Abusive Language Recognition in Russian

Kamil Saitov
Innopolis University
Russian Federation
saitov66@gmail.com

Leon Derczynski
IT University of Copenhagen
Denmark
ld@itu.dk

Abstract

Abusive phenomena are commonplace in language on the web. The scope of recognizing abusive language is broad, covering many behaviours and forms of expression. This work addresses automatic detection of abusive language in Russian. The lexical, grammatical and morphological diversity of Russian language present potential difficulties for this task, which is addressed using a variety of machine learning approaches. We present a dataset and baselines for this task.

1 Introduction

Unfortunately, hate speech and abusive language are prevalent on the internet (Waseem and Hovy, 2016), often creating an aggressive environment for users. This can include cyber-bullying or threats towards individuals and groups. Reducing this content is difficult: it is harmful for humans to moderate.[1] Thus, there is a critical need for abusive language recognition systems, which would help social networks and forums filter abusive language. Moreover, with platforms taking increased control over which content to surface, automatic abuse recognition is more important than ever.

One problem arises when the subjectivity of the matter is considered. Abusive language is hard for humans to recognize universally (Waseem, 2016). This indicates that the collection and labeling of data should be thorough and objective, which could be reached through e.g. large-scale crowd-sourced data annotation (Sabou et al., 2014).

NLP research in the area is nascent, with existing solutions oriented mostly towards English language (Vidgen and Derczynski, 2020), which, despite sometimes being mistakenly considered as "universal" (Bender, 2019), is very different grammatically and lexically from many languages, especially those using non-Latin characters (e.g. Russian, Japanese etc). This paper addresses abusive language detection in Russian. One issue with recognition of abusive language in Russian is the limited number of sources of labeled data relative to English (Andrusyak et al., 2018; Zueva et al., 2020; Smetanin, 2020; Potapova and Gordeev, 2016). Thus, the collection and labeling of data presents an additional challenge, and we present both dataset and models.

2 Abusive Language Definition

In this case, we use the OLID annotation definition of abusive language (Zampieri et al., 2019). This covers profanity, and targeted and untargeted insults and threats, against both groups and individuals. Specifically, in accordance this scheme, we consider the use of racial and other group-targeted slurs abusive.

3 Dataset

3.1 Data collection

We searched for publicly available datasets containing considerable amounts of abusive language.

Russian Troll Tweets is a repository consisting of 3 million tweets.[2] This was filtered to only Cyrillic texts. This data is not labeled, thus a subset of the data was labeled manually for use in this research. During labeling, the data turned out to contain significantly less abusive language than expected. An additional resource, the RuTweetCorp (Rubtsova, 2013), was also annotated for abusive language.

In search for sources rich in abusive language, the "South Park" TV show was found. The Russian subtitles for it embodied a high density of profanity, hate-speech, racism, sexism, various examples of

[1] https://www.theverge.com/2020/5/12/21255870/facebook-content-moderator-settlement-scola-ptsd-mental-health

[2] https://github.com/fivethirtyeight/russian-troll-tweets

Proceedings of the 8th BSNLP Workshop on Balto-Slavic Natural Language Processing, pages 20–25

ethnicity and nationality abuse. The subtitles from more than four seasons of the series yielded many instances of abusive language. This data, Russian South Park (RSP), was annotated manually. Inter-annotator agreement (IAA; computed with Cohen's Kappa) over the whole dataset is 0.68 among three L1[3] Russian annotators.

To complement this, the Kaggle "Russian Language Toxic Comments" dataset (RTC) was also annotated. The dataset contains more than 14 000 labeled samples of hate speech. In Section 4, the performance of models trained on RSP data will be compared to that including RTC. More than 1500 samples are in the RSP dataset, and more than 15 000 samples are in total, adding the RTC data.

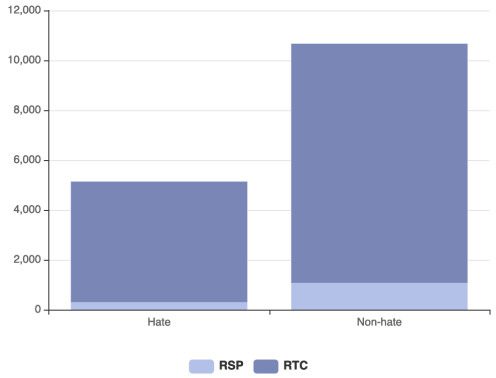


Figure 1: Dataset parts size and balance

		Instances	Tokens
Total	Abusive	5 904	133 180
	Non-abusive	9 893	192 307
RSP	Abusive	307	4 356
	Non-abusive	1 078	18 189
RTC	Abusive	5 597	128 824
	Non-abusive	8 815	174 118

Table 1: Word & token distribution across RSP

As well as in many *in situ* abusive language research, an abusive language lexicon was also constructed. The text data that was collected previously contained a fair amount of such vocabulary, however, the dictionary should not be limited by the dataset. HateBase (Tuckwood, 2017) contains only 17 abusive Russian words. VK, the largest social network in Russia and CIS, has an abusive speech filter dictionary published unofficially, containing a large lexicon of abusive words.[4] Another source is *russki-mat*,[5] an open dictionary of Russian curse words with proper explanations and examples of usage. Overall, the multiple-source lexicon built contains more than 700 unique terms. As can be seen from Table 2, abuse-bearing sentences contain four times more uppercased words and 25 times more abusive words than non-abusive sentences.

Words	mean	deviation
Uppercase (abusive)	0.02	0.007
Profane (absuive)	0.05	0.003
Uppercase (non-abusive)	0.005	0.00019
Profane (non-abusive)	0.002	0.00005

(a) At sentence level

Words	mean	deviation
Uppercase (abusive)	0.02	0.0003
Profane (absuive)	0.03	0.0004
Uppercase (non-abusive)	0.006	1.84E-05
Profane (non-abusive)	0.003	3.80E-06

(b) Across the whole dataset

Table 2: Uppercase and profane word distribution across the dataset

4 Experiments

4.1 Data Preprocessing

The stages of pre-processing are the following:

1. Balance the dataset. The initial dataset no-hate/hate distribution is $1078/307$ for the RSP dataset and $8815/5597$ for the RSP+RTC dataset. The no-hate portion of the dataset is under-sampled so that this proportion is consistent.

2. Strip URLs. Remove the links from texts.

3. Adjust platform-specific text. All Twitter mentions, hashtags and retweet are shown by a set of distinct symbols (# for hasthtag, @ for retweet). These tags might hold information on whether the tweet is targeted at a particular person or not.

4. Orthographic normalisation. Replace Russian ë and Ë to the corresponding e and E. These letters are mostly interchangeable in Russan language, thus it is the standard preprocessing routine when working with Russian text data.

5. Tokenization. Splitting the sentences into separate words and punctuation. The tokenization is done with NLTK library's word_tokenize() method.

6. Lemmatize terms. Lemmatization is reducing the word into its normal form. In case of Russian language, most researchers prefer stemming over lemmatization, however, if stemming is used,

[3] I.e. as first language

[4] Common Knowledge Russian Tweets, http://study.mokoron.com/

[5] http://www.russki-mat.net/home.php

the search for offensive words in sentences would become intractable. The lemmatization is done with pymorphy2 (Korobov, 2015) - a morphological analyzer library specifically for Russian language.

7. Remove stop words from the text. Such words are common interjections, conjugations, prepositions, that do not need to be seen as features in the future modelling of the data.

8. TF-IDF vectorization. Turn the words into frequency vectors for each sample.

9. Train-test split Randomly split the ready data into train and test sets with 80/20 proportion.

4.2 Feature Extraction

Additional features beyond the text itself are included. Since abusive or hateful comments are anticipated to be also negative in sentiment, sentiment analysis is included. The sentiment was automatically predicted for the RTC dataset, for which a FastText (Bojanowski et al., 2017) embedding induced over RuSentiment (Rogers et al., 2018) was used, achieving F1 of 0.71, high for sentiment classifiers for Russian.

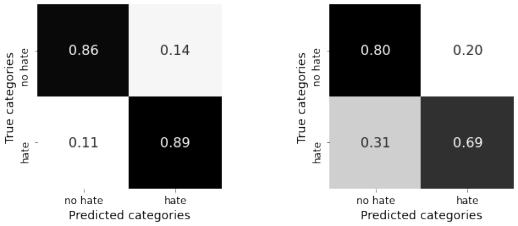
Upper-casing full words is a popular tone-indicating technique (Derczynski et al., 2015). Since one cannot "shout" in the internet, the intent of a higher-tone is expressed with upper-casing. Therefore, the number of fully-uppercased words is counted for each sample.

We also count the number of offensive words (from our lexicon) contained in a sentence. This feature is expected to be important, since abusive language is often combined with profanity, though this kind of sampling is not without bias (Vidgen and Derczynski, 2020).

4.3 Baseline Results [no RTC data]

The baseline model is a binary Linear Support Vector Classifier with default L2 loss and squared-hinge loss. The model was chosen to be an SVC because similar work for other languages suggest that it can be effective for this type of task (Frisiani et al., 2019).

The overall F1-score is up to 0.75, depending on the seed and parameters. The F1-score on the RSP+RTC Comments dataset is higher, up to 0.87, again, depending on the seed and parameters (Figure 2). Analysing the incorrectly classified samples, it turns out that the main difficulty the model has is



(a) with RTC data (b) no RTC data

Figure 2: Confusion matrixes of the baseline model

longer texts as well as texts containing swear words that cannot be converted to initial form due to distortion through slang/word formation. An example of this is the following:

> В чем проблема? Деградируй до неандертальца и х*ярь (heavily distorted slang) п*дарасов (misspelling)
>
> *What is the problem? Degrade to a Neanderthal level and kick those f*ggots*

The following example is a stereotypical hate speech sentence - it is all upper-cased, it uses abusive words and contains numerous insults. The baseline model recognizes it well:

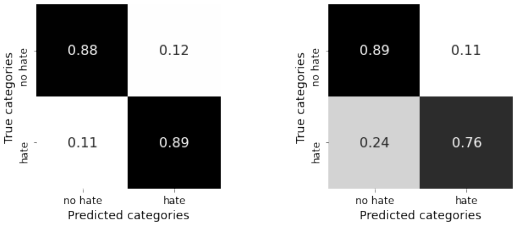
> КРЫМОТРЕД НАРУШАЕТ ПРАВИЛА РАЗДЕЛА Т.К В НЕМ НЕТ ОБСУЖДЕНИЯ ПОЛИТИКИ. СВОБОДНОЕ ОБЩЕНИЕ ЭТО В b. ЭТО ТОЖЕ САМОЕ ЕСЛИ Я НА ДОСКЕ О ПОЛИТИКЕ СОЗДАМ ТРЕД О ШЛ*ХАХ. ТАК ЧТО У*БЫВАЙТЕ В Б ИЛИ НВР СО СВОИМ ЧАТИКОМ ПРЕСТАРЕЛЫХ Г*МОСЕКОВ!
>
> *CRIMEA THREAD VIOLATES THE RULES OF THE FORUM BECAUSE THE RULES DOES NOT ALLOW POLITICS DISCUSSION. THIS IS THE SAME IF I START DISCUSSING SL*TS ON A POLITICS FORUM. SO GET THE F*CK OUT OF HERE AND GO TO ¡another forum¿ AND TAKE YOUR WHOLE OLD F*GGOTS PARTY WITH YOU!*

4.4 Skip stopword exclusion

Although removing stop words from tokenized text is common practice, leaving them in might yield different results. This is the case here. The results are better on both datasets. F1-score over the RTC+RSP dataset is 0.88 (Figure 3).

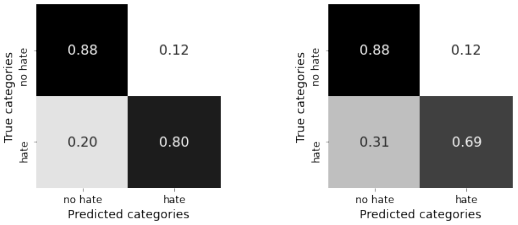
4.5 Without balancing the dataset

In this experiment, the datasets are not balanced, thus the proportion of hate/no-hate is 1/2 in the combined RTC+RSP dataset and 1/10 in the RSP. As can be seen in Figure 4, true positives decrease by a small amount and the false negatives have risen up by a large margin, causing a decrease in overall model performance.



(a) with RTC data (b) no RTC data

Figure 3: Improved recall and precision on both datasets without stopword filtering



(a) with RTC data (b) no RTC data

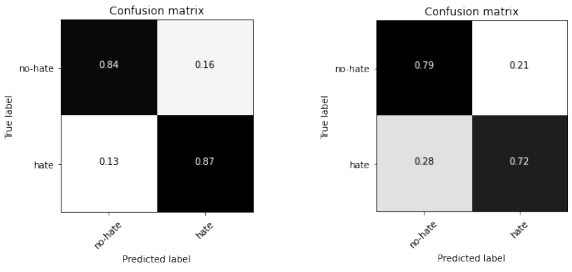
Figure 4: Performance without giving balancing instance weights

4.6 Deep Learning

Neural network-based approaches often show promising results on various NLP tasks. In fact, some of the best methods for hate-speech detection in English are BERT, CNN, GRU/LSTM-based techniques (Zampieri et al., 2020). We investigated these methods over RSP.

Model	F1	Recall	Precision
RuBERT	0.85	0.86	0.84
mBERT	0.76	0.73	0.79

(a) BERT variant performance



(b) RuBERT (c) mBERT

Figure 5: Performance of BERT variations over the combined dataset

4.6.1 RuBERT

RuBERT (Burtsev et al., 2018) is the original Bidirectional Encoder Representations from Transformers (Devlin et al., 2019) model but trained on Russian Wikipedia pages. The fine-tuning needed to be made includes training the last, classifier layer of the network. The results are promising, reaching F1-score of 0.85 on the whole training dataset (confusion matrix in Figure 5).

The model is able to correctly recognize the following sample as hate-speech:

> Посмотрел Утомленных солнцем 2. И оказалось, что это хороший фильм, такая высокобюджетная артхаусятина, к которой могут быть претензии только потому, что сп*здили-распилили и вообще ТАК НЕ БЫВАЕТ. Ну н*хуй этих критиков. Обзоры длиннее фильмов, петросянство хуже рашкомедий, еб*нутая ненависть и до*бки по мелочам.
>
> *Watched Burnt by the Sun 2. Turns out it's a pretty good movie, a high-budget arthouse-ish film, the only downside possible is that most of the budget has been corruptly-stolen and THE PLOT IS NOT REALISTIC. F*ck those critics. The review texts are longer than the movie itself, jokes are worse than <humor in Russian-produced comedies>, f*cked up hate and f*cking nagging about small errors.*

4.6.2 mBERT

mBERT is multilingual BERT (Devlin et al., 2019), again trained on Wikipedia pages of over a hundred languages, mainly of non-Latin alphabets. Russian is Cyrillic, thus the model has the potential in Russian hate-speech recognition domain. The fine-tuning is the same as for RuBERT.

The results (Figure 5) showed worse performance than RuBERT, up to 0.76 F1-score. The reason for the lower performance is probably in the concept of generalisation of BERT to multiple languages, as opposed to RuBERT, which is trained exclusively on Russian language.

The following is an example of a sample which has been incorrectly classified as no-hate with both BERT-based models, as well as the baseline model:

> Вонючий совковый скот прибежал и ноет. А вот и сторонник демократии и свободы слова закукарекал.
>
> *The stinking soviet cattle came running and whining. And here is the supporter of democracy and freedom of speech starting to croak.*

The sentence does not contain any especially abusive vocabulary, but rather the words "stinking", "cattle", "croak" in this context (in relation to people) are abusive.

4.7 Analysis

For the largest dataset of Russian abusive language samples (RSP+RTC) and the LinearSVC model, the best-case is 0.88. This is a good result for

such a simple model compared to typical results in other languages (Zampieri et al., 2020). Our suggestion is that the reason for such a good score is the correct data preprocessing and, even more importantly, feature selection.

Processing	RTC+RSP	RSP only
Base	0.86	0.75
No stopword removal	**0.88**	**0.83**
No dataset balancing	0.85	0.80

Table 3: Ablations over data processing steps, with SVM classifier (F-scores)

RuBERT still struggles mainly with recognizing longer texts and texts with misspellings. Another barrier for this model in particular is when a text contains many named entities, because word representations might not contain entity surface forms (Augenstein et al., 2017) or individual entities may not be representative of the typical context of a given abusive language phenomena.

An example of the above-mentioned criteria is the following long sentence with many named entities (NEs) and misspellings:

> Сторонники бандеровцев (NE) (леваков (NE), выступавших за бесклассовое (misspelling) общество и борьбу с капитализмом) и карлика-душителя котов Степана Бандеры (NE), который, как известно, боролся с расизмом, поддерживал Идель-Урал (NE) и называл побратимами исламских борцов за свободу из Азербайджана (NE), не пользуются симпатиями у правых европейцев.

The mistakes made by mBERT are roughly a superset of those made by RuBERT. This suggests that information mBERT can gain from other languages is not particularly helpful for this task.

5 Conclusion

This paper presented data, models and experiments for abusive language detection in Russian. By choosing the right preprocessing techniques and language-specific feature selection it is possible to achieve state-of-the-art performance on par with best-performing English language models, even using a simple SVM model. This indicates that, given sufficient diversity of data, abusive language detection solutions can be rapidly developed for new languages.

The code and data for this research are publicly available at: https://github.com/Sariellee/Russan-Hate-speech-Recognition

References

Bohdan Andrusyak, Mykhailo Rimel, and Roman Kern. 2018. Detection of abusive speech for mixed sociolects of Russian and Ukrainian Languages. In *Proceedings of RASLAN*, pages 77–84.

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83.

Emily Bender. 2019. The #BenderRule: On Naming the Languages We Study and Why It Matters. *The Gradient*.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yurii Kuratov, Denis Kuznetsov, et al. 2018. DeepPavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127.

Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke Van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. 2015. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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)*, pages 4171–4186.

Nicolò Frisiani, Alexis Laignelet, and Batuhan Güler. 2019. Combination of multiple deep learning architectures for offensive language detection in tweets. *arXiv preprint arXiv:1903.08734*.

Mikhail Korobov. 2015. Morphological analyzer and generator for russian and ukrainian languages. In Mikhail Yu. Khachay, Natalia Konstantinova, Alexander Panchenko, Dmitry I. Ignatov, and Valeri G. Labunets, editors, *Analysis of Images, Social Networks and Texts*, volume 542 of *Communications in Computer and Information Science*, pages 320–332. Springer International Publishing.

Rodmonga Potapova and Denis Gordeev. 2016. Detecting state of aggression in sentences using CNN. In *International Conference on Speech and Computer*, pages 240–245.

Anna Rogers, Alexey Romanov, Anna Rumshisky, Svitlana Volkova, Mikhail Gronas, and Alex Gribov. 2018. RuSentiment: An enriched sentiment analysis dataset for social media in Russian. In *Proceedings of the 27th international conference on computational linguistics*, pages 755–763.

YV Rubtsova. 2013. A method for development and analysis of short text corpus for the review classification task. In *Proceedings of Conferences Digital Libraries: Advanced Methods and Technologies, Digital Collections, RCDL*, pages 269–275.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of LREC*, pages 859–866.

Sergey Smetanin. 2020. Toxic Comments Detection in Russian. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2020"*.

Christopher Tuckwood. 2017. Hatebase: Online database of hate speech. *The Sentinal Project. Available at: https://www. hatebase. org*.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS one*, 15(12):e0243300.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? Annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020).

Nadezhda Zueva, Madina Kabirova, and Pavel Kalaidin. 2020. Reducing unintended identity bias in Russian hate speech detection. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 65–69.

A Data Statement

This appendix describes metadata for RSP, following Bender and Friedman (2018).

A. Curation rationale The texts were taken from the South Park TV series in order to gather a corpus relatively rich in various forms of abusive language.

B. Language variety Scripted Russian translated at high standard from US English. BCP47 `ru-RU`

C. Speaker demographic The text is transcribed from words of Russian actors, mostly male, portraying characters who are both adults and children. The child characters (age eight) make up most of the speech content. The scripts were originally written by two US males from Colorado, over a period where they were aged 20-something to 40-something.

D. Annotator demographic Native Russian speakers, male, twenties, university students.

E. Speech situation This is scripted TV speech; it's not know how much latitude the voice actors were afforded over wording.

F. Text characteristics The content is deliberately somewhat foul-mouthed and very informal; political satire and social commentary are common themes.