Automated Building and Analysis of Ukrainian Twitter Corpus for Toxic Text Detection

Kateryna Bobrovnyk[0000-0003-3358-1035]

Taras Shevchenko National University of Kyiv, Istitute of Philology
Taras Shevchenko Blvd, 14, Kyiv, Ukraine

katherine.bobrovnik@gmail.com

Abstract. Toxic text detection is an emerging area of study in Inter-net linguistics and corpus linguistics. The relevance of the topic can be explained by the lack of Ukrainian social media text corpora that are publicly available. Research involves building of the Ukrainian Twitter corpus by means of scraping; collective annotation of 'toxic/non-toxic' texts; construction of the obscene words dictionary for future feature engineering; and models training for the task of text classification (com-paring Logistic Regression, Support Vector Machine, and Deep Neural Network).

Keywords: toxic text detection, text corpus, Twitter.

The purpose of this study is to create a Ukrainian text corpus based on posts from Twitter and to perform toxic text detection on it. This area of NLP is relatively new so there are few works concerning this topic [1, 2]. The scope of the work is to create a dictionary of Ukrainian obscene words based on posts from Twitter and to train a toxic text classier using methods of Machine Learning.

Text corpus is a central notion of corpus linguistics. Corpora play an essential role in Natural Language Processing (NLP) research as well as a wide range of linguistic investigations: sentiment analysis, topic modeling, machine translation etc. They provide a material basis and a test bed for building NLP systems. There are thousands of corpora in the world, but most of them are created for specic research projects[3] for a particular language and may not be publicly available.

The first stage of the research is to scrape Twitter posts of hand-picked users with Ukrainian tweets. Scraping was based on Kenneth Reitz's library[4]. The resulting corpus consists of 1.87 million tweets with additional meta-information about time, language, replies, retweets, likes, hashtags, URLs and author nick-names.

The second stage of the research involves corpus and text preprocessing. The data cleanup procedure contains the following steps:

– delete empty texts and duplicates;
– detect the language of each text using fastText model[5] and save texts which were detected as Ukrainian, Belarusian, Bulgarian, Serbian, Macedonian (due to inaccuracies of fastText model);

COLINS’2019, Volume II: Workshop. Kharkiv, Ukraine, April 18-19, 2019, ISSN 2523-4013 http://colins.in.ua, online

- perform standard preprocessing procedures such as tokenization, noise re-moval (multiple whitespace/punctuation/new line/quotes, turn all numbers to '0'), substitution (number/html/phone number/email replacers);
- delete texts which contain only URLs, numbers, emails and tags.

The third stage of the research is to annotate texts for further training of the model. Material was distributed amongst 33 people in order to avoid bias. The task was to label tweets as "toxic" (abuses, harassments, threats, obscenity, insults, cyberbullying and identity-based hate texts) or "non-toxic". In total, 55 153 tweets were annotated. To provide features for model training and, conse-quently, improve the accuracy of toxic text detection, a dictionary of obscene words was created. It is based on a list of word roots, word contractions or such combinations as prefix+root or root+suffix. Additionally, Levenshtein edit distance was used. It allows to find the most similar words to those from the dictionary.

The last stage of the research is to train a model that detects toxic texts.

The following steps were made:

- feature engineering (make word embeddings using TF-IDF, bigrams, tri-grams, count number of obscene words in a tweet, number of capitalized words in tweet, number of smiles in a tweet);
- training of models for Text Classification (comparing Logistic Regression, Support Vector Machine and Deep Neural Network);
- evaluation of models' accuracy using cross-validation and F1-score.

The best accuracy is 89% (due to small annotated material and imbalanced classes in training set of 91% of 'non-toxic' and 9% of 'toxic' texts), F1-score is 0.86. There is a room for improvement: to achieve better results, more annotated data is needed. Other possible future directions include generating new features and new classification methods.

References

1. Pradheep, T. and Sheeba, J.I. and Yogeshwaran, T. and Pradeep Devaneyan, S.: Automatic Multi Model Cyber Bullying Detection from Social Networks. In: Proceedings of the International Conference on Intelligent Computing, Salem, Tamilnadu, India. (2017) Available at SSRN: https://ssrn.com/abstract=3123710 or http://dx.doi.org/10.2139/ssrn.3123710
2. Kennedy, G. W., McCollough, A.W., Dixon, E., Bastidas, A.,Ryan, J.,Loo, C., Sahay, S.: Hack Harassment: Technology Solutions to Combat Online Harassment. In: Proceedings of the First Workshop on Abusive Language Online, pp. 73–77, Vancouver, Canada (2017)
3. Rubtsova, Y.: Constructing a corpus for sentiment classification training. SOFT-WARE SYSTEMS 1(109), 72-78 (2015)
4. Twitter Scraper, https://github.com/kennethreitz/twitter-scraper. Last accessed 13 April 2019
5. Language identification, https://fasttext.cc/docs/en/language-identification.html. Last accessed 13 April 2019