TUVD team at SemEval-2019 Task 6: Offense Target Identification

Elena Shushkevich
Social Media
Research Group
Technological University
Dublin, Ireland
e.shushkevich
@yandex.ru

John Cardiff
Social Media
Research Group
Technological University
Dublin, Ireland
john.cardiff
@it-tallaght.ie

Paolo Rosso
PRHLT Research Center
Universitat Politecnica
de Valencia, Spain
prosso@dsic.upv.es

Abstract

This article presents our approach for detecting a target of offensive messages in Twitter, including Individual, Group and Others classes. The model we have created is an ensemble of simpler models, including Logistic Regression, Naive Bayes, Support Vector Machine and the interpolation between Logistic Regression and Naive Bayes with 0.25 coefficient of interpolation. The model allows us to achieve 0.547 macro F1-score.

1 Introduction

Nowadays aggressive language on social media occurs more and more often. Categories of hate speech can be very diverse and can deal with a wide range of issues such as misogyny, sexual orientation, religion and immigration. Such types of speech can be found in posts in social networks, in Internet discussions, in comments on various articles and in responses to posts of famous persons.

This problem is receiving increasing amounts of attention and researchers are making attempts to build systems capable of recognizing such kinds of aggressive speech, offenses and insults in social networks.

This article presented our approach to hate speech detection, which we used for the challenge SemEval-2019 Task 6: OffensEval - Identifying and Categorizing Offensive Language in Social Media (Zampieri et al., 2019a ; Zampieri et al., 2019b).

The task consisted of three sub-tasks and proposed to investigate the data extracted from Twitter for creating a classification system.

Sub-task A had the aim to identify offensive language and there were 860 unmarked English tweets for testing. The post had to be non offensive if it did not contain any offense or profanity.

The main goal of the Sub-task B was to categorize offensive posts from Sub-task A (there were 240 English tweets for testing) to different offensive types:

- Targeted Insults and Threats in cases when a post insults or treats to an individual, a group or an organization;

- Untargeted in cases where a post has a non-targeting profanity and swearing.

Sub-task C focused on offense target identification. There were 213 English tweets which were marked as offensive in Sub-task A and Targeted Insult and Threats in Sub-task B for testing. The classification was for three different groups:

- Individual, when the target of the offensive post was a person;

- Group, when the target of the offensive message was a group of people considered as a unit;

- Other, when the target of the offensive tweet did not belong to any of the previous categories (e.g., a situation, an event, or another issue).

There are two datasets in English and in Spanish languages for analysis, and our team worked with English only. The training dataset included 13200 tweets, 4400 of them were offensive ones, 3876 messages were labeled as 'Target Insult and Threats' and 524 ones as 'Untargeted'. We

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 770–774
Minneapolis, Minnesota, USA, June 6–7, 2019. ©2019 Association for Computational Linguistics

focused our efforts on Sub-task C only, and the training dataset for it consisted of 2407 'individual' offensive posts, 1074 'group' ones and 395 tweets marked as 'other'.

The paper is organized as follows. Some relevant related works in the area are described in Section 2. Section 3 presents the preprocessing we applied for the dataset and the methodology we used for the model creating. In Section 4 the results are described and analyzed. In Section 5 we summarize our work and plan some steps for the future researches.

2 Related Work

Today there are a lot of promising works in the area of the hate speech recognition As was shown in (Fasoli et al., 2015), offensive language can be very diverse and the level of the messages offensiveness can depend on the context and the relationships between users who take part in the conversation.

For example, insults delivered in a sexual context are less offensive in cases where there is a conversation between partners. Some slurs have more offensive meaning in cases of conversations between a superior and a subordinate compared with conversations between friends and some groups of slurs are more acceptable then others.

Expanding the point that offensive speech is heterogeneous, the work (Clarke and Grieve, 2017) presented results which showed that there is a difference between racist and sexist posts: the sexist messages were more interactive (more personal) and more attitudinal (with authors opinion) than racist ones. From this article we can make a conclusion that the most popular linguistic feature in offensive language are question marks and question DO (when a sentence stars with the word do).

The work (Saleem et al., 2017) demonstrated that messages may not include slurs, but still be offensive. The authors took as training dataset messages from potentially vulnerable communities (like groups of Afro-American and plus-size users) and messages from haters of these communities (not included slurs only) and showed that the system of hate speech recognition based on traditional methods like Logistic Regression could indicate insult meaning on the posts without slurs. .

In addition, this work shows that it is possible to test dataset from one source using training set from another one. Authors checked this fact, used the training dataset from one source and the testing dataset from the another source. The results were quite good and it is allow us to say that it could be useful to add to our training dataset some comments from another social media to make predictions better.

At the Automated Misogyny Detection (AMI) Shared Tasks IBEREVAL-2018 (Fersini et al., 2018) and EVALITA-2018 (Caselli et al., 2018), some interesting approaches for offensive language detecting were presented. The main goal in these challenges was to detect misogynistic tweets and to classify tweets for different groups depending on a misogyny type (stereotypes and objectification, dominance, derailing, sexual harassment and threats of violence and discredit) and an insults target (the idea of this type of classification was to recognize misogynous tweets which offend a specific person and tweets which insult a group of people).

In (Pamungkas et al., 2018) it was shown that the results of the model based on Support Vector Machine were quite good and in the research (Frenda et al., 2018) the ensemble of models allow to achieve a high level of accuracy. In work (Shushkevich and Cardiff, 2018) it was presented the ensemble of Logistic Regression. Support Vector Machine and Naive Bayes model which shown quite good results.

It is necessary to add that models based on neural networks show good results of offensive language recognition, as it was shown in (Badjatiya et al., 2017), where the authors created the model based on Long Short-Term Networks (LSTMs) which use internal memory for capture the long range dependencies in sentences and it could be important for the hate speech detection. This approach allowed them to achieve very high results in sexist and racist tweets detection in comparison with classifiers such as Logistic Regression, Random Forest, SVMs and Gradient

Boosted Decision Trees (GBDTs).

3 Methodology and Data

As the preprocessing step we:

- converted the words to the lower case;

- used TF-IDF (Term Frequency - Inverse Document Frequency) for the vectorization;

- marked emojis with the word 'EMOJI';

- labeled some combinations of symbols like '!!! ' and '??? ', because they look like emotional expressions and could be presented as emojis too, and replaced them with the word 'EMOJI'

Our model presents an ensemble of some classic machine learning models:

- The model based on Logistic Regression (LR) (Wright, 1995; Genkin et al., 2007), this type of classifiers apply an exponential function to a lineal combination of objects extracted from the data.

- The model based on Naive Bayes (NB), whose advantages are an absence of big training dataset and speed calculations requirement (Hi and Li, 2007).

- The model presented an interpolation between LR and NB with 0.25 coefficient of interpolation as a form of regulation: trust NB unless the LR. This type of interpolation was shown in (Wang and Manning, 2012) where NB was combined with Support Vector Machine, but in our case the combination LR+NB worked better.

- The model based on Support Vector Machine (SVM), the effectiveness of which in the work with texts was described in (Joachims, 2002).

We blended all above-described models into one which indicated the belonging of a tweet to the classes according to the rule: we summarized probabilities of belonging to all three classed which all four models presented and divided this number by 4. A post was assigned a class with the highest average probability.

4 Results

The predicted results of F1-macro for the all 5 models are presented in Table 1.

As it shown the Blended model achieves the highest score (0.68), so we could conclude that our hypothesis was correct and an ensemble of models presented the best results for the task of offense target identification.

Also, the model which combine Logistic Regression and Naive Bayes achieves good result (0.65), and the worst model for this type of classification was Logistic Regression one.

The results of the challenge are presented in Table 2. Overall Accuracy for the test set was equal to 0.6478 and Macro-F1 was 0.547.

As we can see, the macro F1-score is less when predicted with the training dataset macro F1-score by 0.133, and this difference could be connected with the small number of tweets for training. Also it should be noted, that the results of classification have a strict correlation with the number of testing examples: the IND classifier works better then GRP one and much better then OTH classifier, because in the testing dataset there were more data about individual target of offenses then about group and other targets.

5 Conclusion

To sum up, we created an ensemble of models, which allow as to achieve quite good results being placed 25th out of a total of 65 participants. We showed that the idea of blending simple models based on Logistic Regression, Naive Bayes and Support Vector Machine gives a perspective in the area of hate speech recognition in the identification of the target of offensive messages.

As the next steps in our research, we are planning to expand the preprocessing step and use some dictionaries and lists of offensive language, which could help us to achieve better results. We also intend to additional data for the training datasets.

It is interesting to add, that in these datasets all links were replaced with URL and all usernames

Model	F1 (macro)
Logistic Regression	0.50
Naive Bayes	0.63
LR+NB	0.65
Support Vector Machine	0.60
Blended Model	**0.68**

Table 1: Results for each model with training dataset for the Subtask C

Type of classification	F1-score
GRP	0.6047
IND	0.7615
OTH	0.2759
avg/totall	**0.6243**

Table 2: Results of the classification with testing dataset for the Subtask C

in tweets were replaced with USER. It could be useful to investigate, for example, links, which were mentioned in offensive messages. It could be possible to expand our dataset in cases when link was a respond for another offensive post or we could lable tweets which have links for a blocked content.

In this challenge we faced the problem of the an insufficient quantity of tweets to make our classifier work better: for example, for the class Other there were only 395 post for training. We believe that an increase in the volume of data could make our modeling more effective, and external data sources could be helpful. Also, we intend to experiment with the use of LSTMs.

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion pp. 759760.*

Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso. 2018. Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI)). In *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18) CEUR Workshop Proceedings. CEUR.org*, volume 2263, Turin, Italy.

Isobelle Clarke and Jack Grieve. 2017. Dimensions of abusive language on twitter. In *American Psychological Association*, Washington, DC.

Fabio Fasoli, Andrea Carnaghi, and Maria Paola Paladino. 2015. Social acceptability of sexist derogatory and sexist objectifying slurs across contexts. In *Language Sciences, 52.*

Elisabetta Fersini, Maria Anzovino, and Paolo Rosso. 2018. Overview of the Task on Automatic Misogyny Identification at IberEval). In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018) CEUR Workshop Proceedings. CEUR-WS.org*, volume 2150, Seville, Spain.

Simona Frenda, Bilal Ghanem, and Manuel Montes-y Gomez. 2018. Exploration of Misogyny in Spanish and English tweets. In *CEUR Workshop Proceedings. CEUR-WS.org*, volume 2150, Seville, Spain.

Alexander Genkin, David Lewis, and David Madigan. 2007. Large-scale bayesian logistic regression for text categorization. In *Technometrics, 49(3):291304.*

Zhang Hi and Di Li. 2007. Naive bayes text classifier. granular computing. In *IEEE International Conference on, pages 708708. IEEE.*

Thoarsten Joachims. 2002. Learning to classify text using support vector machines: Methods, theory and algorithms. In *Kluwer Academic Publishers.*

Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile, and Viviana Patti. 2018. P14-ExLab@UniTo for AMI at IberEval2018: Exploiting Lexical Knowledge for Detecting Misogyny in English and Spanish Tweets. In *CEUR Workshop Proceedings. CEUR-WS.org*, volume 2150, Seville, Spain.

Haji Mohammad Saleem, P. Dillon Kelly, Susan Benesch, and Derek Ruths. 2017. A web of hate: Tackling hateful speech in online social spaces. In *CoR abs/1709.10159.*

Elena Shushkevich and John Cardiff. 2018. Classifying Misogynistic Tweets Using a Blended Model: The AMI Shared Task in IBEREVAL 2018. In *CEUR Workshop Proceedings. CEUR-WS.org*, volume 2150, Seville, Spain.

Sida Wang and Christopher D. Manning. 2012. Baselines and bigrams: simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers, ACL, vol. 2, pp. 9094.*

Robert Wright. 1995. Logistic regression. In *Proceedings of the 26th International Conference on World Wide Web Companion pp. 759760.* L.C.Grimm.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL.*

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval).*