bhanodaig at SemEval-2019 Task 6: Categorizing Offensive Language in social media

Ritesh Kumar
Department of CSE
IIT(ISM) Dhanbad
India, 826004
ritesh4rmrvs@gmail.com

Guggilla Bhanodai
Department of CSE
IIT(ISM) Dhanbad
India, 826004
bhanodaig@gmail.com

Rajendra Pamula
Department of CSE
IIT(ISM) Dhanbad
India, 826004
rajendra@iitism.ac.in

M. R. Chennuru
Department of CSE
IIT(ISM) Dhanbad
India, 826004
cmr.mahesh@gmail.com

Abstract

This paper describes the work that our team bhanodaig did at Indian Institute of Technology (ISM) towards OffensEval i.e. identifying and categorizing offensive language in social media. Out of three sub-tasks, we have participated in sub-task B: automatic categorization of offensive types. We perform the task of categorizing offensive language, whether the tweet is targeted insult or untargeted. We use Linear Support Vector Machine for classification. The official ranking metric is macro-averaged F1. Our system gets the score 0.5282 with accuracy 0.8792. However, as new entrant to the field, our scores are encouraging enough to work for better results in future.

1 Introduction

Social media has become most popular among users in these days. Based on survey (Johnson et al., 2011), it has been observed that 70% of teenagers use social media sites on daily basis. Users share their views with help of social media like twitter, facebook, instagram, youtube. Ritesh et al. (Kumar et al., 2018a) tried to identify hate speech. On the one hand Users get benefited from social media by learning or interacting with other users on the other hand they face offensive online contents. With exponential growth of social media it has become quite significant to identify and categorize offensive language in social media.

A key challenge among researchers is to automatically categorization of offense type languages in social media. few research have been performed but it is still a hot topic among researchers. keeping it in mind, we develop a system that could categorize offensive language in social media. The relevant shared task description, data and results are described in the paper (Zampieri et al., 2019b).

In this paper, we use Linear Support Vector Machine (LSVM) for classifying and identifying offensive language in social media. We use snowball stemmer to find out root words. Also, we have used unigram and bigram language models without stopwords.

The rest of the paper is organized as follow. Section 2 describes related work. The proposed methodology and used data is described in section 3. Section 4 describes results obtained after experiment. Finally, we conclude and future work in section 5.

2 Related Work

The interest in identifying and categorizing aggression, cyber-bullying and hate speech, particularly on social media, has been growing in recent years. This topic has attracted attention from researchers interested in linguistic and sociological features of aggression, and from engineers interested in developing tools to deal with aggression on social media platforms. In this section, we review a number of studies and briefly discuss their findings. For a recent and more comprehensive survey on hate speech detection we recommend (Schmidt and Wiegand, 2017) and (Fortuna and Nunes, 2018).

Davidson et al. (Davidson et al., 2017) used crowd source to label a sample of tweets into three categories: hate speech, only offensive and those with neither. The Hate Speech Detection dataset used in (Malmasi and Zampieri, 2017) and a few other recent papers such as (ElSherief et al., 2018; Gambäck and Sikdar, 2017; Zhang et al., 2018).

A proposal of typology of abusive language sub-tasks is presented in (Waseem et al., 2017). For studies on languages other than English has been described in (Su et al., 2017) on Chinese and (Fišer et al., 2017) on Slovene. Finally, for recent discussion on identifying profanity vs. hate speech is discussed in (Malmasi and Zampieri, 2018). This work highlighted the challenges of distinguishing between profanity, and threatening language which may not actually contain profane

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 547–550
Minneapolis, Minnesota, USA, June 6–7, 2019.

System	F1 (macro)	Accuracy
All TIN baseline	0.4702	0.8875
All UNT baseline	0.1011	0.1125
LSVM	**0.5282**	**0.8792**

Table 1: Results for Sub-task B using model LSVM and best result is highlighted with boldface.

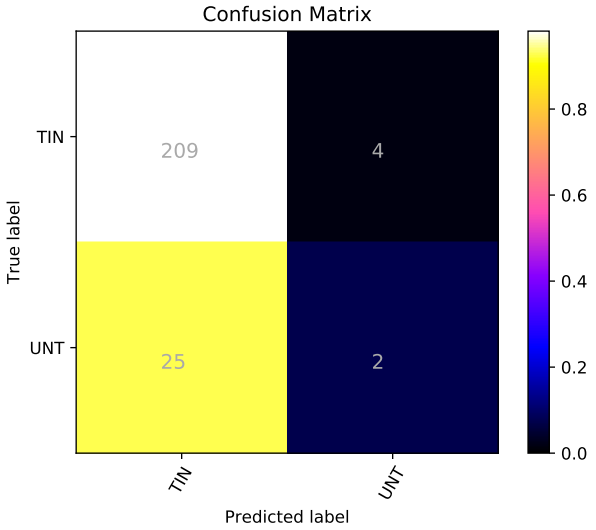

Figure 1: Sub-task B, LSVM

language.

Additionally, the related work has been performed in the related workshops such as TA-COS[1], Abusive Language Online[2], and TRAC[3] and related shared tasks such as GermEval (Wiegand et al., 2018) and TRAC (Kumar et al., 2018b).

3 Methodology and Data

The description of our system and different runs has been described in this section. We have been provided training dataset with 13,240 tweets, a trial set with 320, and a test set with 860 (Zampieri et al., 2019a). Each instance is composed of a tweet and its respective labels for tasks A, B and C. The three levels/subtasks are as follows:
Task A : Whether the tweet is offensive (OFF) or non-offensive (NOT).
Task B : Whether the tweet is targeted (TIN) or untargeted (UNT).
Task C : If the target is an individual(IND), group (GRP) or other (OTH; e.g., an issue or an organisation).
We have focused on subtask B. In our methodology, tweets are preprocessed by replacing following words with corresponding words shown below:
what's → what is
've → have
can't → can not
n't → not
i'm → i am
're → are
'd → would
'll → will
'scuse → excuse
followed by stemming words with snowball stemmer. LSVM is used for classification. We perform the task for only categorizing offensive language in social media, whether tweet is targeted insult or untargeted. SVM is categorizing offensive language in social media. For this, tf-idf of words unigrams and bigrams (without stopwords) that are occured at least 3 times are considered as features with 12 normalization.

[1] http://ta-cos.org/
[2] https://sites.google.com/site/abusivelanguageworkshop2017/
[3] https://sites.google.com/view/trac1/home

4 Results

In this section, we describe our experimental results. The official ranking metric is macro-averaged F1. We have included accuracy here as well for comparison. In table 1 , we see that we get the best result **0.5282** with accuracy **0.8792** using LSVM model. All TIN (targeted insult) baseline has got the score 0.4702 with accuracy 0.8875 and All untargeted baseline has got the score 0.1011 with accuracy 0.1125. The confusion matrix has been shown in figure 1.

5 Conclusion and future work

This year we participated in OffensEval sub-task B i.e. automatic categorizing offensive language in social media. We use LSVM model for classification. While there can be no denial of the fact that our overall performance is average, initial results are suggestive as to what should be done next. As we have taken ngrams for training set, our model unable to handle OOV (out of vocabulary) words. Pretrained word embeddings would have handled this problem. LSTM with these word embeddings might give better results. We explore these models in coming future.

References

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of ICWSM*.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media. *arXiv preprint arXiv:1804.04257*.

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2017. Legal Framework, Dataset and Annotation Schema for Socially Unacceptable On-line Discourse Practices in Slovene. In *Proceedings of the Workshop Workshop on Abusive Language Online (ALW)*, Vancouver, Canada.

Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys (CSUR)*, 51(4):85.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using Convolutional Neural Networks to Classify Hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90.

Timothy Johnson, Robert Shapiro, and R Tourangeau. 2011. National survey of american attitudes on substance abuse xvi: Teens and parents. *The National Center on Addiction and Substance Abuse*, 2011.

Ritesh Kumar, Guggilla Bhanodai, Rajendra Pamula, and Maheshwar Reddy Chennuru. 2018a. Trac-1 shared task on aggression identification: Iit (ism) @ coling18. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 58–65.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018b. Benchmarking Aggression Identification in Social Media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbulling (TRAC)*, Santa Fe, USA.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting Hate Speech in Social Media. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 467–472.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in Discriminating Profanity from Hate Speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30:1–16.

Anna Schmidt and Michael Wiegand. 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*, pages 1–10, Valencia, Spain.

Huei-Po Su, Chen-Jie Huang, Hao-Tsung Chang, and Chuan-Jie Lin. 2017. Rephrasing Profanity in Chinese Text. In *Proceedings of the Workshop Workshop on Abusive Language Online (ALW)*, Vancouver, Canada.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Langauge Online*.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of GermEval*.

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)*.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In *Lecture Notes in Computer Science*. Springer Verlag.