DA_Master at HASOC 2019: Identification of Hate Speech using Machine Learning and Deep Learning approaches for social media post

Apurva Parikh[1], Harsh Desai[1], and Abhimanyu Singh Bisht[1]

DA-IICT, Gujarat, India
{apurvakparikh,hsd31196,bisht2492}@gmail.com

Abstract. This paper describes the research that our team, DA_Master, did on the shared task HASOC, conducted by FIRE-2019, which involves identification of hate and offensive language in Twitter and Facebook posts. The task is divided into three sub-tasks. Our team conducted our experiments on an English dataset. We employed Machine Learning techniques like Logistic Regression and Navies Bayes classifier and a Deep Learning based approach which utilizes Convolutional Neural Networks. Our best model obtained Macro F1 score of 0.6472 for SubTask-A, 0.4068 for SubTask-B and 0.4303 for SubTask-C.

1 Introduction

In the past few years, there has been an exponential rise in the number of people who have uninhibited access to the internet. The arrival of the "Netizen" populace, and their affinity for interacting via social media platforms like Twitter, Facebook, Instagram etc. has led to the manifestation of diverse online user behaviour. Sometimes the online behaviour exhibited by individuals on social media platforms is hostile, targeting an individual(s) or a community by posting insults or threats. Studies have shown that hate/offensive messages are becoming extremely common on social media platforms, accounting for almost 500 million posts daily on Twitter[1]. Thus, the problem of detecting and stifling the propagation of hateful or offensive posts on social media is a pertinent issue for research.

To promote research in this field several competitions have been organized, such as OffenseEval[1], GermanEval[2]. Recently, HASOC [3] was organized as a shared task for FIRE 2019. The task is aimed at identifying hateful and offensive language in social media posts. The task was organized for three languages namely English, Hindi and German, but our team conducted our investigation only on the English dataset.

[1] https://sproutsocial.com/insights/social-media-statistics/

Organizers proposed a hierarchical model which consists of three sub-tasks:

- **TASK-A: Identification of hate and offensive language**, posts are classified as follows:
  - Hate and Offensive(HOF)
  - Not Hate and Offensive(NOT)
- **TASK-B: Categorization of Hate and Offensive**, posts identified as HOF in task-A are further categorized as:
  - Hate Speech(HATE)
  - Offensive(OFFN)
  - Profane(PRFN)
- **TASK-C: Identification of target**,in which posts identified as HOF in task-A are further classified on the basis of whether the post targets a particular individual/group or not.
  - Targeted Insult(TIN)
  - Untargeted Insult(UNT)

The rest of paper is organized as follows. In Section 2 we discuss related works and the dataset is discussed in Section 3. Proposed methods are presented in Section 4 and the results in Section 5. Finally, we conclude our work in Section 6.

2 Related Work

The past few years have seen an increase in aggressive user behaviour on social media platforms, usually exhibited via the use of targeted or untargeted hateful and offensive language. This has encouraged NLP researchers to work in this field and develop automatic detection systems to preemptively discard such posts. The term hate speech was coined by [4]. Hate speech detection proposed by [5] used a Convolutional Neural Network with Word2Vec for classifcation. Schmidt and Wiegand [6] gave a detailed survey on hate speech detection using Natural Language Processing. They identified various features for hate speech like words, sentiment, linguistic, etc. After performing various experiments they observed that, bag of words or word embedding based representation gave better classification results, a character-level approach works better than token-level, sometimes the text may not contain hate speech but it may be accompanied by images or other media which might contain hateful messages.

Recent work on the topic by [7] has focused on the type and target of the offensive language and was the first to categorically define offensive language. They created a dataset based on these distinctions which was used to train models utilizing SVM, BiLSTM and CNN.

3 Dataset

The datasets provided by the organizers [3] were annotated in a hierarchical fashion. Table 1 shows detailed analysis of the provided datasets.

Table 1. Details of Dataset Provided

Details	# Posts in Train Data	# Posts in Test Data
TASK A	5852	1152
Hate and Offensive posts (HOF)	2261	
Non Hate and Offensive posts (NOT)	3591	
TASK B	2261	
Hate (HATE)	1143	
Offensive (OFFN)	667	
Profane (PRFN)	451	
TASK C	2261	
Targeted Insult (TIN)	2041	
Untargeted insult (UNT)	220	

It can be observed from Table 1 that the train data for TASK-A consists of 61% Non Hate posts and 39% were classified as hateful, offensive or profane, whereas TASK-B has a distribution of approximately 50% Hate , 30% Offensive and 20% Profane posts, and more than 90% of the TASK-C dataset are posts which are targeted insults and the remaining are untargeted insults.
Therefore, only 39% of the given data can be used for training further tasks. There is a huge imbalance in TASK-C data.
The authors have provided 1153 posts for testing, which were to be filled based on a prediction of model.

4 Proposed Method

For our experiment we used two approaches. For the first approach we represented the input features using CountVectorizer and Tfidf, which were then used to train logistic regression and Naive-Bayes classifiers. The models used were imported from the Scikit-learn Python package.
Our second implementation took inspiration from [8]. Our model used a hierarchical CNN i.e 3 layers of 1D convolution. The words were represented using pre-trained GloVe [9] 200d vectors. No data pre-processing was performed by in any of the experiments.

5 Results

The organizers had allowed a maximum of 3 submission for each sub-tasks. Macro F1 Score was the primary metric for evaluation. Table 2 shows results obtained by our methods and that of top performed of each task. Figure 1, 2, 3 show confusion matrix of our best performing system for all three sub-tasks.

6 Conclusion

In this paper we have presented two approaches to solve the hierarchical task of detecting and classifying hate speech and offensive posts. The accuracy of our

Table 2. Results obtained for English Dataset

Method/ Technique used	Obtained Result (Macro F1 score)		
	TASK-A	TASK-B	TASK-C
Convolutional Neural Network	0.6216	0.4021	**0.4303**
CountVectorizer + Navie Bayes	0.6387	0.3137	0.4236
CountVectorizer + Logistic Regression	**0.6472**	**0.4068**	
Top Performer of each task	0.7882	0.5446	0.5111

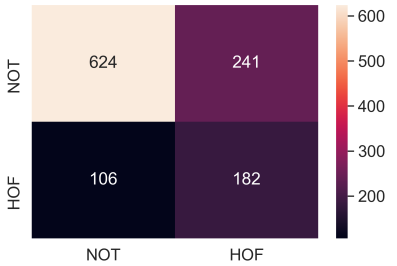

Fig. 1. TASK-A CountVectorizer + LR

Fig. 2. TASK-B:CountVectorizer + LR

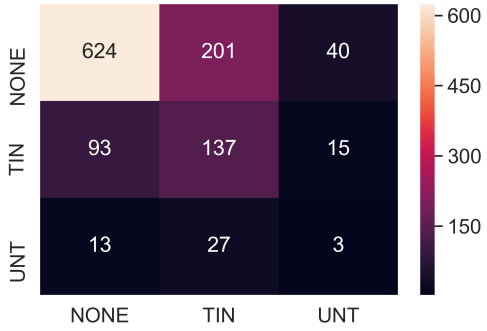

Fig. 3. TASK-C: Convolutional Neural Network

methods did not cross 70% for any of the sub-tasks. The methods used did not perform well when for the sub-task B, a possible reason for it may have been that, the model could not differentiate between profane, hateful and offensive posts properly. Our future experiments will try to solve the problem of class imbalance by using data augmentation and explore the efficiency of pre-trained language models like ULMFit, XLNet, RoBERTa for tackling this hierarchical task.

References

[1] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, 2019.

[2] Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. Overview of the germeval 2018 shared task on the identification of offensive language. 2018.

[3] Sandip Modha, Thomas Mandl, Prasenjit Majumder, and Daksh Patel. Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. In *Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation*, December 2019.

[4] William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26. Association for Computational Linguistics, 2012. URL https://www.aclweb.org/anthology/W12-2103.

[5] Björn Gambäck and Utpal Kumar Sikdar. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90. Association for Computational Linguistics, 2017. https://doi.org/10.18653/v1/W17-3013. URL https://www.aclweb.org/anthology/W17-3013.

[6] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, April 2017. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1101. URL https://www.aclweb.org/anthology/W17-1101.

[7] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Predicting the type and target of offensive posts in social media. pages 1415–1420, 01 2019. https://doi.org/10.18653/v1/N19-1144.

[8] Kim Yoon. Convolutional neural networks for sentence classification. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, 2014. URL http://aclweb.org/anthology/D/D14/D14-1181.pdf.

[9] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.