%	985
Attitude dataset (Initial + Enrichment)	2,040	17%	8,697	72%	1,315	11%	12,052

4. Methodology

To solve the task of ethnicity-targeted hate speech detection, we adopt the following strategy. Hate speech detection is performed in two settings:

1. Binary attitude detection (**BAD**): a *text-level* approach, where each text is classified as hateful/negative or non-hateful. The binarization procedure is organized as follows: if the text

contains hate speech towards at least one of ethnic groups mentioned in it (i.e., "negative" label), it is labeled as hateful, otherwise non-hateful.

2. Instance-based attitude detection for specific ethnic groups (**IBAD**): an *instance-level* approach, where a pair *(ethnic group, text)* becomes our *instance* of analysis. *Ethnic group*, in turn, can be represented by one or more coreferent *ethnonyms*. For each ethnic group mentioned in the text, we solve a three-class classification task: the attitude of the author towards the ethnic group is classified as positive, neutral or negative (implying hate speech).

Consider the following example from our dataset:

(3) ***грузинские*** *бл*ди всегда отличались подлостью и трусостью совсем не давно изза одной* ***грузинской*** *сучки тренером* ***россии*** *вольной борьбе по имени когоушвилли двое* ***чеченцы*** *не поехали в чемпионат мира по вольной борьбе именно* ***грузинская*** *сучка гвишиани приказала заживо сожжать в конюшне 705* ***чеченских*** *женшин и стариков именно* ***грузинская*** *шлюха сталин выслала* ***чеченцев*** *и других* ***кавказских*** *народов именно* ***грузинская*** *дочь сучки берия уничтожила много* ***чеченских*** *архивных исторических документов и как после этого не еб*ть эти черножопых родственников* ***тюрков****?*

Georgian *whores have always stood out by their meanness and cowardice. Just recently because of a* ***Georgian*** *bitch, a wrestling coach of* ***Russia*** *named Kogoushvilli, two* ***Chechens*** *have missed the world wrestling championship. It is the* ***Georgian*** *bitch Gvishiani who commanded to burn 705* ***Chechen*** *women and elderly people alive in a barn. It is a* ***Georgian*** *bitch Stalin who deported* ***Chechens*** *and other* ***Caucasian*** *peoples. It is a* ***Georgian*** *daughter of a bitch Beria who exterminated many* ***Chechen*** *archives and historical documents. How can one but f*ck these black-ass relatives of* ***Turks****?*

Ex. (3) clearly contains hate speech towards the Georgian people and individuals. According to the **BAD** approach, there is hate towards certain ethnic group(s) present in the text. In contrast, the **IBAD** approach allows us to identify hate towards Georgians, positive attitude (based on compassion) towards Chechens, and neutral attitude towards Russians and Turks, who are only mentioned in passing in the argument. **IBAD** would thus enable preserving information about the contrasting attitudes towards different ethnic groups. Thus, in the **IBAD** approach this text would be represented by 4 different instances: (Georgians, text), (Chechens, text), (Russians, text) and (Turks, text).

Our target evaluation metrics in both settings are F1 for negative class (**F1-hate**), as we are mostly interested in detecting the negative class. We also calculate average weighted F1 (**F1-ave**) and macro-averaged F1 (**F1-macro**). **F1-macro** is calculated as an unweighted mean F1 across classes, thus treating the classes as balanced and resulting in bigger penalization of minority class errors, including the *negative attitude* class. **F1-ave** is different from **F1-macro** in that it takes into account the classes distribution as weights in the mean calculation, whereas each class is represented proportionally.

We use traditional machine learning approaches and feature engineering as a baseline. We apply deep learning techniques and supplement these with linguistic features, to obtain high quality of hate speech detection towards ethnic groups.

4.1 Classical machine-learning models

As baselines, we use the following classifiers: Naive Bayes (**NB**, baseline), Logistic Regression (**LR**), Support Vector Machine (**SVM**) and Voting Classifier (**VC**), the latter being essentially an ensemble of **NB**, **LR** and **SVM**.

Our linguistic features are as follows:

- Word **unigram** features (155,081 features);

- Counts of emoticons (one feature for positive emoticons + one feature for negative emoticons), exclamation marks (one feature), total number of words in text (one feature), words in capital letters (one feature);
- The following features from the context window (size = +-3) of the target *ethnonyms* (only used in **IBAD**):
    - Negative polarity words from the PolSentiLex sentiment dictionary [Koltsova et al., 2020], as sentiment is considered an important feature in hate speech detection [Fortuna, Nunes, 2018] (884 features);
    - Character n-grams, with n in range [2, 3, 4] (28,306 features);
    - POS n-grams, with n in range [1, 2, 3] (1,418 features);
    - Word n-grams, with n in range [2, 3] (127,055 features).

Thus, we used 312,748 features for the IBAD representation. The texts were lemmatized with PyMorphy2, and all the word features apply to normal word forms (lemmas). No stop words removal, frequent or rare words removal was conducted because in a series of preliminary experiments keeping all the words led to better performance of the models.

Context window size = 3 is only reported, as other window sizes result in similar or lower performance. We did not carry out any optimization procedures for our baseline models and used their default configuration from scikit-learn[5] implementation.

4.2 Deep learning models

We experiment with LSTM and GRU models, and feed them with pre-trained word embeddings as input. We select LSTM/GRU as a second baseline because it was previously shown to be the best solution in Russian ABSA [Arkhipenko et al., 2016]. In our **IBAD** setting which is closely related to the ABSA task we also experiment with the third baseline - a series of state-of-the-art ABSA models including MemNet

[5] https://scikit-learn.org/

[Tang et al., 2016b], attention-based LSTM [Wang et al., 2016], interactive attention networks [Wang et al., 2016], and some others - using their open-source implementation[6]. Finally, we fine-tune a state-of-the-art Russian BERT model for ethnically targeted hate speech detection. While ABSA models are used with the default parameters from their implementation, the hyperparameters for all the other deep learning models result from a careful selection based on several runs.

4.2.1 Word embeddings for LSTM/GRU

For LSTM/GRU we use three types of pre-trained word embeddings:

- Word2vec CBOW [Rehurek & Sojka, 2010]:
    - **Word2vec-RNC**: provided by the Webvectors project [Kutuzov & Kuzmenko, 2016] trained on the Russian National Corpus;
    - **Word2vec-Ethno**: trained on **RuEthnics** (2,6M messages);
- Conversational RuBERT (**RuBERT-emb**): word embeddings based on the Multilingual BERT model by Google and pre-trained by DeepPavlov[7] [Kuratov & Arkhipov, 2019] on social media texts from OpenSubtitles [Lison & Tiedemann, 2016], Dirty, Pikabu, and Social Media segment of the Taiga corpus [Shavrina & Shapovalova, 2017]).

Common RuBERT and Conversational RuBERT are multilingual initializations of BERT [Devlin et al., 2018] trained on Russian datasets and shown to improve performance over multilingual BERT in a variety of NLP tasks in Russian [Kuratov & Arkhipov, 2019]. Both previous works and our preliminary experiments have shown that Conversational RuBERT gives a more accurate representation of our data than Common RuBERT trained on written texts in Russian. Therefore, we only report Conversational RuBERT-based results. Based on Google's Multilingual BERT-base, Conversational RuBERT naturally inherits its configuration parameters, such as maximum sequence length of 512 tokens, 12 attention heads, 768-dimensional token vectors. Word2Vec-RNC had a fixed set of parameters. For Word2Vec-Ethno, we

[6] https://github.com/AlexYangLi/ABSA_Keras
[7] http://deeppavlov.ai/

ran several preliminary experiments optimizing vector dimension parameters and selected the value of 200 dimensions.

The main characteristics of the selected embeddings are compared in Table 2.

Table 2. Embeddings characteristics

	Word2Vec-Ethno	**Word2Vec-RNC**	**RuBERT-emb**
Size (dimensions)	200	300	768
Vocabulary	200K	270K	120K
Preprocessing	uncased, lemmatized with POS tags	uncased, lemmatized with POS tags	cased

4.2.2 Deep learning models with LSTM and GRU layers

Our deep learning architecture is as follows. The model consists of LSTM and GRU layers (**LSTM+GRU**), with hard sigmoid and sigmoid activation functions respectively. The model is trained for 20 epochs with Adam optimizer and categorical cross-entropy loss. As a result of the preliminary experiments, the LSTM layer size is 50 for **Word2vec-Ethno** and **Word2vec-RNC** embeddings and 200 for **RuBERT-emb**, and a dropout rate of 0.7 is selected for the LSTM layer to prevent the model from overfitting.

Word2vec-Ethno and **Word2vec-RNC** embeddings were updated during training, while **RuBERT-emb** were frozen (no fine-tuning was done at this stage). For **Word2vec-Ethno** and **Word2vec-RNC** we used lemmatized uncased versions of texts, with reversed order of words in texts (an approach shown to be helpful in [Arkhipenko et al., 2016]).

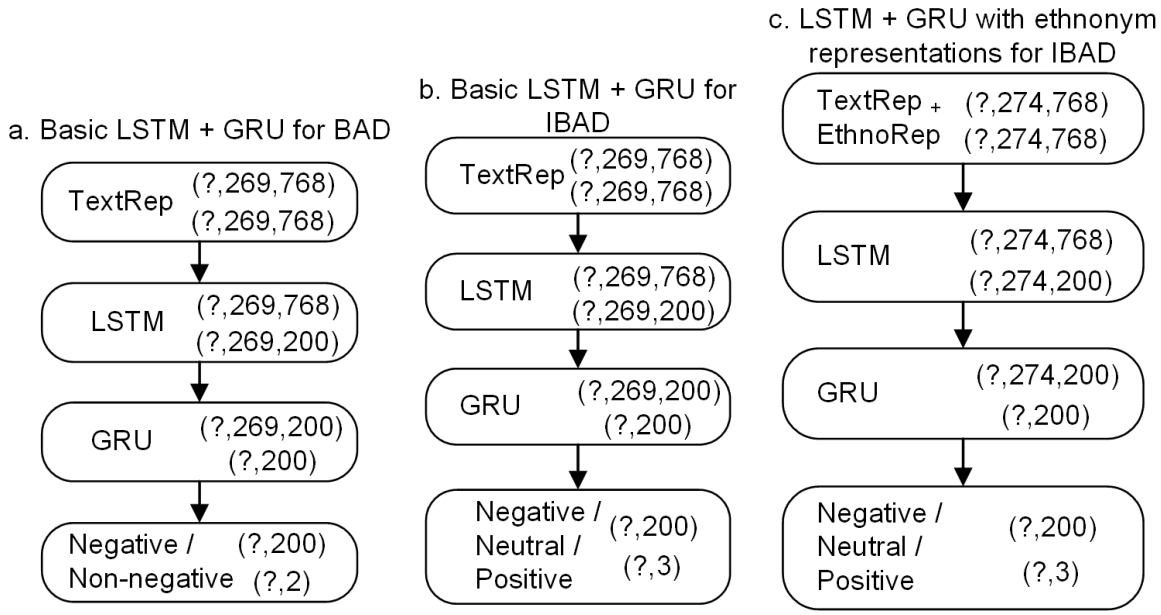


Figure 1. LSTM+GRU architectures for ethnicity-targeted hate speech detection.

The deep learning architectures employed are illustrated in **Figure 1**. In **BAD**, we use the basic **LSTM+GRU** architecture, where the whole text is represented as the input (**TextRep**), Fig. 1a. In **IBAD**, this architecture (Fig. 1b) results in an oversimplification, as the output attitude towards all the *ethnic groups* mentioned in the text is the same. To overcome this issue and specify *ethnic-group*-related information in **IBAD**, for each *instance* (represented by the pair (*ethnic group, text*)), we insert up to 5 ethnonym representations (**EthnoRep**) referring to the target ethnic group at the beginning of the input, followed by **TextRep** (Fig. 1c).

4.2.3 State-of-the-art ABSA deep learning models

We use state-of-the-art ABSA models as our third baseline in the **IBAD** setting. These models include

- Content Attention Model (Cabasc) [Liu et al., 2018],
- Recurrent Attention Network on Memory (RAM) [Chen et al., 2017],
- Interactive Attention Network (IAN) [Wang et al., 2016],
- Deep Memory Network (MemNet) [Tang et al., 2016b],
- Attention-based LSTM [Wang et al., 2016],

- Target-dependent LSTM [Tang et al., 2016a].

For each of the models listed above, we tried both fixed and trainable word and aspect embeddings. As for the other parameters, we used the default ones[8]. Since this implementation relies on the GloVe vectors trained on the Common Crawl data[9] which are not available for the Russian language, we substitute them with fastText vectors trained on the Russian part of GeoWac corpus [Dunn & Adams, 2020] which consists of Russian-language documents from Common Crawl. The fastText vectors are provided by the Webvectors project [Kutuzov & Kuzmenko, 2016]. They are 300-dimensional, uncased and non-lemmatized[10].

4.2.4 Conversational RuBERT model

We experiment with the following Conversational RuBERT (**Convers-RuBERT**) architectures. The **Convers-RuBERT** model is used with sequence length = 256 (covering 99.6% of our texts).

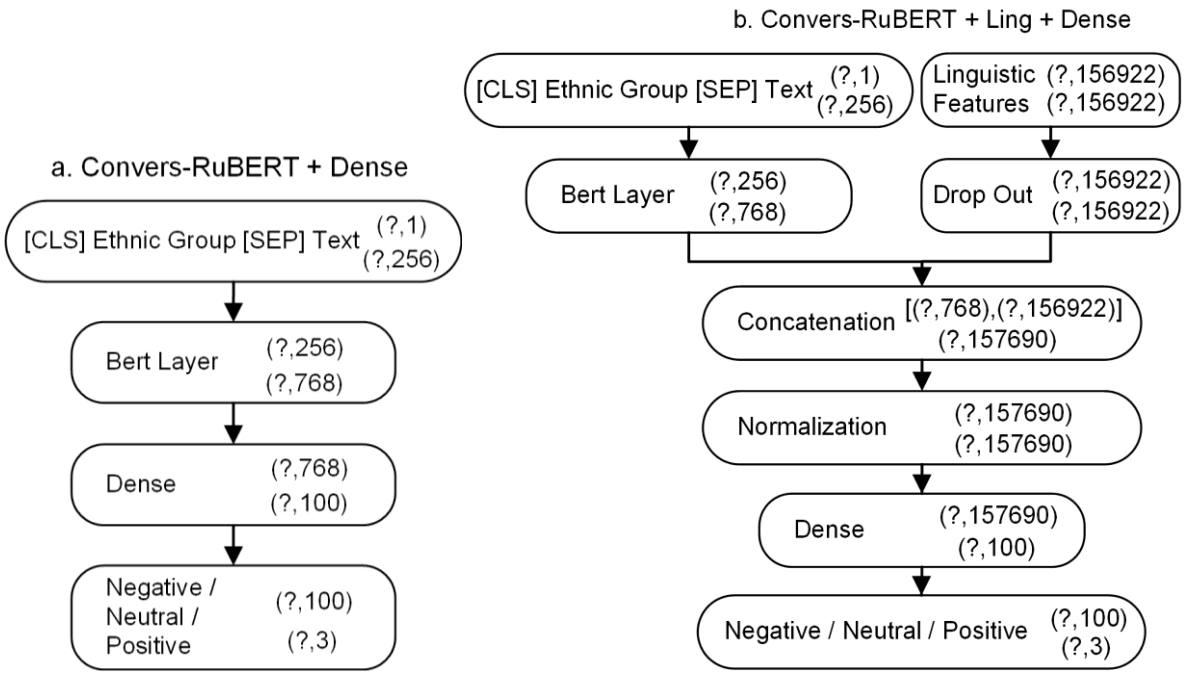
In **BAD**, we again apply the basic architecture by adding an output dense layer with sigmoid activation on top of the pre-trained **Convers-RuBERT (Convers-RuBERT+Dense)**.

In **IBAD**, we treat the attitude detection task as a natural language inference task by specifying the *ethnic group* information in an auxiliary sentence, followed by the text representation in the second sentence (**EthnicGroup+Text** representation). We apply sentence pair classification architecture in our task: the input to BERT consists of two sentences, where the first sentence is an ethnic group representation, while the second one is the text mentioning the ethnic group. The resulting **Convers-RuBERT** model is leveraged in the following architectures, illustrated in Figure 2:

- A Dense classification layer (sizes = {30, 50, 100}) is added to RuBERT (**Convers-RuBERT+Dense**, Fig. 2a).

[8] as implemented in https://github.com/AlexYangLi/ABSA_Keras
[9] https://commoncrawl.org/
[10] like the original GloVe ones from https://github.com/AlexYangLi/ABSA_Keras

- RuBERT output is concatenated with 157,667[11] linguistic features (see Section 3.1), with the concatenation followed by a dense layer (**Convers-RuBERT+Ling+Dense**, Fig. 2b).
- RuBERT output is fed into a LSTM layer (sizes={100, 200}), which is concatenated with the linguistic features, with the concatenation followed by a dense layer (**Convers-RuBERT+LSTM+Ling+Dense**, Fig. 2c)
- Additional dense layer (sizes = {30, 50, 100}) is added to architectures 1-3 (+**Dense2**).



[11] The full list of linguistic features used in Convers-RuBERT models includes 157667 features; however in Fig. 2b-c a smaller number of features is shown (156,922). Since in our experiments we applied 10-fold CV, we learned features only from the training part of the data at each of the 10 folds thus resulting in a smaller amount of features.

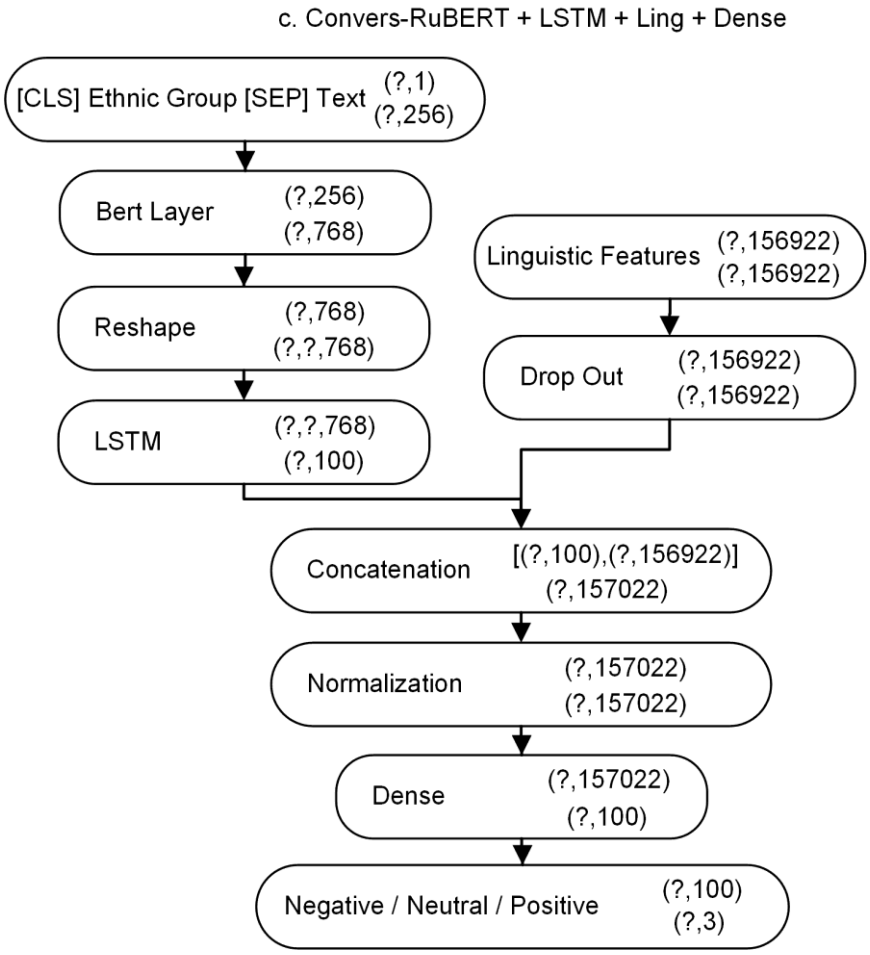


Figure 2. Convers-RuBERT architectures for instance-based ethnicity-targeted hate speech detection.

We experimented with the following RuBERT parameters:

- the number of fine-tuned layers (1-4);
- output consisting of one layer or a concatenation of several layers (one layer performed better);
- pooling strategy ([CLS] token, mean vector, no pooling);
- BERT output layer number (last or second-to-last, where the last performed better);
- Pre-training Conversational RuBERT on **RuEthnics**.

5. Results of ethnicity-targeted hate speech detection

We report the results for ethnicity-targeted hate speech detection (with **BAD** and **IBAD** techniques) obtained with 10-fold cross-validation. Significance of the difference between the results is calculated with

the Mann-Whitney U-test [Mann & Whitney 1947]. The experiments were performed with the following libraries in Python: *keras* [Chollet et al., 2015], *scikit-learn* [Pedregosa et al., 2011], *scipy* [Virtanen et al., 2020] and *tensorflow* [Abadi et al., 2016].

5.1 Binary hate speech detection

The best results for **BAD** by traditional classifiers, **LSTM+GRU** and **Convers-RuBERT** are shown in Table 3. The best result is highlighted in bold.

Table 3. Binary hate speech detection (BAD) results

Models	Features	Parameters	F1-hate	F1-ave	F1-macro
NB	**Unigrams**		0.701	0.828	0.790
LSTM+GRU	**RuBERT-emb**	dimensions = 200	0.736[a]	0.851[a]	0.816
Convers-RuBERT+Dense	**RuBERT-emb**	**mean pooling, concatenation 1-4**	**0.760[a]**	**0.864[a]**	**0.833[a]**

a - significant difference from the **NB** baseline ($p < 0.01$)

Surprisingly, among the traditional classifiers the most simple technique, **NB** with word unigram features performed the best in **BAD**. However, both **LSTM+GRU** and **Convers-RuBERT** models with mean pooling and concatenation of the last four layers have significantly outperformed **NB**. At the same time, there was no significant difference between **LSTM+GRU** and **Convers-RuBERT** model performance in the binary approach to hate speech detection.

The results of **BAD** are comparable to the results achieved by other binary hate speech detection approaches on the datasets of similar size. The closest setting to our experiment is presented by the HatEval task at SemEval-2019. The best model in our task scored F1-macro=0.833, while the highest results at HatEval for Spanish was F1-macro=0.73 (a dataset of 6.6K tweets) and for English F1-macro=0.65 (a dataset of 13K tweets).

5.2 Instance-based ethnicity-targeted hate speech detection

5.2.1 Machine learning models

Results of **IBAD** with machine learning models are presented in Table 4. As Voting Classifier (**VC**) has typically outperformed the other approaches (**NB**, **LR** and **SVM**), we only report the results for **VC**. The best result is highlighted in bold.

Table 4. Instance-based hate speech detection (IBAD) results with machine learning models

Run	Feature set	F1-hate	F1-ave	F1-macro
0	**Baseline** (lemma unigrams)	0.614	0.822	0.696
1	0 + Negative polarity words (from context [-3; 3])	0.628	0.825	0.702
2	1 + Character n-grams (from context [-3; 3])	0.690[a]	0.839[a]	0.733[a]
3	**2 + POS n-grams, Lemma n-grams (context [-3; 3]), Emoticons, Exclamation marks, words count, words in capital letters counts**	**0.702[a]**	**0.842[a]**	**0.734[a]**

a - significant difference from the lemma unigrams baseline ($p < 0.01$)

Only character n-gram features have improved the performance of the model significantly against lemma unigrams ($p < 0.001$ for F1-neg and $p < 0.007$ for F1-ave). But since all of the feature sets added up to the overall F1 scores, we further use the full feature set (run 3) as a contribution to our further experiments with deep learning approaches.

We conducted feature importance analysis for the full (run 3) feature set. Since our model is a voting classifier and it is thus unfeasible to obtain feature importance scores for it, we did it for Logistic Regression which appeared to be the best one out of the three models inside VC (SVM, LR and NB). The top-30 most informative features in LR for the hateful class are presented in Table 5. It can be seen that the most informative features are lemma unigrams expressing negative emotions, they are either slurs or ethnophaulisms (the latter ones are denoted by "*" in Table 5 and also in this paragraph). Char n-grams are clearly parts of ethnonyms ("зер" is a part of both normal and ethnophaulistic names for Azerbaijani, while "ец" is a typical suffix of ethnonyms in Russian, e.g., in "ногаец" / Nogai, and "сс" can be a part of ethnonyms as well, e.g., in "русский" / Russian). Interestingly, pronouns also appear among the most informative features for ethnic hate speech detection. Indeed, hateful texts often represent accusations and threats starting with "you", e.g., "*Михаил, ты азербот* сдаешь себя тем что стоишь за азерботов**

так как русские их ненавидят, а ты готов ху.. их сосать" ("*Mikhail, you, an Azerbaijani*, you give yourself away by standing for the Azerbaijani* because the Russians hate them, and you are ready to suck their d*cks*")

Table 5. Top-30 most important features for hate detection with Logistic Regression (IBAD setting). Ethnophaulisms are marked with *.

Feature type	Feature	Coefficient
Lemma unigram	жид* (Jew)	1.478
	черножопый* (black ass)	1.436
	мразь (scum)	1.143
	чурка* (black ass)	0.840
	ты (you)	0.790
	азер* (Azerbaijani)	0.775
	свинья (pig)	0.724
	х** (d*ck)	0.712
	укр* (Ukrainian)	0.664
	вы (you)	0.632
	хохол* (Ukrainian)	0.620
	узкоглазый* (narrow-eyed)	0.618
	есть (eat)	-0.612
	а (but)	0.601
	азербот* (Azerbaijani)	0.597
	тупой (stupid)	0.555
	этот (this)	0.551
	национальность (nationality / ethnicity)	-0.540
	брат (brother)	0.537
	хохлов* (Ukrainians)	0.535
	нерусский (non-Russian)	-0.533
	хохлушка* (Ukrainian (female))	0.527
Lemma ngram	какой то (some)	0.581
	чурка и* (black ass and)	0.574
Char ngram	зер	0.663

	ец_	-0.575
	ц_	-0.571
	ец	-0.569
	сс	-0.549
Punctuation mark	! (exclamation_mark)	0.557

5.2.2 Deep learning models

Results of the most prominent **IBAD** experiments with deep learning models are presented in Table 6. The best results for the **LSTM+GRU** model were obtained with **RuBERT-emb** word embeddings. As for the ABSA models (our third **IBAD** baseline), we are reporting the best results which were achieved by MemNet with trainable word and aspect embeddings. The overall best result is highlighted in bold.

Table 6. Instance-based hate speech detection (IBAD) results with deep learning models

Run	Model	Architecture	F1-hate	F1-ave	F1-macro
1	**LSTM+GRU**	TextRep	0.670[a]	0.834	0.727[a]
2	**LSTM+GRU**	EthnoRep + TextRep	0.732[a,c]	0.853[a,b,c]	0.750[a,b,c]
3	**MemNet**	See [Tang et al., 2016]	0.732[a,c]	0.849[a,b,c]	0.752[a,b,c]
4	**Convers-RuBERT+Dense**	Text, size = 100, mean pooling, fine-tuned 4 last layers	0.785[a,b,c,d,j,f]	0.877[a,b,c,d,j,f]	0.797[a,b,c,d,j,f]
5	**Convers-RuBERT+Ling+Dense+Dense2**	Text, pre-trained, layer sizes = 100, mean pooling, fine-tuned 4 last layers	0.732[a,c]	0.860[a,b,c,g]	0.768[a,b,c,e,g]
6	**Convers-RuBERT+Ling+Dense+Dense2**[12]	**EthnicGroup+Text, pre-trained, layer sizes = 100, mean pooling, fine-tuned 4 last layers**	**0.813[a,b,c,d,f,i,j]**	**0.892[a,b,c,d,f,h,j]**	**0.824[a,b,c,d,f,h,j]**
7	**Convers-RuBERT+LSTM+Ling+Dense+Dense2**	EthnicGroup+Text, pre-trained, layer sizes = 100, mean pooling, fine-tuned 4 last layers	0.813[a,b,c,d,f,i,j]	0.889[a,b,c,d,f,i,j]	0.820[a,b,c,d,h,f,j]

a - significant difference from the **Baseline** (Table 4 run 0, $p < 0.01$)

b - significant difference from fine-grained features with **VC** (Table 4 run 3, $p < 0.01$)

c - significant difference from **LSTM+GRU** (**TextRep**) (Table 6 run 1, $p < 0.01$)

d - significant difference from **LSTM +GRU** (**TextRep+EthnoRep**) (Table 6 run 2, $p < 0.01$)

e - significant difference from **LSTM +GRU** (**TextRep+EthnoRep**) (Table 6 run 2, $p < 0.05$)

[12] Detailed training configuration of our best performing model can be found in Appendix. 4

f - significant difference from MemNet (Table 6, run 3, $p < 0.01$)

g - significant difference from MemNet (Table 6, run 3, $p < 0.05$)

h - significant difference from fine-tuned **Convers-RuBERT** (Table 6 run 4, $p < 0.01$)

i - significant difference from fine-tuned **Convers-RuBERT** (Table 6 run 4, $p < 0.05$)

j - significant difference from fine-tuned **Convers-RuBERT** with Text representation only (Table 6 run 5, $p < 0.01$)

Based on the results reported in Table 6, the following conclusions about the models performance can be made:

- **LSTM+GRU** model with word embeddings as text representation is comparable to classical machine learning models with hand-crafted linguistic features;
- Adding ethnonym representation to both **LSTM+GRU** and **Convers-RuBERT** models significantly increases their performance;
- **MemNet** (state-of-the-art model for ABSA) performs at the same level as **LSTM+GRU** with ethnonym representation;
- Fine-tuning **Convers-RuBERT** with **EthnicGroup+Text** representation outperforms other models (**LSTM+GRU**, classical machine learning models and state-of-the-art ABSA models);
- Linguistic features (including sentiment and other contextual information), additional pre-training on in-domain data and an additional dense layer further increase the performance of the **Convers-RuBERT model** significantly.

In terms of F1-macro, our best run in the three-class **IBAD** approach to hate speech detection scored 0.824. Finally, one of our assumptions was that the underlying structure of ethnicity-targeted hate speech is better described with three (positive, negative, neutral) rather than two (hate, non-hate) classes. Indeed, the assumption has been confirmed: hate speech detection with the three-class **IBAD** approach resulted in higher performance in terms of F1-hate: 0.813 (**IBAD**) against 0.760 (**BAD**).

6. Error analysis & Discussion

To interpret the results of instance-based hate speech detection, we manually analyzed the errors of our model performance in the instance-based attitude detection approach, as it performed better than the binary attitude-detection approach. Additionally, to demonstrate the contribution of ethnic information included in the input representation of the deep learning models, we performed a deeper analysis of the difference between our best-performing BERT-based model with and without adding ethnonym information as an auxiliary sentence.

In the analysis of the model errors, first the errors were identified automatically against the ground truth, as part of the model evaluation (see Section 4. Methodology). Next, we annotated the errors manually in terms of their linguistic nature, similar to the qualitative analysis of hate speech detection errors performed by Corazza et al. (2020).

6.1 Error annotation

We performed a second annotation as part of the error analysis, in order to further understand the linguistic phenomena causing errors.

First, a sample output of the models was selected randomly from our test datasets (each test fold in the 10-fold cross-validation). We selected four representative models for our analysis: the best traditional classifier (**VC**, run 3 from Table 4), the best **LSTM+GRU** run (run 2 from Table 6) and two runs of BERT: run 4 (**Convers-RuBERT+Dense**) and the best run 6 from Table 6 (**Convers-RuBERT+Ling+Dense+Dense2**). Having four models, we resulted in 2^4 possible error combinations of correct/incorrect predictions. For each of these 16 types of combinations, we randomly selected three instances (one for each of the three true classes: positive, negative and neutral) from each of the ten test folds. Since our dataset is unbalanced, for some error combinations there were not all the three classes present in the examples. In this case in our error sample we gave preference to the negative class, implying hate speech. The resulting error annotation dataset consists of 473 instances: 182 negative, 176 neutral and 115 positive instances.

Second, we engaged four annotators: two researchers in Social Informatics (one PhD, one BSc-level), and two in Computational Linguistics (one PhD, one MSc-level). The task included annotating each instance represented by a pair *(ethnic group, text)* in the sample output with labels of relevant linguistic phenomena present in the corresponding text. More than one label could be assigned to one instance. The labels and the examples of texts representing them are given in Table 6. The total set of labels with their brief description is presented in Appendix 2. Each annotator also assigned a specific mark to each instance indicating their agreement with the initial class annotation.

Next, we calculated the inter-annotator agreement in the second annotation between the label sets assigned by four annotators (see Appendix 3). Krippendorff's alpha was applied. The agreement was generally medium (0.39-0.56) between the pairs of annotators, with the exception of two computational linguists reaching the agreement of 0.81. The annotators agreed on the correct class labels with each other and with the golden annotation only in 67% cases. This additionally confirms the complexity of both interpreting the author's attitude towards ethnic groups and detecting specific linguistic phenomena in text.

6.2 Error analysis

Taking into account the complexity and low agreement of our annotation, we have selected the frequent labels which, when applied by at least one annotator to an instance, were agreed upon by at least three of the four annotators in more than 50% of the cases. We report statistics for these labels in Table 6. For each of the four selected models (**VC**, **LSTM+GRU**, **Convers-RuBERT+Dense**, **Convers-RuBERT+Ling+Dense+Dense2**) we present the percentage of cases, where the respective model gave an incorrect prediction. In other words, the percentages indicate how difficult the current linguistic phenomenon is for a specific classifier.

The preliminary error sample analysis has shown that there is no strong variation in model performance across different labels. However, there are some tendencies.

Firstly, the traditional classifier (**VC**) is worse at detecting hate speech towards ethnic groups when the group is referred to as a **noun phrase**, than neural networks. It can be explained by the fact that VC uses

one-hot encoding of words and phrases and ethnicity-based features are extracted using a dictionary of ethnonyms, whereas in LSTM and BERT approaches words are represented by their embeddings. The latter obviously allows the models to identify ethnic groups represented by more complex means than by single-word ethnonyms.

Table 6. Error labels statistics and examples (in Russian and English translation)

Label	VC	LSTM +GRU	Convers-RuBERT+ Dense	Convers-RuBERT+Ling+ Dense+Dense2	Example
actions	58	52	48	52	Исторически **чувашский народ** никогда не конфликтовал, жил в мире и согласии. / Historically, the **Chuvash people** have never been in conflict, they lived in peace and quiet.
caps	40	56	32	48	ТУПЫЕ **РУСАКИ** НЕ ЗНАЮТ, ЧТО ЭТО ФЛАГ ЧЕЧНИ И НАЗЫВАЮТ ЕГО ТУРКМЕНСКИМ/КЫРГЫЗСКИМ/ТАТАРСКИМ ФЛАГОМ / STUPID **RUSSAKS** DON'T KNOW THAT THIS IS THE FLAG OF CHECHNYA AND CALL IT THE TURKMEN/KYRGYZ/TATAR FLAG
context_neg	49	41	46	49	**якуты** бл* кричат что Киеву нельзя быть с Европой.... этож каким нада быть е*нутым **якутом**)) / The **Yakuts** f*cking cry that Kiev can't be with Europe… what a f*cked **Yakut** this must be!
contrast	65	60	73	40	Мы воюем против Нового **Израиля**. Я отдаю себе отчет в том, что наш враг не украинцы, а **еврейские** олигархи и их американские кураторы (с) / We are fighting against the New **Israel**. I am aware that our enemy is not the Unkranians, but the **Jewish** oligarchs and their American supervisors (c)
discussion	42	35	60	60	Так значит, и **русский народ** такой плохой, т.к. там тоже есть плохие, не воспитанные люди. Но и в Туве, и у **Русского народа** есть талантливые люди, которыми можно гордиться. / So, the **Russian people** is so bad then, because it includes bad, rude people. But in both Tuva and the **Russian people** there are talented people, which one can be proud of.
ethnophaulism	48	49	47	56	ага и ещё вспомним на чей стороне воевали **мамалыжники** в ВОВ!!! / Yeah, let us now remember, on which side the **mamalyzhnicks** fought in WWII!!!
exclam	51	45	45	55	
irony	53	57	57	60	Безумно популярным могло бы стать приложение Яндекс.**Хачи**", сообщающее актуальные данные о плотности кучкования кавказцев на станциях метро / The Yandex.Khachi application could become insanely popular, providing up-to-date data on the clustering density of **Caucasians** clumping together at metro stations У **русских** привычка считать что победили они))" / "**Russians** have a habit of thinking that they won))
noun phrase	64	45	56	47	Исторически **чувашский народ** никогда не конфликтовал, жил в мире и согласии. / Historically, the **Chuvash people** have never been in conflict, they lived in peace and quiet.
obscene	54	50	56	54	**якуты** бл* кричат что Киеву нельзя быть с Европой.... этож каким нада быть е*нутым **якутом**)) / The **Yakuts** f*cking cry that Kiev can't be with Europe… what a f*cked **Yakut** this must be!
other_neg	49	55	55	54	Виталий, да мало ли что кто сказал ? бред !!!! русские сами разваливают свою страну...даже с простого начнём- кто срёт на природе после отдыха ? засерая озёра и реки ? кто засерает всё вокруг себя ? америкосы и **жиды** ..всё это херня" / "Vitaly, but do you care what somebody said? nonsense !!!! Russians themselves are destroying their country ... take for a example, who litters in nature on a vacation? littering lakes and rivers? who will litter everything around him? Americans and **Jews** .. this is bullshit
strong	50	36	41	48	а самое смешное **чечены** пишут постоянно мол вы русские нация алкашей и наркоманов. мне лично писали всегда. а сами то)) наркобароны, еще и русских сажают. гниды черножопые / the funny thing is, the **Chechens** always write: you, Russians, are a nation of alcoholics and junkies. they wrote this to me personally always. Look at yourselves)) drug lords, and they put Russians on drugs. black-ass nits

Secondly, **VC** is better for texts with contrasting opinions where both positive and negative lexicon is used (**other_neg**). Indeed, **VC** mainly focuses on the nearest context words of ethnonyms. When several ethnic groups are mentioned in the text, and hate speech towards another ethnicity is used in the text, **VC** appears to be better than neural network-based models (**LSTM+GRU** and **Convers-RuBERT**).

Thirdly, texts where hate targeted at ethnicities is expressed with negative lexicon in the nearest context of ethnonyms (**context_neg**) or with strong negative sentiment words (**strong**) are easier to classify for all models, than texts where such attitude is expressed in a more implicit way: describing **actions** attributed to ethnic groups, or using **ironic** language. Hate expressed with **obscene** words is also relatively difficult for the models to detect, perhaps, due to the lack of creative derivatives of obscene words and expressions in the sentiment dictionary and in word embeddings.

Finally, **Convers-RuBERT** enhanced with linguistic features and pre-training appears by far the best model at detecting ethnicity-targeted hate speech represented by a **contrast** to another ethnic group.

Indeed, a number of hate speech examples in **RuEthnoHate** demonstrate a certain level of creativity, making the hate speech subtle and difficult to classify. In Ex. (4) below, the author first expresses ironic admiration and gratitude towards his interlocutor and generalizes their attitude over the Ukranian people. Then the author proceeds to mutual derogatory comments between Russians and Ukranians, ironically promising that the Russians will meet the derogatory expectations.

> *(4) Эдик, ну ты и впрям великий **Укр**)). Прям такой умный, важный, уважаемый, богатый, великодушный. Спасибо Вам, за то что учитете нас **ватников** и **азиатов**. Спасибо Вам за Мир, за человечество и Чёрное море. Мы **кацапы** будем вести себя хорошо и газ вам будем давать бесплатно. А Путина мы выгоним в Ростов, пусть там сидят тираны пенсионеры. Обещаем сжечь все покрышки и обосать все дворы Москвы. Обещаем как и вы встать раком и смазать зад и ждать друзей из **Америки**. Нет ну не бесплатно, пусть нам за это кредиты дают.*

*Ed, you are really a great **Ukr**)). So clever, influential, respected, rich, generous. Thank you for teaching us, **vatniks** and **Asians**. Thank you for Peace, for humanity and for the Black sea. We, **katsaps**, will behave well and will give you gas for free. We will send Putin away to Rostov, to other tyrants and retirees. We promise to burn all the tires and piss in all the yards in Moscow. We promise to get down on all fours, like you, lubricate our ass and wait for our friends from **America**. Not for free, let them give us money loans for that.*

*(5) И это только одно из тысяч предприятий, которые понастроили озверелые **русаки** на вильних зэмлях нэньки. Оголтелая, вечно пьяная и немытая **русня**, смесь **татар** и **мордвы**, загадила **Украинушку** своими фабриками, заводами, школами, больницами, электростанциями и многоэтажными жилыми домами. Не было предела этой циничной, бесчеловечной оккупации. Народ был запуган и порбащён. Но справедливость восторжествовала! Майдан!!!!!!! Скинули кляте ярмо! Долой душные панельки! Назад, к природе, к мазанкам и копанкам. Слава **Уркаине**.*

*And this is a single enterprise out of thousands, which were built by the outraged **Rusacks** on the free lands of nenka* [misspellings imitating the Ukranian language]. *Unbridled, eternally drunk and dirty **Rusnya**, a mix of **Tatars** and **Mordva**, has shitted the beloved **Ukraine** with its factories, plants, schools, hospitals, power stations and multi-storey houses. The cynical, inhuman occupation was unlimited. The people were terrified and enslaved. But justice has prevailed! Maidan!!!!!!! The damn yoke is thrown off* [misspellings imitating the Ukranian language]*! Get rid of the stuffy multi-storey houses! Back to nature, to wattle and daub huts. Long live **Ukraine*** [originally: U**rk**aine, a derogatory misspelling reminiscent of "Urki", a Russian slang name for "jailbirds"].

Example (5) imitates accusations of Russians by Ukranians in an exaggerated and ironic manner, in effect demonstrating a strong negative attitude towards Ukranians, also mockingly imitating the Ukranian language and pronunciation and using offensive misspellings.

The examples above demonstrate that there are texts in our corpus containing highly complex hate speech instances, not only targeting different ethnic groups with opposite attitudes, but also containing irony, imitative and derogatory misspellings, unconventional forms of obscene lexicon and references to specific historical actions and political statements. These findings confirm the linguistic error categories identified by Corazza et al. (2020) in English, German and Italian hate speech detection, involving implicit abuse: sarcasm, jokes, the usage of negative stereotypes, or supposedly objective statements implying some form of offense; and complex syntactic structure: more than one negation or questions, anaphoric elements referring to previous messages, and examples requiring external knowledge to be understood.

6.3 Effect of adding ethnic information to Convers-RuBERT

In this paper we address ethnicity-targeted hate speech detection, and the ethnic aspect is important in this task. Interestingly, the results show that both Convers-RuBERT and LSTM/GRU models benefit from including ethnic information in the input representation of the data. To confirm that ethnic representation plays an important role in our best-performing BERT-based model, we compared our best-performing BERT model **Convers-RuBERT+Ling+Dense+Dense2** with **EthnicGroup+Text** (see the best run 6 in Table 5) with the same model accounting for **Text** representation only (see run 5 in Table 5). Specifically, we obtained the highest results by adding an ethnic group to the input representation, thus making it a sentence-pair classification task instead of a single-sentence classification one.

The results demonstrate that adding specific ethnonym information as an auxiliary sentence and treating the problem as a sentence-pair classification task significantly improves the performance of instance-based ethnicity-targeted hate speech detection with BERT.

To obtain a deeper understanding of the improvement, we illustrate the performance of the **EthnicGroup+Text**- and **Text**-based models with their confusion matrices in Table 7.

Table 7. Confusion matrices for Convers-RuBERT+Ling+Dense+Dense2 models EthnicGroup+Text and Text only representations

Classified by: EthnicGroup+Text / Text			**Correct**
-1	**0**	**1**	
1,610/1,497	311/390	119/153	-1
283/390	8,110/8,029	307/281	0
75/160	275/313	970/847	1

From Table 7 it is obvious that adding the ethnonym information as **EthnicGroup+Text** representation to our best model is especially useful in correcting positive and neutral instances misclassified as hateful instances, i.e. it increases the Precision of the "hate" class classification. This reflects the intended effect of the added ethnonym representation: to discern the different (sometimes polarized) attitude towards the different ethnic groups mentioned in the same text. We demonstrate this case with examples 6-8, where the target instances (emphasized in bold) are correctly classified as positive or neutral by our best model with **EthnicGroup+Text**, and misclassified as hateful by the model with **Text** - evidently, because of the hate directed at other ethnic groups mentioned in the text or general hate towards an interlocutor (respective hateful passages are underlined):

*(6) Qa, а конкретно, кто вы по национальности? Вот я ,например, по матери-**Русский**,по отцу-Табасаранец. А вы,чать, <u>азербот</u> или еще какая <u>семитская блядь</u>? Ты даже не пренадлежишь к кавказской расе, ты <u>ебаный семит</u>. От тебя <u>воняет джудами</u>.*
*Qa, what is your nationality specifically? For example, I am **Russian** by my mother, Tabasaran by my father. Are you probably <u>Aserbot</u> [ethnophaulism for Aserbaijani] or another <u>Semitic bitch</u>? You don't even belong to the Caucasian race, you <u>fucking Semit</u>. You <u>reek of Jews</u>.*

*(7) Игорь, да кавказцы по натуре гораздо частней чем укры, у меня кумовья **армяне** уже 30 лет, друзья разных национальностей и все нормальные люди, но <u>нет у меня ни одного</u>*

друга из хохлов, только знакомые и те говно не люди, жрут русский хлеб и не довольны русскими, когда говорю валите к себе в укропию, так нет там плохо.

*Igor, yes, Caucasians are much more honest than Ukrs, my relatives are **Armenian** for 30 years, friends of various nationalities and all are normal people, but I have no friends among Khokhols [ethnophaulism for Ukranian], I only know some people and they are shit, They eat Russian bread and are unhappy with Russians, when I say 'go back to your Ukropia [derogatory misspelling for Ukraine], no, it is too bad there.*

*(8) Крутой, А что с вами в дебаты вступать. У вас вест мир не прав только русские правы. От росси батько отказался ,**казахи** нахер послали , и впредачу вы еще у Армян на косячили и считаете себя правыми.*

*Krutoy, why should I even bother discussing with you. For you, the whole world is wrong, only Russians are right. Bat'ko [informal for the president of Belorussia] has given you up, **Kazakhs** have fucked you, and you have also screwed up with Armenians and are still thinking you are right.*

At the same time, adding ethnic information to text representation in BERT increases the Recall of the "hateful" class by capturing specific patterns of relations between ethnic groups and other important tokens in the text. To illustrate such patterns captured by the model, we prepared visualization of the attention mechanism in our best-performing **Convers-RuBERT+Ling+Dense+Dense2** with **EthnicGroup+Text** model using the bertviz[13] tool. We selected hateful instances which **Convers-RuBERT+Ling+Dense+Dense2** with **EthnicGroup+Text** was able to predict correctly, while the same model with **Text** representation only misclassified them.

[13] https://github.com/jessevig/bertviz

In Figure 3, attention visualization for one of the selected hateful instances is presented using the "neuron view" mode from bertviz[14]. This view illustrates the flow of attention from the token on the left to the complete sequence of tokens on the right: tokens which are paying attention are shown in the left column, while the tokens being paid attention to are in the right column. The brighter the color of the square next to the rightmost tokens, the more attention is being paid to those tokens. The text can be translated as *"liu, ha-ha-ha a Ukrainian will be teaching me the history of the Armenians? did we hide behind the backs of the Turks? do you even understand what you said? And behind the back of the Russians…"*. As it can be seen in Figure 3, the [CLS] token which is known for capturing the core semantic information about our text does pay attention to the first sentence "хохол" (ethnophaulism for the Ukrainian) - the word denoting the target of the above comment. It also pays attention to the pronoun referring to the ethnic group in question ("ты" / "you"), and the nouns and pronouns referring to the other ethnic groups in the text ("турков" / "the Turks", "армян" / "the Armenians", and "меня" / "me" obviously referring to the Armenians). The text contains several mentions of different ethnic groups, and the author's attitude towards them is different: the author is hateful towards the Ukrainians but neutral towards the other ethnicities. Thus, to differentiate among these ethnicities it is important to pay attention to the ethnicity in the first one out of the two sentences ("хохол" / "the Ukrainian" in our case), since in this instance, the model is predicting attitude towards this ethnic group.

[14] There are disagreements in the NLP community about the value of attention diagrams for interpreting attention-based models [Jain & Wallace 2019; Wiegreffe & Pinter 2019]. However, we present the attention diagrams for illustration, rather than verification purposes.

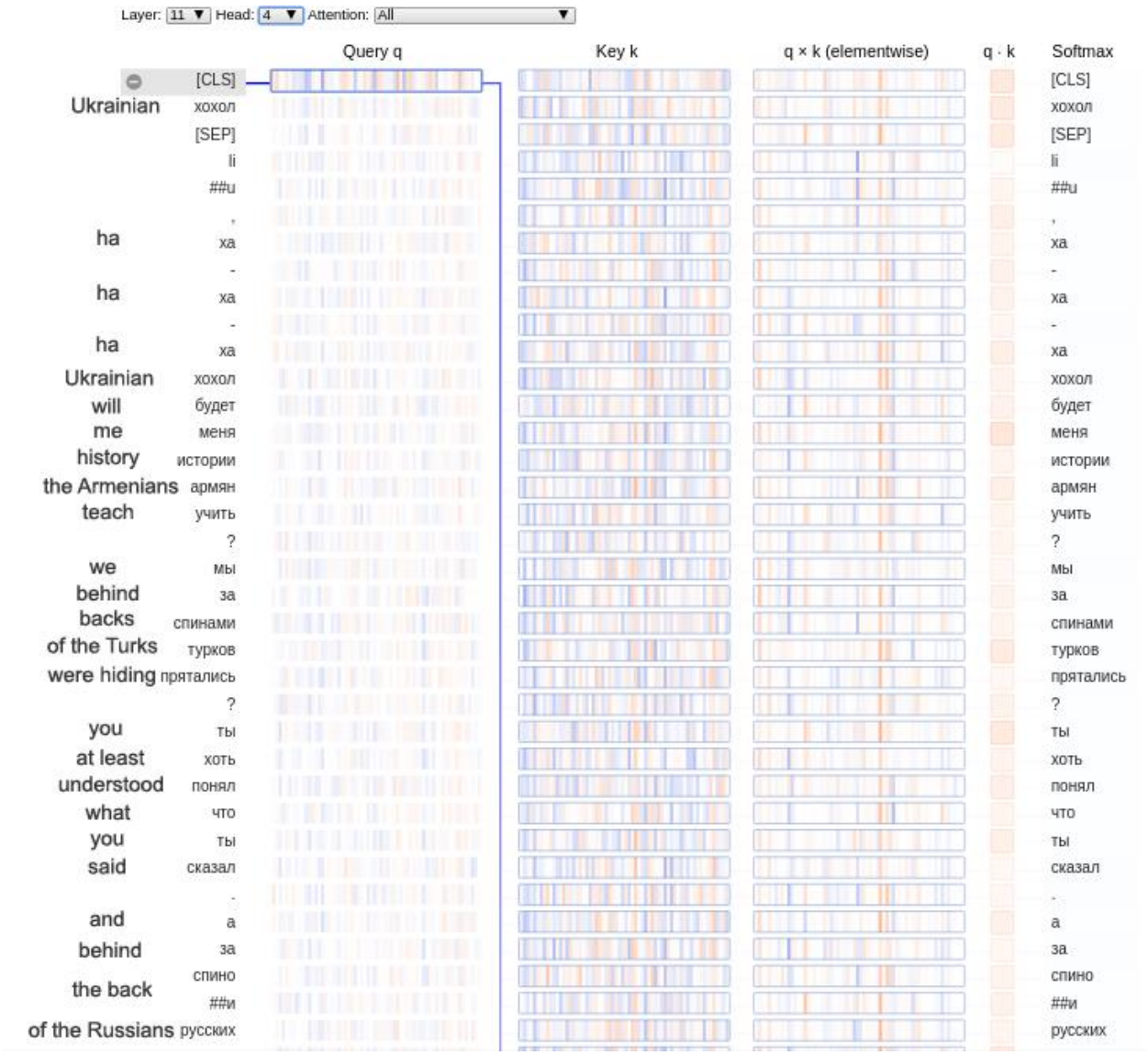


Figure 3. Visualization of attention in **Convers-RuBERT+Ling+Dense+Dense2** with **EthnicGroup+Text**: attention directed at "хохол" (ethnophaulism for the Ukrainian)

7. Limitations and ethical considerations

Despite the significant theoretical and experimental achievements, our study has a number of limitations, which should be taken into account when generalizing the results to ethnicity-targeted hate speech detection, also in other languages, and to detection of hate speech directed at other groups or individuals. First, the inter-annotator agreement at all stages of our work, including dataset annotation and error analysis,

was modest. Probably, this has to do with the fact that ethnicity-related hate speech is still an evolving notion with no clear boundaries, especially in cases where creative and unconventional language is used. It is also important to note that we are currently identifying ethnic groups in text with a simple lexical approach: including more complex cases of ethnic group mentions is a separate task, which obviously has to be solved in a real-life scenario, and it is beyond the scope of this paper. Finally, we completely miss cases of ethnicity-targeted speech where the targets are not mentioned explicitly - either due to co-reference, metonymy or other indirect indications of ethnic groups.

Our study has some important ethical considerations as well. First of all, the models automatically detecting hate speech should by no means be used to stigmatize the authors. These tools should only be applied in addition to, and not in replacement of, expert judgement. Current work on hate speech detection is primarily aimed at obtaining scientific insights into the diverse phenomena of hate speech, not at automatic penalization of authors in social media. Second, neither this research is aimed at stigmatizing hate speech targets; the availability of hate speech examples in our publications and in our dataset does not imply our agreement with the judgements of hateful authors. Nor our negative attitude towards certain ethnic groups is implied in the cases that were marked up as non-negative by annotators, but may sound negative to some of our readers. As we are making our dataset public, we believe that the best ways to avoid the listed stigmatization dangers are (1) to restrict its use for research purposes only, (2) to anonymize authors and (3) to make available initial diverse annotations instead of classes, in order to fully illustrate the disagreement and the complexity of the issue.

8. Conclusions and future work

In this work, we aimed at detecting ethnicity-targeted negative attitudes, implying hate speech, in Russian social media texts.

To achieve this, we have created the **RuEthnoHate** dataset containing texts mentioning numerous Russian ethnic groups, and annotated the corpus in a fine-grained instance-level manner. We have adopted a broad

definition of hate speech based on the negative attitude towards ethnic groups. The annotation involved 3 classes: positive, neutral, and negative attitude towards ethnic groups, with the latter implying ethnicity-targeted hate speech.

We have carried out experiments on hate speech detection with text-level binary attitude detection (**BAD**) and trinary instance-based attitude detection (**IBAD**) approaches with classical machine learning and deep learning models. Text representation included simple unigrams, Word2vec trained on our large RuEthnics dataset (**Word2vec-Ethno**), the Russian National Corpus (**Word2vec-RNC**), and Conversational RuBERT embeddings (**RuBERT-emb**). The classical machine learning models applied were Naive Bayes, Logistic Regression, SVM and ensemble thereof (Voting Classifier, **VC**). Deep learning models were built with **LSTM+GRU** and Conversational RuBERT (**Convers-RuBERT**) architectures. Our best results were obtained on the **IBAD** approach. **Convers-RuBERT** outperformed both classical machine learning and **LSTM+GRU** models. However, the results of **Convers-RuBERT** were significantly improved with hand-crafted linguistic features, including sentiment lexicon, in-domain pre-training and an additional dense layer, reaching **F1-hate** = 0.813, **F1-macro** = 0.824.

To the best of our knowledge, our research is the first study of hate speech in the Russian language targeted at ethnic minorities. Our results lead to the following conclusions:

- Ethnicity-targeted hate speech should be addressed with the instance-based three-class approach including negative, neutral and positive attitudes (**RQ1)**;
- Instance-based ethnicity-targeted hate speech detection significantly benefits from including ethnic information into the input text representation in BERT (**RQ2**), which is a novel approach to this task;
- In instance-based ethnicity-targeted hate speech detection, state-of-the-art deep learning models significantly benefit from a combination of linguistic and sentiment features with BERT pre-training and an additional dense layer, but not from linguistics features separately (**RQ3**).

Moreover, we are making available to the research community the **RuEthnoHate** dataset containing 5,5K social media texts, the first dataset annotated with ethnicity-targeted hate speech in Russian.

As future work, we plan to increase the performance of ethnicity-targeted hate speech detection models in Russian by enriching our dataset and taking into account the following phenomena: (a) meaningful misspellings, including unconventional obscene forms and imitations of foreign accent and language; (b) irony; and (c) other contrasting expressions involving complex narrative logic.

Finally, we plan to perform experiments on ethnicity mention detection and classification, and integrate our models into a fully automatic hate speech detection tool identifying both hate speech and its ethnic group targets in Russian social media texts.

Acknowledgements

This work is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University).

This work is the result of the collaboration with Paolo Rosso in the framework of his virtual online internship at the National Research University Higher School of Economics due to COVID-19. This research was supported in part through computational resources of HPC facilities at NRU HSE.

The authors are grateful for the invaluable help of Gretel Liz De la Peña Sarracén, who shared advice on deep learning models, and Maxim Terpilovsky, who developed the annotation interface.

Competing Interests Statement

The authors have no competing interests to declare.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Kudlur, M. (2016). Tensorflow: A system for large-scale machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)* (pp. 265-283).
2. Akhter, M. P., Jiangbin, Z., Naqvi, I. R., Abdelmajeed, M., & Sadiq, M. T. (2020). Automatic Detection of Offensive Language for Urdu and Roman Urdu. *IEEE Access*, *8*, 91213-91226.
3. Andrusyak, B., Rimel, M., & Kern, R. (2018). Detection of abusive speech for mixed sociolects of russian and ukrainian languages. *Proceedings of Recent Advances in Slavonic Natural Language Processing*. (pp.77–84).

4. Aragón, M. E., Carmona, M. Á. Á., Montes-y-Gómez, M., Escalante, H. J., Pineda, L. V., & Moctezuma, D. (2019, September). Overview of MEX-A3T at IberLEF 2019: Authorship and Aggressiveness Analysis in Mexican Spanish Tweets. In *IberLEF@ SEPLN* (pp. 478-494).
5. Arkhipenko, K., Kozlov, I., Trofimovich, J., Skorniakov, K., Gomzin, A., & Turdakov, D. (2016). Comparison of neural network architectures for sentiment analysis of Russian tweets. In *Computational Linguistics and Intellectual Technologies* (pp. 50-58).
6. Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 759-760).
7. Basile, V., Bosco, C., Fersini, E., Debora, N., Patti, V., Pardo, F. M. R., ... & Sanguinetti, M. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation* (pp. 54-63). Association for Computational Linguistics.
8. Belchikov, A. Russian language toxic comments, https://www.kaggle.com/blackmoon/russian-language-toxic-comments
9. Bodrunova, S. S., Koltsova, O., Koltsov, S., & Nikolenko, S. (2017). Who's bad? Attitudes toward resettlers from the post-Soviet south versus other nations in the Russian blogosphere. *International Journal of Communication*, *11*, 23.
10. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135-146.
11. Bosco, C., Felice, D. O., Poletto, F., Sanguinetti, M., & Maurizio, T. (2018). Overview of the EVALITA 2018 hate speech detection task. In *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian* (Vol. 2263, pp. 1-9). CEUR.
12. Bursztyn, L., Egorov, G., Enikolopov, R., & Petrova, M. (2019). Social media and xenophobia: evidence from Russia (No. w26567). *National Bureau of Economic Research*. https://home.uchicago.edu/bursztyn/SocialMediaXenophobia_December2019.pdf
13. Chen, P., Sun, Zh., Bing, L., Yang, W. (2017). Recurrent Attention Network on Memory for Aspect Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 452–461).
14. Chollet, F. & others (2015). Keras. Available at: https://github.com/fchollet/keras.
15. Cimino, A., De Mattei, L., & Dell'Orletta, F. (2018). Multi-task learning in deep neural networks at evalita 2018. *Proceedings of the Wvaluation Campaign of Natural Language Processing and Speech tools for Italian*, 86-95.
16. Corazza, M., Menini, S., Cabrio, E., Tonelli, S., & Villata, S. (2020). A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT)*, *20*(2), 1-22.
17. Wigand, C., & Voin, M. (2017). Speech by Commissioner Jourová—10 years of the EU Fundamental Rights Agency: A call to action in defence of fundamental rights, democracy and the rule of law. Retrieved from http:// europa.eu/rapid/press-release_SPEECH-17-403_en.htm.
18. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media (ICWSM 2017)*, AAAI Press (2017) 512-515.
19. Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on Facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)* (pp. 86-95).
20. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
21. Dinakar, K., Jones, B., Havasi, C., Lieberman, H., & Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *2*(3), (pp. 1-30).
22. Dunn, J., Adams, B. (2020) Geographically-Balanced Gigaword Corpora for 50 Language Varieties. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)* (pp. 2528-2536)
23. Elmadany, A., Zhang, C., Abdul-Mageed, M., & Hashemi, A. (2020). Leveraging Affective Bidirectional Transformers for Offensive Language Detection. *arXiv preprint arXiv:2006.01266*.
24. Facebook (2013). What does Facebook consider to be hate speech? Retrieved from https://www.facebook.com/help/ 135402139904490
25. Farha, I. A., & Magdy, W. (2020). Multitask Learning for Arabic Offensive Language and Hate-Speech Detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection* (pp. 86-90).
26. Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, *51*(4), 1-30.
27. Fortuna, P., da Silva, J. R., Wanner, L., & Nunes, S. (2019). A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online* (pp. 94-104).
28. Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(4) (pp. 215-230).

29. Golubev, A., & Loukachevitch, N. (2020). Improving Results on Russian Sentiment Datasets. In *Conference on Artificial Intelligence and Natural Language* (pp. 109-121). Springer, Cham.
30. Gu, S., Zhang, L., Hou, Y., & Song, Y. (2018). A position-aware bidirectional attention network for aspect-level sentiment analysis. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 774-784).
31. Haas, J. (2012). Hate speech and stereotypic talk. *The handbook of intergroup communication. London: Routledge*, 128-40.
32. Hernandez, J., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. F. (2013). An empirical study of oversampling and undersampling for instance selection methods on imbalance datasets. In *Iberoamerican Congress on Pattern Recognition* (pp. 262-269). Springer, Berlin, Heidelberg.
33. Hoang, M., Bihorac, O.A., Rouces, J. (2019) Aspect-Based Sentiment Analysis Using BERT. Proceedings of the 22nd Nordic Conference on Computational Linguistics. 187-196.
34. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.
35. Jain, S., & Wallace, B. C. (2019). Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 3543-3556).
36. Kapil, P., Ekbal, A., & Das, D. (2020). Investigating Deep Learning Approaches for Hate Speech Detection in Social Media. *arXiv preprint arXiv:2005.14690*.
37. Karpov, I. A., Kozhevnikov, M. V., Kazorin, V. I., & Nemov, N. R. (2016). Entity Based Sentiment Analysis Using Syntax Patterns and Convolutional Neural Network. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016", 2*, 13.
38. Kiritchenko, S., Zhu, X., Cherry, C., & Mohammad, S. (2014). NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)* (pp. 437-442).
39. Koltsova, O. (2018). Methodological challenges for detecting interethnic hostility on social media. In *International Conference on Internet Science* (pp. 7-18). Springer, Cham.
40. Koltsova, O. Y., Alexeeva, S. V., Nikolenko, S. I., & Koltsov, M. (2017a). Measuring Prejudice and Ethnic Tensions in User-Generated Content. *Annual Review of CyberTherapy and Telemedicine*, *15*, 76–81. http://www.arctt.info/volume-15-summer-2017
41. Koltsova, O., Nikolenko, S., Alexeeva, S., Nagornyy, O., & Koltcov, S. (2017b). Detecting Interethnic Relations with the Data from Social Media. In *Digital Transformation and Global Society* (pp. 16–30). Springer, Cham. https://doi.org/10.1007/978-3-319-69784-0_2
42. Koltsova, O., Alexeeva, S., Pashakhin, S., & Koltsov, S. (2020). PolSentiLex: Sentiment Detection in Socio-Political Discussions on Russian Social Media. In *Conference on Artificial Intelligence and Natural Language* (pp. 1-16). Springer, Cham.
43. Kuratov, Yu., & Arkhipov, M. (2019). Adaptation of deep bidirectional multilingual transformers for Russian language. *Computational Linguistics and Intellectual Technologies*, (18), 333-339.
44. Kutuzov, A., & Kuzmenko, E. (2016). WebVectors: a toolkit for building web interfaces for vector semantic models. In *International Conference on Analysis of Images, Social Networks and Texts* (pp. 155-161). Springer, Cham.
45. Lison, P., & Tiedemann, J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 923-929.
46. Liu, Q., Zhang, H., Zeng, Y., Huang, z., Wu, Z. (2018). Content Attention Model for Aspect Based Sentiment Analysis. In *Proceedings of the 2018 World Wide Web Conference* (pp. 1023–1032).
47. Loukachevitch, N. V., & Rubtsova, Y. V. (2016). SentiRuEval-2016: overcoming time gap and data sparsity in tweet sentiment analysis. In *Computational Linguistics and Intellectual Technologies* (pp. 416-426).
48. Ma, D., Li, S., Zhang, X., Wang, H. (2017). Interactive Attention Networks for Aspect-Level Sentiment Classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (pp. 4068–4074).
49. Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50-60.
50. Mehdad, Y., & Tetreault, J. (2016). Do characters abuse more than words?. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 299-303).
51. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
52. Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., & Patel, A. (2019). Overview of the HASOC track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval*

Evaluation (pp. 14-17).

53. Mollas, I., Chrysopoulou, Z., Karlos, S., & Tsoumakas, G. (2020). ETHOS: an Online Hate Speech Detection Dataset. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*.
54. Moon, J., Cho, W. I., & Lee, J. (2020). BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection. *arXiv preprint arXiv:2005.12503*.
55. Mubarak, H., Darwish, K., Magdy, W., Elsayed, T., & Al-Khalifa, H. (2020). Overview of OSACT4 Arabic Offensive Language Detection Shared Task. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection* (pp. 48-52).
56. Müller, K., & Schwarz, C. (2019). Fanning the flames of hate: Social media and hate crime. *Available at SSRN 3082972*.
57. Niemann, M., Riehle, D. M., Brunk, J., & Becker, J. (2019). What Is Abusive Language?. In *Multidisciplinary International Symposium on Disinformation in Open Online Media* (pp. 59-73). Springer, Cham.
58. Nockleby, J. T. (2000). Hate speech. *Encyclopedia of the American constitution*, *3*(2), 1277-1279.
59. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *The Journal of machine Learning research*, *12*, 2825-2830.
60. Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).
61. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.
62. Pitenis, Z., Zampieri, M., & Ranasinghe, T. (2020). Offensive Language Identification in Greek. In *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 5113-5119).
63. Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2020). Resources and benchmark corpora for hate speech detection: a systematic review. In *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 5113-5119)
64. Ptaszynski, M., Pieciukiewicz, A., & Dybała, P. (2019). Results of the PolEval 2019 Shared Task 6: First Dataset and Open Shared Task for Automatic Cyberbullying Detection in Polish Twitter. *Proceedings of the PolEval 2019 Workshop*, 89.
65. Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.*
66. Rosenthal, S., Atanasova, P., Karadzhov, G., Zampieri, M., & Nakov, P. (2020). A large-scale semi-supervised dataset for offensive language identification. *arXiv preprint arXiv:2004.14454*.
67. Russian Census-2010 statistics (2010). *On the website of the Russian Federal Agency of National Statistics*. Retrieved from https://rosstat.gov.ru/free_doc/new_site/perepis2010/croc/vol4pdf-m.html .
68. Saeidi, M., Bouchard, G., Liakata, M., & Riedel, S. (2016). Sentihood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. *arXiv preprint arXiv:1610.03771*.
69. Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., & Stranisci, M. (2018). An italian twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
70. Shavrina, T., & Shapovalova, O. (2017). To the Methodology of Corpus Construction for Machine Learning: "Taiga" Syntax Tree Corpus and Parser. In *Proceedings of "CORPORA'2017"*, 78.
71. Siegel, A. A. (2019). Online hate speech. *Social Media and Democracy*, 56.
72. Smetanin, S. (2020). Toxic comments detection in Russian. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2020"*.
73. Struß, J. M., Siegel, M., Ruppenhofer, J., Wiegand, M., & Klenner, M. (2019). Overview of GermEval Task 2, 2019 shared task on the identification of offensive language. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, 352-363.
74. Sun, C., Huang, L., & Qiu, X. (2019). Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. *arXiv preprint arXiv:1903.09588*.
75. Tang, D. Qin, B., Feng, X., Liu, T. (2016a). Effective LSTMs for Target-Dependent Sentiment Classification. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (pp. 3298-3307)
76. Tang, D., Qin, B., & Liu, T. (2016b). Aspect Level Sentiment Classification with Deep Memory Network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 214-224).

77. Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., & Daelemans, W. (2016). A dictionary-based approach to racism detection in dutch social media. *arXiv preprint arXiv:1608.08738*.
78. Twitter (2017). The Twitter Rules. Retrieved from https://support.twitter.com/articles/ .
79. Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., ... & Hoste, V. (2015). Detection and fine-grained classification of cyberbullying events. In *International Conference Recent Advances in Natural Language Processing (RANLP)* (pp. 672-680).
80. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & van der Walt, S. J. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, *17*(3), 261-272.
81. Wagner, J., Arora, P., Vaíllo, S. C., Barman, U., Bogdanova, D., Foster, J., & Tounsi, L. (2014). DCU: Aspect-based Polarity Classification for SemEval Task 4. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)* (pp. 223-229).
82. Wang, Y., Huang, M., Zhu, X., Zhao, L. (2016). Attention-based LSTM for Aspect-level Sentiment Classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 606-615).
83. Wang, B., & Lu, W. (2018). Learning Latent Opinions for Aspect-level Sentiment Classification. In *AAAI* (pp. 5537-5544).
84. Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the World Wide Web. In *Proceedings of the second workshop on language in social media* (pp. 19-26).
85. Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop* (pp. 88-93).
86. Wiedemann, G., Yimam, S. M., & Biemann, C. (2020). UHH-LT & LT2 at SemEval-2020 Task 12: Fine-Tuning of Pre-Trained Transformer Networks for Offensive Language Detection. *arXiv preprint arXiv:2004.11493*.
87. Wiegreffe, S., & Pinter, Y. (2019). Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 11-20).
88. Williams, M. L., Burnap, P., Javed, A., Liu, H., & Ozalp, S. (2020). Hate in the machine: anti-Black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, *60*(1), 93-117.
89. Wullach, T., Adler, A., & Minkov, E. (2020). Towards Hate Speech Detection at Large via Deep Generative Modeling. *arXiv preprint arXiv:2005.06370*.
90. Yuan, S., Wu, X., & Xiang, Y. (2016). A Two Phase Deep Learning Model for Identifying Discrimination from Tweets. In *Proceedings of the International Conference on Extending Database Technology EDBT* (pp. 696-697).
91. Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., ... & Çöltekin, Ç. (2020). SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). *arXiv preprint arXiv:2006.07235*.
92. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 75-86).
93. Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting hate speech on twitter using a convolution-GRU based deep neural network. In *European semantic web conference* (pp. 745-760). Springer, Cham.
94. Zueva, N.; Kabirova, M., & Kaliadin, P. (2020) Reducing Unintended Identity Bias in Russian Hate Speech Detection. https://arxiv.org/abs/2010.11666

Appendix 1. Overview of approaches towards online hate speech detection

Paper	Dataset	Hate group(s)/target(s)	Classes	Method	F1 (hate)
(Dinakar et al., 2012)	Youtube (4.5K) and Formspring	sexuality; race & culture; intelligence	2[15]	SVM + features (lexicon, tf-idf, POS, abusive words)	0.77; 0.638; 0.58

[15] In this paper separate binary models are trained and evaluated for 3 types of hate speech based on sexuality, race and intelligence

(Warner & Hirschberg, 2012)	Yahoo! and the American Jewish Congress (1K paragraphs)	jews, black, asian, women, muslims, immigrant, other	2	SVM + features (n-grams, Brown clusters, POS templates)	0.63
(Gitari et al., 2015)	180 + 320 labeled paragraphs from blogs	ethnicity + religion + nationality	3[16]	Rule learning + dependency patterns + lexicon features	0.708[17]
(Van Hee et al., 2015)	ask.fm (85K) - Dutch	women; any people (hate types: threat/blackmail, insult, curse/exclusion, defamation, sexual talk, defense, encouragement to the harasser)	2	SVM + features (n-grams, char n-gams, lexicon)	0.554
(Tulkens et al., 2016)	Facebook (6K) - Dutch	ethnicity + nationality + religion + culture	2	SVM + sentiment lexicon features expanded by word2vec	0.46
(Waseem and Hovy, 2016)	Twitter (16K)	race, gender	4[18]	Logistic Regression + char-ngrams + gender information	0.739
(Mehdad and Tetreault, 2016)	Yahoo Finance (951K)	any target (abusive language detection)	2	NBSVM + RNNLM (char-level)	0.79
(Badjatiya et al., 2017)	Twitter (16K)	race, gender	2	LSTM + randomly initialized GloVe embeddings + GBDT	0.930
(Davidson et al., 2017)	Twitter (25K)	any target	3[19]	Logistic Regression + fine-grained features	0.51
(Del Vigna et al., 2017)	Facebook (17K) - Italian	religion, physical and/or mental handicap, socio-economical status, politics, race, gender, other	2	BiLSTM + 2 types of word embeddings + features	0.728
(Fortuna and Nunes, 2018)	Twitter (5K) - Portuguese	gender, body, origin, sexuality, ethnicity, ideology, religion, health, lifestyle	2	MLP + hateful n-gram features	0.76
(Zhang et al., 2018)	Twitter (2.4K)	refugees, muslims	2	CNN + GRU + Word2Vec Skip-gram embeddings	0.92
(Wiedemann et al, 2020)	Twitter (14.1K)	any target	2	Ensemble of ALBERT models	0.891
(Wullach et al., 2020)	Existing datasets augmented by GPT-2 (200K)	any target	2	CNN + GRU	0.678[20]

[16] The three classes are strong hate, weak hate and no hate.

[17] Results are only reported for the strong hate class in [10]. The authors also evaluated their model on the two different corpora separately, and we only selected the best result out of the two reported ones.

[18] The four classes in question are racism, sexism, both, none.

[19] The three classes are hateful, offensive and none.

[20] In [25] the authors performed both intra-dataset and cross-dataset experiments with 5 different datasets. We include their average intra-dataset F1 score in Table 1.

(Mollas et al., 2020)	Youtube + Reddit (1K)	gender, race, national origin, disability, religion, sexual orientation	2	Fine-tuned BERT	0.744
(Moon et al., 2020)	News comments (9.4K) - Korean	gender, other	2	KoBERT + BiLSTM	0.681

Appendix 2. Error annotation labels

Label	Description
not_clear	Is it not clear what the text is about
irony	Author is ironic towards target ethnic group
implicit	Author's attitude is implicit; ethnic group may not be even mentioned in the text; we can guess what the attitude is based on our knowledge of the world and political agenda, etc.
indirect	Attitude towards ethnic groups is expressed by showing attitude of other people/nations towards this ethnic group
context_pos	Words conveying positive sentiment towards ethnic group in the context of ethnonym(s)
context_neg	Words conveying negative sentiment towards ethnic group in the context of ethnonym(s)
noncontext_pos	Words conveying positive sentiment towards ethnic group out of the context of ethnonym(s) (far from ethnonym in the text)
noncontext_neg	Words conveying negative sentiment towards ethnic group out of the context of ethnonym(s) (far from ethnonym in the text)
general_pos	General positive sentiment expressed in text (NOT towards ethnicities) - including sentiment towards author's interlocutor
general_neg	General negative sentiment expressed in text (NOT towards ethnicities) - including sentiment towards author's interlocutor
other_pos	Positive sentiment expressed towards OTHER ethnic group mentioned in the same text
other_neg	Negative sentiment expressed towards OTHER ethnic group mentioned in the same text
negation_pos	Positive attitude towards ethnic group is expressed using phrases preceded with negation (e.g., "they could never offend anyone")
negation_neut	Neutral attitude towards ethnic group is expressed using phrases preceded with negation (e.g., "it's not that they did something bad...")
negation_neg	Negative attitude towards ethnic group is expressed using phrases preceded with negation (e.g., "they don’t' want to work")
actions	Attitude towards ethnic group is expressed by the description of its actions including those taken during some historical events (e.g., a particular ethnic group showed its prowess and strength in the military confrontation with other ethnic groups/nations; or an ethnic group is well-known for its hospitality)
question_pos	Positive attitude towards ethnic group is expressed by a question

question_neut	Neutral attitude towards ethnic group is expressed by a question
question_neg	Negative attitude towards ethnic group is expressed by questioning positive qualities of it (e.g., "have you seen them do anything good?")
call	Text contains call for aggression against ethnic group
ethnophaulism	Ethnic group is described using ethnophaulism(s)
contrast	Attitude towards ethnic group is shown by contrasting a particular ethnic group to other ethnic group(s)
noun phrase	Ethnic group or a person of this group is referred to by a noun phrase "adjective + noun" (e.g., "russian people, turkish nation, american businessman etc.")
discussion	Ethnic group is mentioned in a discussion where both positive and negative lexicon is used and contrasting opinions are described (WITHOUT expressing author's attitude towards ethnic group)
quote	Text is a quotation from a historical novel / poem / song / film / etc.
anaphora	Attitude towards ethnic group is expressed using phrases with anaphoric reference towards this ethnic group
caps	CAPS LOCK expressing strong sentiment
strong	Strong sentiment
exclam	Sentiment is expressed using exclamation marks
prejudice	Negative attitude towards ethnic group is due to prejudice against this ethnic group
obscene	Negative attitude towards ethnic group is expressed using obscene words
no_ethnonym	Ethnic group is not mentioned in the text (it was annotated by mistake)

Appendix 3. Experts agreement (Krippendorff's alpha)

Assessor	Expert 1	Expert 2	Expert 3	Expert 4
Expert 1	1.00	0.56	0.39	0.51
Expert 2	0.56	1.00	0.50	0.81
Expert 3	0.39	0.50	1.00	0.43
Expert 4	0.51	0.81	0.43	1.00

Appendix 4. Convers-RuBERT+Ling+Dense+Dense2 training parameters

Parameters	Value
Epochs	20
Batch size	24
Optimizer	Adam
Learning rate	1e-5

Activations	Dense: relu
	Dense2: sigmoid
Layer sizes	Dense: 100
	Dense 2: 100
Loss function	categorical crossentropy
Bert layer pooling strategy	mean pooling
Bert layer: fine-tuned layers	last four layers

Ekaterina Pronoza: Methodology, Software, Formal Analysis, Investigation, Data Curation, Visualization, Writing - Original Draft

Polina Panicheva: Methodology, Software, Validation, Formal Analysis, Investigation, Writing - Original Draft, Writing - Review & Editing

Olessia Koltsova: Conceptualization, Methodology, Data Curation, Writing - Review & Editing, Supervision, Project administration, Funding acquisition

Paolo Rosso: Conceptualization, Methodology, Supervision