Detection of Abusive Speech for Mixed Sociolects of Russian and Ukrainian Languages

Bohdan Andrusyak[1], Mykhailo Rimel[2], and Roman Kern[1]

[1] Know Center,
Inffeldgasse 13 8010 Graz AUSTRIA
badnrusyak@know-center.at
rkern@know-center.at
know-center.at
[2] Grenoble INP
46 Avenue Félix Viallet 38031 Grenoble FRANCE
Mykhailo.Rimel@grenoble-inp.org
grenoble-inp.fr

Abstract. Uncontrolled use of abusive language is a problem in modern society. Development of automatic tools for detecting abusive and hate speech has been an active research topic in the past decade. However, very little research has been done on this topic for Russian and Ukrainian languages. To our best knowledge, no research considered surzhyk [3]. We propose to use unsupervised probabilistic technique with a seed dictionary for detecting abusive comments in social media in Russian and Ukrainian languages. We demonstrate that this approach is feasible and is able to detect abusive terms that are not present in the seed dictionary.

Keywords: Russian; Ukrainian; abusive speech

1 Introduction

Abusive language, including different swear words and hate speech is not a new phenomenon on the Internet. There are numerous reasons why such language might be considered undesirable, including but not limited to legal issues, issues related to the ease of searching information and preventing people (especially adolescents) from the negative impact harmful content. Before the rapid growth of the social media and lower volume of user generated content, the effort to manually detect and moderate comments or forum posts containing such language was manageable. Nowadays, however, due to the prevalence of the social media, it becomes economically infeasible to manually moderate all the comments. This has resulted in an active development of techniques of automated detection of abusive language [9] [2]. A lot of published papers are dedicated to the detection of hate speech and/or abusive language in comments

[3] We refer to surzhyk as to a specific mixed sociolect in Ukrainian-Russian bilingualism [5]

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2018, pp. 77–84, 2018.

in English. According to our best knowledge, there are no adequate methods for automatic detection of abusive language and/or hate speech for Russian and Ukrainian language. Development of such methods is a difficult task because several reasons:

- There are no labeled databases or corpora of comments/posts in social media with abusive content in these languages;
- Due to the peculiarities of word formation in Russian and Ukrainian languages, it is practically impossible to define a finite list of abusive words; [6] – people often invent new, obviously obscene words, by linking two common stems by an infix or attaching prefixes and/or suffixes.
- Use of surzhyk creates great variety of abusive words.

Similarly to other languages, people sometimes try to mask the use of the abusive language by using euphemisms, inverting the order of letters in a word, replacing letters with stars or other symbols etc. Due to the fact that often social media comments in Ukrainian social media environment is written in surzhyk, we treat Russian and Ukrainian languages the same. This approach is reasonable, because most of the stems of abusive words are the equal in these languages. The goal of our research is to develop an automated approach for detecting abusive and hateful speech in Russian and Ukrainian languages in social media. Given very limited availability of the labeled data, in this research we focus on the unsupervised probabilistic techniques by using a seed dictionary of abusive terms as input.

2 Related work

Literature describes many approaches for automatic detection of hate speech and abusive language in social media. There are solutions for very specific types of hate speech (for example, Jihadist hate speech [12]) as well as for general offensive speech and hate speech detection (for example, [3]).

For the cases when there is a limited availability of labelled data sets (as with Russian and Ukrainian languages) unsupervised learning methods are often used. Thereby we mainly focus on unsupervised methods in this section.

Unsupervised learning techniques are commonly used for the natural language processing task. For example in [13], authors use these techniques for the sentiment classification of Chinese Text. They showed that the unsupervised techniques produced results that are surprisingly close to the ones obtained by the use of supervised learning. At the same time, unsupervised techniques are not usually susceptible of being domain- or language-specific.

Many approaches for extracting features from the natural text rely on some sort of external knowledge. Seed dictionaries is a form of such an external knowledge, which are relatively cost-effective to be assembled. There are different application areas of such seed dictionaries. For example, they may be used for entity extraction, automatic translation and many other application. It is common to use seed dictionaries in conjunction with unsupervised learning

techniques. For example, [4] used unsupervised learning with a seed dictionary of entities of multiple classes for the task of pattern learning and entity extraction in informal text corpuses. Furthermore, this approach was used for detection of abusive and hate language by [8].

There is little research dedicated to the automated detection of the offensive language in Russian and Ukrainian texts. Existing work for these languages focus on other tasks. For example, [11] developed a system for classifying Russian text into thematic categories. However, their main focus was to detect the content that is illegal in Russia (e.g. related to selling narcotics on the Internet) rather than detecting abusive speech. They also targeted all web sites which usually contain large blocks of texts with a small number of orthographic and grammatical errors rather than social media resources, where most of the comments are short and often contain slang, mistakes and/or obscured offensive speech.

3 Methodology

The core idea of our approach is to use a seed dictionary of abusive terms in combination with unsupervised assignment of labels (abusive or not abusive) to social media comments and then iteratively expand the initial seed dictionary with abusive and obscene words. Our hypothesis is that inappropriate comments contain in many cases more than one abusive word, thus the likelihood of abusive words appearing together in a single comment is sufficiently high to extend the dictionary. An overview of the workflow of our method is depicted in Diagram 1.

Before applying any algorithms to the textual data, we applied the following pre-processing steps: we removed punctuation, numbers, emoticons and other non-alphabetic symbols. We choose to remove those features, even though they can be strong indicator of toxic speech, because we consider words to be the most important features. Moreover, it allows us to manually evaluate the correctness of our method. In order to check the influence of word formation in Ukrainian and Russian languages, we first trained our model with words as they were present in the dataset and thereafter by using the words reduced to their stem. For evaluation of the results, we split the dataset into training, validation and test datasets, where the validation and test datasets were manually labeled by multiple native speakers.

After the dataset was pre-processed, each comment was automatically labelled as abusive or non-abusive by applying the seed dictionary, where a comment was assigned to contain abusive language, if at least a single word from the dictionary was present in the comment. This approach may introduce some false positives in case when abusive word is used in decent context, such as quote, however such occurrences are extremely rare in the context of social media. Then based on these initial labels, the dataset was split into two classes: abusive and non-abusive. Then for every word that appeared in the dataset more times than a threshold, a likelihood of being in the abusive or non-abusive class

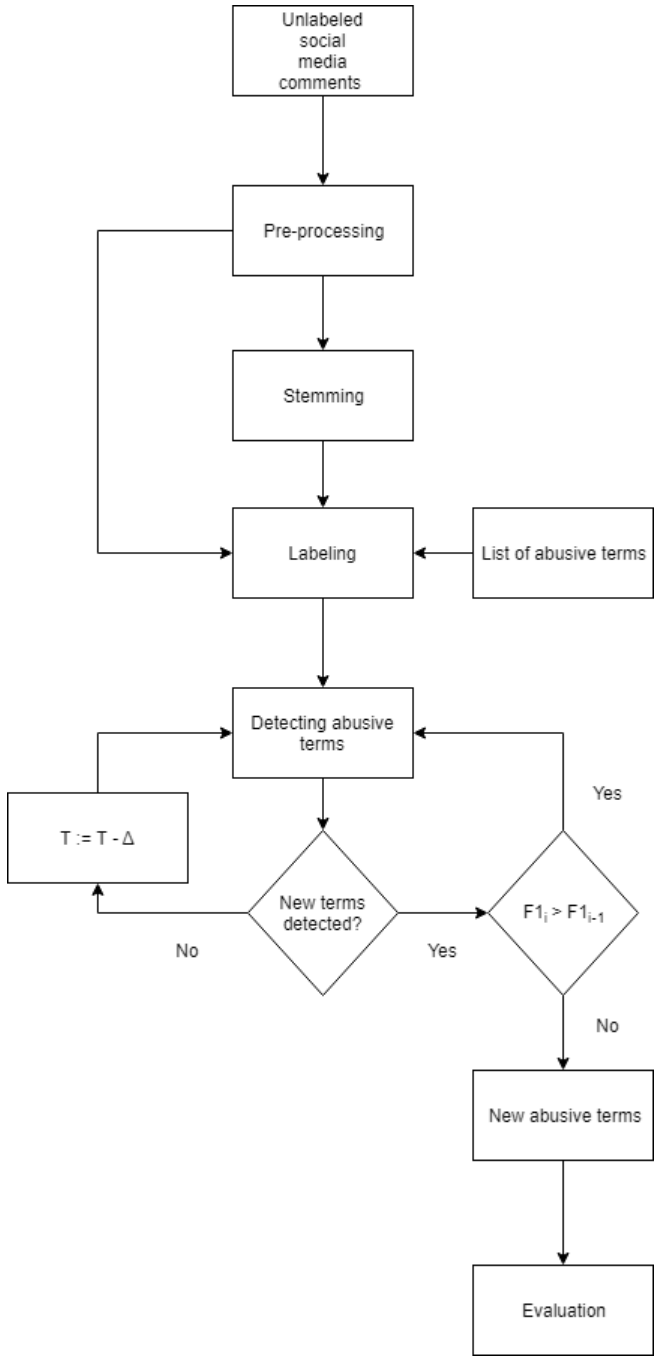

Fig. 1: High-level workflow

was computed. To that end, we used the frequency of the occurrences based on a threshold to filter out rare words that can be falsely classified as abusive.

Having initial likelihoods for each term we estimated its probability to be abusive. For the estimation we used relative distance and log odds ratio metrics. The formula of relative distance and log odds ratio are presented on Eq. (1) and Eq. (2) respectively:

$$P(x) = \frac{p_b(x) - p_g(x)}{max(p_b(x), p_g(x))} \tag{1}$$

$$L(x) = \frac{\frac{p_g(x)}{1-p_g(x)}}{\frac{p_b(x)}{1-p_b(x)}} \tag{2}$$

Where x is selected term, $p_b(x)$ - likelihood of x being in abusive part, $p_g(x)$ - the likelihood of x being in the non-abusive part.

In order to decide if term is abusive or not, we compare its estimation of being abusive to a given threshold T. This threshold was set to 0.95 based on the empirical data. If the estimation metric exceeds T, then the selected term is added to the dictionary of abusive words.

After all newly found abusive terms were added to the list, we evaluated our classifier on validation dataset. For evaluation we use precision, recall and F_1 score, which are calculated as in Eq. (3):

$$\begin{aligned} P &= \frac{tp}{tp+fp} \\ R &= \frac{tp}{tp+fn} \\ F_1 &= \frac{2 \cdot P \cdot R}{P+R} \end{aligned} \tag{3}$$

Where P is precision, R — recall, F_1 — F_1 score, tp — number of true positives, fp — number of false positives and fn — number of false negatives.

If the F_1 score is bigger than the F_1 score calculated on previous step, training dataset is re-labeled with expanded dictionary and estimation for each term are recalculated. If none of estimations exceeds threshold T, the threshold T is lowered by delta. If the F_1 score is lower than on the previous step, the iterative process stops and results are evaluated on the test dataset. The final dictionary then contains all words from the initial seed dictionary together with all newly found abusive words.

4 Experimental setup

In our experiments we assembled a dataset based on YouTube comments. We have chosen YouTube[4] since it was shown by [7] that flaming and abusive

[4] https://www.youtube.com/

language are common on this particular social media platform. We manually selected videos related to the topic of Ukrainian Revolution of Dignity, also know as Euromaidan [10]. Events of this revolution stirred a high controversy in all types of media, thus leading us to believe that YouTube videos on this topic will have high percentage of abusive language.

We collected comments from 329 videos, which were all related to the topic of Euromaidan. During exploratory analysis of collected comments, we noticed that some comments are written in transliteration - using English alphabet to write Russian or Ukrainian words. Such comments were deleted from our dataset. After cleaning and pre-processing our final dataset comprised more than 50,000 comments [5]. From this dataset 2,000 comments were randomly selected for manual labeling by native speakers [6]. The manual annotators found 32.7% of all comments contained abusive language.

For the seed dictionary we used crowdsourced list, which contained over 600 abusive and obscene words [7]. In addition, we also evaluated our approach on a minimal seed dictionary comprising only the 5 most top-used abusive words [8] in order to check whether our approach will still work with such a small seed dictionary.

The pre-processing, e.g. stemming, of the text of the comments were performed using the Natural Language Toolkit [1]. We used the settings for the Russian language for all terms in our dataset.

5 Results

We report the results of our approach for a number of different configurations, in order to assess the influence of various parameters and settings on the final performance. In Table 1 the results of our evaluation of our approach using different probabilistic estimation metrics are presented.

In our evaluation we found that stemming improves recall but at the same time reduces precision and overall slightly improves F_1 score. When comparing metrics of relative distance and log odds ratio, it can be seen that in all cases when using log odds ratio a high recall has been achieved, but the precision did drop considerably. With relative distance we achieved an improvement in recall and slight decrease in precision. When using the micro seed dictionary, we noticed that the use of log odds ratio metric leads to a rapid growth of the seed dictionary such that in the end it contained many non-abusive words. At the same time, the use of relative distance metrics provided more conservative results: the size of initial seed dictionary grew from 6 to 23 terms, where each term was indeed abusive.

We were surprised to find that our algorithm was able to pick up several ethnophaulisms that were not a part of initial seed dictionary but are definitely

[5] https://github.com/bohdan1/AbusiveLanguageDataset/blob/master/data.csv
[6] https://github.com/bohdan1/AbusiveLanguageDataset/blob/master/labled.csv
[7] https://github.com/bohdan1/AbusiveLanguageDataset/blob/master/bad_words.txt
[8] https://github.com/bohdan1/AbusiveLanguageDataset/blob/master/bad_words_seed.txt

Table 1: Results achieved using our approach
SD – seed dictionary
RD – relative distance
LOR – log odds ratio
STM – stemmed dictionary
MSD – micro seed dictionary
Words – number of new abusive terms added to the seed dictionary

Method	P	R	F_1	*Words*
SD	0.875	0.510	0.644	—
SD, RD	0.736	0.629	0.678	8
SD, LOR	0.678	0.629	0.652	11
STM, SD	0.742	0.629	0.681	—
STM, SD, RD	0.667	0.675	0.671	10
STM, SD, LOR	0.474	0.741	0.578	13
MSD	0.857	0.158	0.268	—
MSD, RD	0.588	0.463	0.518	44
MSD, LOR	0.303	1.000	0.465	573
STM, MSD	0.897	0.231	0.368	—
STM, MSD, RD	0.684	0.344	0.458	17
STM, MSD, LOR	0.308	0.920	0.462	145

offensive. Examples include words “хохлы” (khokhly) and “кацапы” (katsapy) which are derogatory names for people of Ukrainian and Russian nationalities respectively.

6 Conclusion and Future work

We demonstrated that unsupervised automatic labeling approach is a feasible choice for automatic detection of abusive speech in social media comments in Russian and Ukrainian languages as well as for surzhyk. As expected, we observed a slight drop in precision for the automatic population of the abusive word dictionary, there was a considerable gain in terms of recall. We found that the available pre-processing tools for the Slavic languages lag behind their counterparts for languages like English. For example, the stemming approach is based on a simple heuristic, which is not fully capable to match the Russian and Ukrainian languages. Considering everything mentioned above, we have outlined the following steps for our future research:

(i) develop a robust tool for lemmatization and stemming for both languages
(ii) develop algorithm for detecting hate speech aimed at nationality.
(iii) expanding dataset of social media comments and training word embedding models.

References

1. BIRD, S., KLEIN, E., AND LOPER, E. *Natural Language Processing with Python*, 1st ed. O'Reilly Media, Inc., 2009.
2. CHEN, Y., ZHOU, Y., ZHU, S., AND XU, H. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing* (Sept 2012), pp. 71–80.
3. DAVIDSON, T., WARMSLEY, D., MACY, M. W., AND WEBER, I. Automated hate speech detection and the problem of offensive language. *CoRR abs/1703.04009* (2017).
4. GUPTA, S., AND MANNING, C. D. Improved pattern learning for bootstrapped entity extraction. In *CoNLL* (2014).
5. LEWCZUK, P. Socjolingwistyczny status surżyka. *LingVaria*, 21 (2016), 177–189.
6. MOKIENKO, V. M. Russkaya brannaya leksika: tsenzurnoye i nyetsenzurnoye. *Rusistika* (1994).
7. MOOR, P. J., HEUVELMAN, A., AND VERLEUR, R. Flaming on youtube. *Computers in Human Behavior 26*, 6 (2010), 1536 – 1546. Online Interactivity: Role of Technology in Behavior Change.
8. MUBARAK, H., DARWISH, K., AND MAGDY, W. Abusive language detection on arabic social media. In *Proceedings of the First Workshop on Abusive Language Online* (2017), Association for Computational Linguistics, pp. 52–56.
9. PITSILIS, G. K., RAMAMPIARO, H., AND LANGSETH, H. Detecting offensive language in tweets using deep learning. *CoRR abs/1801.04433* (2018).
10. SHVEDA, Y., AND PARK, J. H. Ukraine's revolution of dignity: The dynamics of euromaidan. *Journal of Eurasian Studies 7*, 1 (2016), 85 – 91.
11. SIDOROVA, E., KONONENKO, I., AND ZAGORULKO, Y. An approach to filtering prohibited content on the web. In *DAMDID/RCDL'2017* (2017).
12. SMEDT, T. D., PAUW, G. D., AND OSTAEYEN, P. V. Automatic detection of online jihadist hate speech. *CoRR abs/1803.04596* (2018).
13. ZAGIBALOV, T., AND CARROLL, J. Automatic seed word selection for unsupervised sentiment classification of chinese text. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1* (Stroudsburg, PA, USA, 2008), COLING '08, Association for Computational Linguistics, pp. 1073–1080.