Hate-speech recognition in Russian language

Kamil Saitov
Innopolis University
k.saitov@innopolis.ru

Abstract

Hate speech and offensive expressions are a commonplace on the Internet, creating a deep academic and industrial interest in recognizing such language. However, the task of recognizing hate speech varies greatly. In this paper, I define the problem as detection of insults towards a particular person(s), place, occurrence, or other notion. The lexical, grammatical and morphological diversity of Russian language is what characterizes the great challenge of this task. In this paper, I describe collection of the dataset, mixing it with the existing datasets, processing and experimenting with data, before applying the Machine Learning algorithms on it.

1 Introduction

Unfortunately, hate speech and abusive language are extremely common on the Internet, creating an aggressive environment for the users. The apex of hateful tone on the Web is cyber-bullying or even severe threats towards an individual. Thus, there is a critical need for hate speech recognition systems, which would help social networks and forums filter the abusive language and possible offenses out. Moreover, with the recent ubiquitous introduction of recommendation systems that are supposed to deliver the most relevant content to the user, the importance of automatic offence recognition is escalated more than ever.

One problem arises when the subjectivity of the matter is considered. Hate speech is hard to recognize for humans. It is known that younger annotators are more likely to label a given sample as offensive. That proves that the collection and labeling the data should be thorough and objective, which could be reached through large-scale crowd-sourced data annotation.

One issue with recognition of hate speech in Russian is that there is a very limited number of sources for the hate speech data in Russian language, and there are no sources of labeled data. Thus, the collection and labeling of data is another challenge to be dealt with.

The NLP research in the area is still rather insufficient, with existing solutions oriented mostly towards English language, which, despite being considered as ”universal”, still is very different grammatically and lexically from most languages.

2 Data

2.1 Data collection

The initial research was done to find the publicly available datasets containing a considerable amounts of dense offensive speech data.

Russian Troll Tweets is a repository consisting of 3 million tweets. While being a large resource with potentially high portion of offense, it needed to be filtered massively, leaving only Cyrillic texts and leaving out the unimportant information. The data is not labeled, thus a small fraction of the repository's data was labeled manually and used in this research. During labeling, the data turned out to contain significantly less hate speech than expected, therefore revealing the demand for another data source. Another resource, the RuTweetCorp, was attempted to extract hate speech from, with mixed succcess. The extraction of a small amount of samples of hate speech implied labeling an incomparably large amount of text. Manual extraction of abusive language from the news, forums and social network comments was intangible. In search for a densely-packed offensive data, the ”South Park” TV show was found. The Russian subtitles for it embodied an unrivaled amount of offense, hate-speech, racism, sexism, offenses, various examples of ethnicity and nationality abuse. The subtitles from more than 4 seasons of the series yielded 280 samples of highly valuable data. In the last part of my research, the Kaggle ”Russian Language Toxic Comments” was found. The dataset contains 10 000 labeled samples of hate speech. In this paper, the performance on the manually collected data will be compared to the Toxic Comments plus manually collected data performance.

In total, more than 1500 samples are in the manually labeled dataset, and more than 15 000 samples are in total, adding the Toxic Comments data.

Another data resource needed is the offensive (abusive) words and expressions vocabulary. The text data that was collected previously contained a fair amount of such vocabulary, however, the dictionary should not be limited by the dataset. The HateBase contains only 17 abusive Russian words, 5 of them are actually relevant. VK, the largest social network in Russia and CIS, has its abusive speech filter dictionary published unofficially, containing a large lexicon of abusive words. Another source is russkimat, which is an open dictionary of Russian curse words with proper explanations and examples of usage. Overall, the offensive words vocabulary, collected from various sources, contains more than 700 unique terms.

2.2 Data Preprocessing

The collected corpus partially consists of tweets, but mostly is a corpus of forum comments and subtitles. The stages of pre-processing are the following:

1. Balance the dataset. The initial dataset no-hate/hate distribution is $1077/386$ for the manual dataset and $10663/5211$ for the manual+toxic dataset. I stratify the whole dataset so that this proportion is more equal.

2. Remove all the URL links from the texts. They do not carry any useful information, thus are just disposed of.

3. Replace all Twitter mentions, hashtags and retweet tags by a set of distinct symbols (”#” for hashtag, ”@” for retweet). These tags might hold valuable information on whether the tweet is targeted at a particular person or not.

4. 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. Tokenize the text. Tokenization is the task of splitting the sentences into separate words and punctuation.

6. Lemmatize the 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, the search for offensive words in sentences would become an unbearable task.

7. Remove the 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.

2.3 Feature Extraction

For feature extraction, the focus is on the textual features, which are contained in the text. Number of chars can indicate the tone in the sentence, total number of chars and words can help distinguish the relationship between the length of the text and the potential for being offensive.

Sentiment analysis is an important feature, since abusive or hateful comments are usually expected to be also negative in sentiment. The sentiment was automatically predicted for the Toxic Comments dataset, for which the RuSentiment-trained FastText model was used (reportedly up to 0.71 F1-score, which is in the top-tier sentiment classifiers for Russian language)

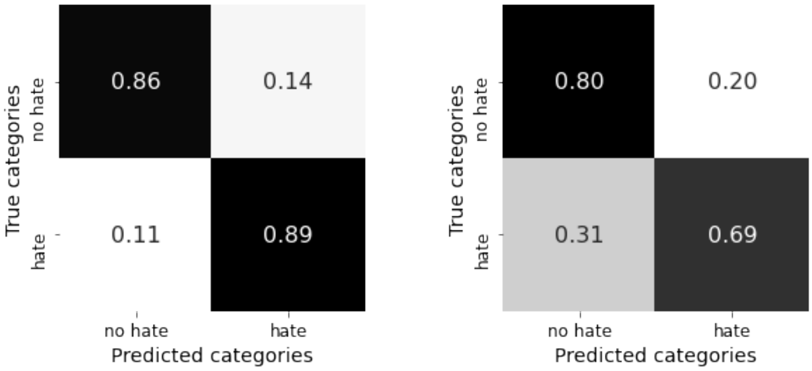
CAPS-ing, or upper-casing full words is a popular tone-indicating technique. Since one cannot ”shout” in the internet, the intent of a higher-tone is expressed with upper-casing. Therefore, the number of fully CAPS-ed words is counted for each sample.

Number of offensive words, containing in the sentence. This feature is expected to be among the most important ones, since hate speech is often used with a proportional amount of swear words.

3 Model baseline

The baseline model is a binary-classifying Linear Support Vector Classifier with default l2 loss and squared-hinge loss. The model was chosen to be an SVC because similar researches for other languages suggest that it is optimal for this type of task.

3.1 Baseline Results [no Toxic Comments data]



(a) with Toxic Comments data

(b) no Toxic Comments data

Figure 1: Confusion matrixes of the baseline model

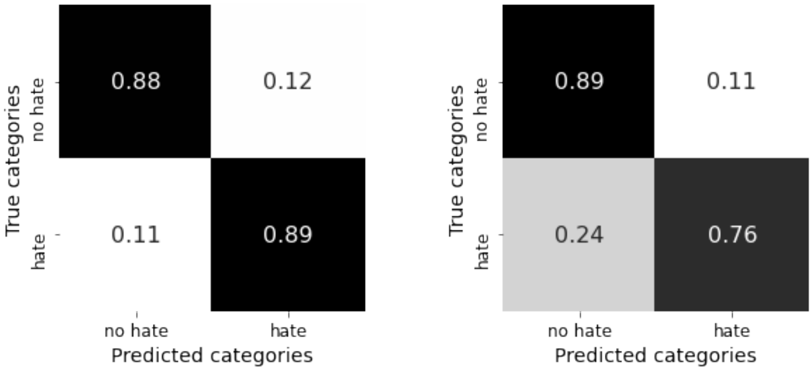
The overall accuracy is up to 81%, depending on the seed and parameters.

The accuracy on the Manual+Toxic Comments dataset is higher, up to 84%, again, depending on the seed and parameters.

4 Experiments

4.1 Without stop words

Although removing the stop words from the tokenized text is a common practice, in some cases removing this preprocessing step might yield unexpected results. This is definitely an example of such behaviour. The results are better on both datasets.



(a) with Toxic Comments data

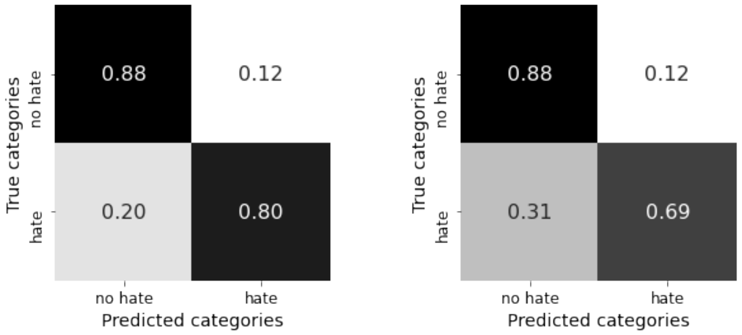
(b) no Toxic Comments data

Figure 2: Improved recall and precision on both datasets if no stopword filtering used

The total accuracy for the Toxic+Manual dataset is 89.9

4.2 Without balancing the dataset

This experiment is bound to fail, but it is a good observation nonetheless. In this experiment, the datasets are not balanced, thus the proportion of hate/no-hate is 1/2 in the Toxic+Manual dataset and 1/10 in the Manual dataset.



(a) with Toxic Comments data

(b) no Toxic Comments data

Figure 3: Decreased performance due to imbalance in the dataset

As can be seen, the true positives have gone down by a small amount and the false negatives have risen up by a large margin, causing the decrease in overall model performance.

5 Results and Analysis

For the largest dataset of Russian hate speech samples (Manual+Toxic Comments) and the LinearSVC model, the best-case accuracy is 89.9%. This is an impressive result for such a simple model. My suggestion is that the reason for such a good score is the correct data preprocessing and, even more importantly, feature selection.

5.1 Failures

1. Initially, the manual dataset labeling was not binary, but consisted of 4 classes(from 0 to 3), which represented the scale of hate-speech and offensiveness of the sample. However, when run on the same pipeline, the model's performance was not higher than 55The reason for that is that it is very hard to distinguish such slight variations in offensiveness. The binary hate/non-hate classification is much more approachable way, at least for the baseline model, which is LinearSVC.

2. There was an attempt on data augmentation, instead of stratifying to balance the dataset. However, the model's performance was much worse. One explanation of this phenomenon is that the augmented data did not follow the patterns that are present in the existing data, thus confusing the model instead of providing more useful data.

5.2 Future Development potential

The baseline model described in this paper is fairly simple. It was unexpected to see that it would yield such great results. It is very curious to see how the more complex models will behave. Suggested in another relevant paper (Guler et al., 2019) is the Convolutional Neural Network model and the Bi-LSTM Neural Network model, which have a great potential to produce results that are better than described in this paper.

6 Conclusion

I found that, even with a relatively small self-made Russian language dataset and a simple SVM model, by choosing the right preprocessing techniques and smart language-specific feature selection it is possible to achieve state-of-the-art performance that is on par with the best-performing English language models. This proves that the shortage of research and new solutions are not grounded and needs to be, and can be, eliminated.

The code for this paper is publicly available at:
https://github.com/Sariellee/Russan-Hate-speech-Recognition

References

Russian Troll Tweets repository. https://github.com/fivethirtyeight/russian-troll-tweets

Рубцова Ю., 2012 Автоматическое построение и анализ корпуса коротких текстов (постов микроблогов) для задачи разработки и тренировки тонового классификатора

Common Knowledge Russian Tweets data http://study.mokoron.com/

Batuhan Guler et al.. 2019. *Frisio41 at SemEval-2019 Task 6: Combination of multiple Deep Learning architectures for Offensive Language Detection in Tweets*. Imperial College London

Pinkesh Badjatiya et al.. 2017. *Deep Learning for Hate Speech Detection in Tweets*.

Association for Computational Linguistics 2016. *Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter*,

Davidson, Thomas and Warmsley, Dana and Macy, Michael and Weber, Ingmar 2017. *Automated Hate Speech Detection and the Problem of Offensive Language*. Proceedings of the 11th International AAAI Conference on Web and Social Media

Zeerak Waseem, Dirk Hovy 2016. *Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter*. Association for Computational Linguistics