Hope at SemEval-2019 Task 6: Mining social media language to discover offensive language

Gabriel-Florentin Pătraș[1], Diana-Florina Lungu[1]
Daniela Gîfu[1,2,3], Diana Trandabăț[1]
[1]Alexandru Ioan Cuza University of Iași, Romania
[2]Institute of Computer Science of the Romanian Academy, Iași Branch, Romania
[3]Cognos Business Consulting S.R.L., Romania
{patras.gabriel.florentin, lungu.diana.florina,
daniela.gifu, dtrandabat}@info.uaic.ro

Abstract

User's content share through social media has reached huge proportions nowadays. However, along with the free expression of thoughts on social media, people risk getting exposed to various aggressive statements. In this paper, we present a system able to identify and classify offensive user-generated content.

1 Introduction

With the constant spread of social media, users are spending increasing amounts of time on various social networking sites aiming to connect with peers, to share information or common interests. While users benefit from their use of social media by interacting with and learning from others, they are also at the risk of being exposed to large amounts of offensive contents.

Considering that people are negatively affected by harmful contents, detecting online offensive language to protect users online safety becomes an urgent task. To address concerns on people's access to offensive content over the internet, social media administrators often need to manually review online texts to detect and delete offensive materials. However, manually reviewing and identifying offensive messages is a highly human and time consuming task. Some automatic content filtering software packages have been developed to detect and filter offensive WebPages or paragraphs, mostly word-based approaches.

The "OffensEval: Identifying and Categorizing Offensive Language in Social Media" task at the SemEval 2019 competition (Zampieri *et al.*, 2019a) focuses on detecting and classifying offenses, pervasive in social media.

In this paper, we present a system able to identify whether a tweet is abusive language or not, and if abusive, if it is offensive or not. We trained a model to differentiate between these categories and then analyzed the results to better understand how we can improve the system.

The rest of the paper is organized as follows: section 2 presents other projects related to offensive language identification, section 3 presents the project's data set and methods, section 4 presents the results we have obtained and a short analysis, followed by our last point represented by section 5 with the conclusions.

2 Related Work

This topic has attracted significant attention in recent years, evidenced by increasing number of recent publications and a several scientific events such as ALW and TRAC workshops.

Offensive language is often subdivided into various intercalated categories, since different subtasks have been grouped under this label. One of the most analyzed such language is "hate speech", i.e. discriminative remarks, such as the racist or sexist ones (Norbata *et al.*, 2016).

Based on work on hate speech, cyberbullying and online abuse, Waseem *et al.*, 2017 proposses a typology that captures central similarities and differences between subtasks and discuss its implications for data annotation and feature construction. Additionally, Waseem *et al.* (2017) emphasize the practical actions that can be taken by researchers to best approach their abusive language detection subtask of interest.

Lexical detection methods for the offensive language tend to have low precision because they fail to classify messages not containing listed offensive terms. On the other hand, various

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 635–638
Minneapolis, Minnesota, USA, June 6–7, 2019.

machine learning methods are used in the literature, from Logistic regression, Naïve Bayes, Decision Trees, Random forests, SMVs to neural networks. Previous analysis of hate speech modeling (Schmidt and Wiegand, 2017) shows that there is a too wide range of features used, and a more advanced feature relevance analysis was needed (Waseem *et al.*, 2017).

A first shared task on aggression identification aiming to classify aggressive speech into overt, covert or no aggression was held at the TRAC Workshop collocated with COLING 2018 (Kumar *et al.*, 2018). 130 teams registered to participate in the task, 30 teams submitted their test runs and 20 teams sent their system description paper, which are included in the TRAC workshop proceedings.

The problem of distinguishing general profanity from hate speech is not a trivial task (Malmasi and Zampieri, 2018) and requires features that capture a deeper understanding of the text not always possible with surface grams.

3 Data set and Methods

The data set for SemEval 2019 task 6 was formed from 14100 tweets, 13240 training instances, retrieved from social media and distributed in tab-separated format and 860 tweets for testing (Zampieri *et al.*, 2019b). Using this data set, we were able to identify offense, aggression and hate speech in user generated content.

This section presents our approach for the different subtask, for each submission we uploaded.

Sub-task A: Offensive language identification

Submission 1. We analyzed the training data to identify specific words or expressions for offensive, respectively non-offensive tweets. Based on these expressions, we crafted a set of rules consisting in exact or partial matches of these expressions in the test corpus. Tweets that have complied with these rules have been annotated as offensive. Tweets containing such expression only in a negated form were annotatd as non-offensive. The rest of the tweets were randomly classified in offensive or non-offensive. The application code was written in the Java programming language and the results are presented in Table 1.1.

Submission 2. We created a lexicon based on two lists of words of offensive lexicons[1], freely available online, along with the list resulted from the analysis of the set of training tweets, as described above. Using these offensive words or expressions, we developed patterns and we classified the tweets in offensive tweets and non-offensive tweets. If the tweet was containing at least one word from the lists, it means that the tweet is offensive, otherwise the tweet would be considered not offensive. The results are presented in Table 1.2.

Submission 3: For this submission, we used the same lists of offensive words obtained from external sources, along with the list of offensive words found in the training data, but we put a restriction on the size of the words (more than 4 letters). This constraint was considered due to the fact that we noticed that they introduced noise in the non-offensive tweets. Additionally, we used WordNet to obtain the synonyms of the words we had in our lists. The results are presented in Table 1.3.

Sub-task B: Automatic categorization of offense types

Submission 1: We tokenized the tweets annotated with targeted offensive words and collected different lists of cue words. Additionally, we noticed that if the tweet contained a proper name towards the middle of the sentence, the tweet was marked as a targeted tweet; otherwise it was marked as an untargeted tweet. We used this restriction and made the first submission, with the results presented in Table 2.1.

Submission 2: For the second submission, we tokenized the test tweets and checked if those words were found in the list of pronouns[2]. If a tweet was containing a pronoun from that list, then that tweet was marked as a targeted offensive one, otherwise it was marked as an untargeted offensive tweet. The results are presented in Table 2.2.

[1] One available at the GitHub repository for the paper (Davidson *et al.*, 2017) and one from Luis von Ahn (2018), consisting on English terms that could be found offensive on websites.

[2] https://www.really-learn-english.com/list-of-pronouns.html

Submission 3: We separated the tweets in words and counted how many words begin with a capital letter. We didn't take into consideration the "#" (hashtags) and @(@USER) because the vast majority were written with a capital letter. If a tweet was containing at least 2 words with capital letter, then the tweet was marked as being a targeted offensive tweet, otherwise was marked as an untargeted offensive tweet. The results are presented in Table 2.3.

Sub-task C: Offense target identification

We created two lists with pronouns. One list was used for the personal pronouns in singular for and the second one for the personal pronouns in plural. Therefore, we obtained 3 scenarios:

- If the tweet contains a personal pronoun from the singular pronoun list, then the tweet is marked IND.
- If the tweet contains a personal pronoun from the plural pronoun list, the tweet is marked GRP.
- If the tweet does not contain any pronouns from the above lists then the tweet is marked as OTH. The results are presented in Table 3.

4 Results

Below are the results for each individual level using the test set. We report Precision (P), Recall (R), and F-measure (F) for each baseline on all classes along with weighted averages and Macro-F1. The result for sub-task A are presented in table 1, the results for sub-task B are presented in table 2 and the results for sub-task C are presented in Table 3.

Sub-Task A: Offensive language identification

	P	R	F	Samples
NOT	0.7398	0.4952	0.5932	620
OFF	0.2966	0.5500	0.3854	240
Avg./ Total	0.6161	0.5105	0.5352	860

Table 1.1: Results Sub-Task A – Submission 1.

	P	R	F	Samples
NOT	0.7876	0.5323	0.6352	620
OFF	0.7324	0.6292	0.4435	240
Avg./ Total	0.6634	0.5593	0.5817	860

Table 1.2: Results Sub-Task A – Submission 2.

	P	R	F	Samples
NOT	0.7718	0.6984	0.7333	620
OFF	0.3746	0.4667	0.4156	240
Avg./ Total	0.6610	0.6337	0.6446	860

Table 1.3: Results Sub-Task A – Submission 3.

Sub-Task B: Automatic categorization of offense types

	P	R	F	Samples
TIN	0.8571	0.4789	0.6145	213
UNIT	0.0826	0.3704	0.1351	27
Avg./ Total	0.7700	0.4667	0.5605	240

Table 2.1: Results Sub-Task B – Submission 1.

	P	R	F	Samples
TIN	0.9091	0.4695	0.6192	213
UNIT	0.1308	0.6296	0.2166	27
Avg./ Total	0.8215	0.4875	0.5739	240

Table 2.2: Results Sub-Task B – Submission 2.

	P	R	F	Samples
TIN	0.9211	0.3286	0.4844	213
UNIT	0.1280	0.7778	0.2199	27
Avg./ Total	0.8318	0.3792	0.4547	240

Table 2.3: Results Sub-Task B – Submission 3.

Sub-Task C: Offense target identification

	P	R	F	Samples
GRP	0.3333	0.0128	0.0247	78
IND	0.4815	0.3900	0.4309	100
OTH	0.1860	0.6857	0.2927	35
Avg./ Total	0.3787	0.3005	0.2595	213

Table 3: Results Sub-Task C.

5 Conclusions

The offensive language in social media commonly comes from an unpleasant condition or something that is disgusting or forbidden. We discussed the challenges in detecting offensive language including the abusive words writing patterns in social media.

This paper presents our system participating at SemEval Task 6. We present simple baseline scores on all classes in all of the three sub-tasks.

In the future, we would like to make a comparison between our system and datasets annotation for similar tasks such as aggression or abusive identification and hate speech detection.

As further work, we have already started to study how to use the datasets for applying deep learning techniques to improve our results, based on word embedding, similar to the work presented in (Badjatiya *et al.*, 2017).

Acknowledgments

This survey was partially supported by a grant of the Romanian Ministry of Research and Innovation, CCCDI – UEFISCDI, project number PN-III-P1-1.2-PCCDI-2017-0818/73PCCDI (ReTeRom), within PNCDI III and by the README project "Interactive and Innovative application for evaluating the readability of texts in Romanian Language and for improving users' writing styles", contract no. 114/15.09.2017, MySMIS 2014 code 119286.

References

Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. 2017. *Deep learning for hate speech detection in tweets*. In Proceedings of the 26th International Conference on World Wide Web Companion, pages 759–760

Davidson, T., Warmsley, D., Macy, M. and Weber, I. 2017. *Automated Hate Speech Detection and the Problem of Offensive Language*. In Proceedings of ICWSM.

Kumar, R., Ojha, A.K., Malmasi, S. and Zampieri, M. 2018. *Benchmarking Aggression Identification in Social Media*. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC), pages 1-11.

Luis von Ahn Research Group (accessed 2018) *Offensive/Profane Word List*, available online at https://www.cs.cmu.edu/~biglou/resources/bad-words.txt?fbclid=IwAR3yLdiB5lsgoQjWIJXgYLPb6Pl4jK-MCT5INw_Lfkfet6A8mvHsB-hyJVY

Malmasi, S., Zampieri, M. 2018. *Challenges in Discriminating Profanity from Hate Speech.* Journal of Experimental & Theoretical Artificial Intelligence, Vol. 30, Issue 2, pages 187-202. Taylor & Francis.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. 2016. Abusive language detection in online user content. In Proceedings of the 25th International Conference on World Wide Web. pages 145–153.

Schmidt, A. and Wiegand, M. 2017. A survey on hate speech detection using natural language processing. In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Association for Computational Linguistics, Valencia, Spain, pages 1–10.

Waseem, Z., Davidson, T., Warmsley, D. and Weber, I. 2017. *Understanding Abuse: A Typology of Abusive Language Detection Subtasks*. In: Proceedings of the Abusive Language Online Workshop.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R. (2019a) *SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)* in Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval) 2019.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R. (2019b) *Predicting the Type and Target of Offensive Posts in Social Media*, in Proceedings of N AACL 2019.