DEEP at HASOC2019 : A Machine Learning Framework for Hate Speech and Offensive Language Detection

Conference Paper · December 2019

Hamada Nayel
Benha University

H. L. Shashirekha
Mangalore university

DEEP at HASOC2019 : A Machine Learning Framework for Hate Speech and Offensive Language Detection

Hamada A. Nayel[1] $^{[0000-0002-2768-4639]}$ and Shashirekha H. L.[2]

[1] Department of Computer Science
Faculty of Computers and Artificial Intelligence
Benha University, Egypt
hamada.ali@fci.bu.edu.eg

[2] Department of Computer Science
Mangalore University, India
hlsrekha@gmail.com

Abstract. In this paper, we describe the system submitted by our team for Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC) shared task held at FIRE 2019. Hate speech and offensive language detection have become an important task due to the overwhelming usage of social media platforms in our daily life. This task has been applied for three languages namely, English, Germany and Hindi. The proposed model uses classical machine learning approaches to create classifiers that are used to classify the given post according to different subtasks.

Keywords: Multi-task Classification · Multi-lingual Text Analysis · Hate Speech and Offensive Detection [3]

1 Introduction

Wide spread use of internet is giving more freedom to people to express their thoughts freely and anonymously in different forms such as blogs, business networks, forums and social media such as Twitter and Facebook. Internet has enabled the interaction among people from different culture, race, religion, origin, gender and nationality. It has also paved the way for the online information to reach the mass within a matter of seconds. Social media is generating lot of content as it has opened up a new galaxy of opportunities for people to express their opinions online resulting in the exchange of ideas in a positive way as well as contributing to the propagation of hate speech in a negative way. The anonymity of users provided by the internet is also impacting the society in generating the web contents in a negative way by the usage of profane words, hate speech, derogatory terms, racists slurs, toxic, abusive and offensive language. This negative content may be aggressive and potentially harmful lowering the self-esteem of people leading to mental illness and suicidal attempts and has also forced people to deactivate their social media accounts. The increase of cyberbullying and cyberterrorism, and the use of hate speech content on the Internet, make the identification of hate speech an essential ingredient for anti-bullying policies of social media [5]. Many terrorist activities, which are related to hate speech, are gaining huge attention in social media in terms of posts and suggestions [1].

Hate speech and other offensive and objectionable content online are posing threats and challenges to the society. A lot of countries prohibit hate speech in social media subject to the condition that it should not target any group or trigger any crime. As hate speech acts as a opinion builder, many online forums such as YouTube, Face book and Twitter, have their own policies to remove hate speech content or anything which is impacting the society in a negative way.

It is therefore important to take preventive measures to cope up with hate speech and objectionable content or remove such content. Detecting hate speech is a difficult task as there is no clear definition for hate speech and can vary according to the context. For example, speech which contain subtle and nuance sentences are difficult to detect as hate speech. Further, the usage of slang terms may be found neutral, but may trigger hate crime or offense in some context. Although there is no universal definition for hate speech, the most accepted definition is provided by Nockleby [12]: "any communication that disparages a target group of people based on some characteristic such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic" [5].

Detection and removal of hate speech and objectionable content manually is a tedious task due to the massiveness of the web and the increasing number of online users. Further, the task becomes difficult due to the anonymity of users on the internet. Hence, there is a great demand for tools and techniques that automatically detect the hate speech and objectionable content quickly on the web to reduce the spread of hate speech.

In this paper, we describe the system submitted by our team for the shared task of Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC) track at FIRE 2019 [7]. Rest of the paper is organized as follows: In Section 2 we review the related work on hate speech and offensive language detection. Description of the task and a summary of dataset is given in Section 3. Our proposed model is discussed in Section 4 followed by the experiments and results in Section 5. Finally, we conclude the paper in Section 6.

2 Related work

Though, defining and understanding hate speech is difficult, several online forums, IT industries and researchers have explored many algorithms to detect hate speech, offensive and abusive content on the web [5, 2, 6, 3, 15, 4]. Thomas et.al. [2] trained a multi-class classifier to classify tweets into one of three categories, namely, hate speech, offensive but not hate speech, neither offensive and nor hate speech. They performed experiments using logistic regression, Naïve Bayes, decision trees, random forests, and linear SVM classifiers with 5-fold cross validation and obtained an overall precision 0.91, recall of 0.90, and F1 score of 0.90 for the best performing model. Gibert et.al. [5], constructed a manually labeled hate speech dataset composed of thousands of sentences obtained from Stormfront, a white supremacist online forum which contains hateful and no hateful sentences. Their model was built using Support Vector Machines (SVM), Convolutional Neural Networks (CNN), Recurrent Neural Networks with Long Short term Memories to annotate the test data.

Sean MacAvaney et.al. [6], examines the challenges faced by online automatic approaches for hate speech detection in text in addition to the existing approaches and various datasets for hate speech detection. They also have proposed a multi-view SVM approach that achieves near state-of-the-art performance, while being simpler and producing more easily interpretable decisions than neural methods. The experiments were conducted on HatebaseTwitter, Stanfront and TRAC datasets. Avishek Garain and Arpan Basu [3], presents a description of their system to detect offensive language in Twitter submitted to "SemEval-2019 Task 6 [14]" which uses bidirectional LSTM, a neural network based model to capture information from both the past and future context. Ziqi and Lei [15], discusses about the typical features responsible for classification between hate speech and no hate content. In their work they proposed Deep Neural Network structures serving as feature extractors that are particularly effective for capturing the semantics of hate speech. Their methods were evaluated on the largest collection of hate speech datasets based on Twitter, and are shown to be able to outperform the best performing method by up to 5 percentage points in macro-average F1, or 8 percentage points in the more challenging case of identifying hateful content. Aditya et.al., [4] proposes an approach to automatically classify tweets on Twitter into three classes: hateful, offensive and clean by considering features like n Grams and TF/IDF values. They performed a comparative analysis of the models considering several values of n in n-grams and TF/IDF normalization methods on Twitter dataset and achieved 95.6% accuracy. Areej Al-Hassan and Hmood Al-Dossari [1] in their survey paper, gives detailed information about hate speech in terms of insufficient dataset, complexities and challenges and the complete knowledge about text mining and Natural Language Processing techniques to find and classify hateful and no-hateful contents.

3 Task Description and Corpus

Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC) shared task aims at identifying the hate speech and offensive content in three Indo-European languages namely English, Germany and code-mixed Hindi. The shared task is divided into three sub-tasks to identify the hate speech and offensive content posts and classify them into various predefined categories. The first sub-task, Sub-task A is a coarse-grained binary classification that focuses on the identification of hate speech and offensive language in the given posts and is offered for English, German and Hindi language posts. Sub-task B is a fine-grained multi-class classification that classifies the hate speech and offensive language posts into one of the three predefined categories, namely hate speech, offensive and profane. This subtask is offered for English, German and Hindi language posts. The third sub-task, Sub-task C aims at checking the type of offense as targeted insult or un-targeted insult and is offered for English and Hindi. The corpus consists of posts collected from Twitter tweets and Facebook comments in English (EN), German (GR) and Hindi (HI) languages. Each post contains three labels for Sub-task A, Sub-task B and Sub-task C respectively. Table 1 gives a glimpse of the shared task and the associated categories in each sub-task.

Table 1. A glimpse of shared task and the associated categories in each sub-task

Sub-task	Class Labels	Description
A	NOT HOF	Contains neither hate speech nor offensive content Contains hate, offensive and profane content
B	HATE OFFN PRFN	Contains hate speech content Contains offensive language content Contains profane words
C	TIN UNT	Contains an insult to an individual, group, or others Contains non targeted profanity

4 Methodology

A detailed description of our model and the classifiers used are given in this section.

4.1 Problem Formulation

Given a set of posts $P = \{p_1, p_2, ..., p_n\}$, where each post is composed of a set of words $p_i = \{w_1, w_2, ..., w_k\}$, in each language $L = \{EN, GR, HI\}$ and the classes $A = \{NOT, HOF\}$, $B = \{HATE, OFFN, PRFN\}$ and $C = \{TIN, UNT\}$ for the sub-tasks A, B and C respectively, the shared task is to formulate a multi-label classification problem, where an unlabeled post/instance is assigned with multiple class labels one from each class A, B and C. ie., given a unlabeled post p_k, multi-label classification problem will assign the triple $< a, b, c >$ such that, $a \in A$, $b \in B$ and $c \in C$.

4.2 Model

The general structure of the proposed model is given in Fig. 1. This model creates sub-models for each subtask and then assigns multiple class labels, one for each subtask, by combining the output of each sub-model. The model consists of the following phases:

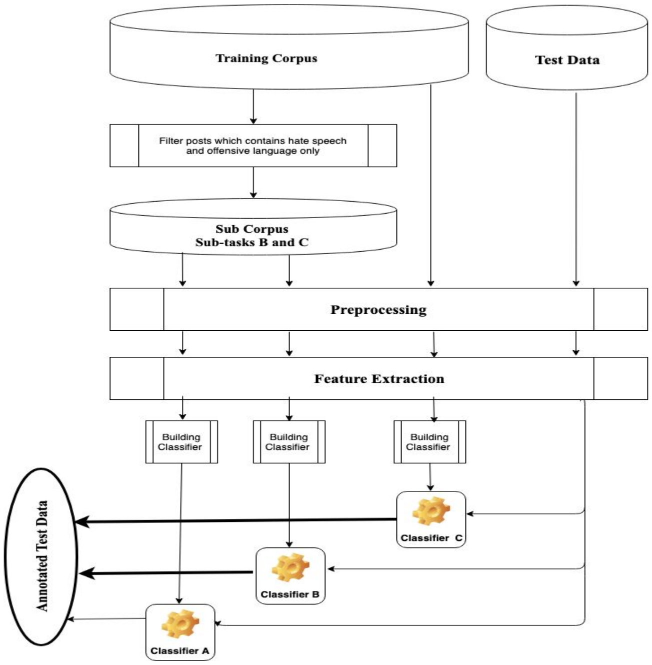

Fig. 1. The general structure of our model.

I. Preprocessing Preprocessing also known as corpus cleaning is a crucial and first step in building any classifiers. Each post p_k has been tokenized into a set of words to get n-gram ($n = 2$) bag of words. In this step, all un-informative tokens such as urls, digits and special characters have been removed from all the posts.

II. Feature Extraction In this phase, a TF/IDF vector has been computed for all the posts in the training set. This vector will be used as an input for training the classifier. TF/IDF has been calculated as described in [10].

III. Training the Classifier In this phase, the features that have been extracted in previous phase are used as input for training the classifiers. Three classification algorithms namely, Linear classifier, SVM and Multilayer Perceptron (MLP) have been used separately to train the proposed model. Linear classifier is a simple classifier that uses a set of linear discriminant functions to distinguish between different classes [13]. SVM is a kind of linear classifier which uses the samples close to the classes's boundaries for training and these samples are known as support vectors [11]. SVM has been used effectively in different NLP tasks [9, 8]. MLP is a deep learning approach that uses back propagation for training the neural network [10]. It is characterized by several layers of input nodes connected as a directed graph between the input and output layers.

Three different classifiers namely, Linear classifier, SVM and MLP are created for each subtask. Classifiers for Sub-task A will classify the unlabeled instance into one of the two predefined categories as mentioned in Table-1. While the entire data is used to construct classifiers for Sub-task A, a sub-corpus (only posts labelled as HOF in Sub-task A) that contain only the hate speech and offensive language posts is used as training corpus to create classifiers for Sub-tasks B and C. Classifiers for Sub-task B will classify the unlabeled instance into one of the three predefined categories and classifier for Sub-task C will classify the same instance into one of the two predefined categories as mentioned in Table 1.

5 Experiments and Results

In our proposed model, the Stochastic Gradient Descent (SGD) optimization algorithm has been used for optimizing the parameters of linear classifier, while SVM uses Lagrange multipliers to solve the optimization problem [13]. The loss function used in linear classifier was ”Hinge” loss function. Linear kernel has been used for SVM classifier. In MLP classifier the logistic function has been used as activation function using 20 neurons in the hidden layer.

We have used cross-validation approach with five folds to train different sub-models for each subtask and the outputs of each sub-task have been combined to generate the final output. The organizers of the shared task used Macro-averaged F1-score (M-f1) and weighted F1-score (W-f1) [10] as performance evaluation metrics for all sub-tasks. The dataset comprises of a set of tweets and Facebook comments in the three languages labelled with different categories. Statistics of the dataset is shown in Table 2.

Table 3 shows the M-f1 and W-f1 of our submissions for all sub-tasks over English, Germany and Hindi language posts. From the table, it is clear that SVM outperforms other classifiers for all subtasks for English. For German language posts, MLP outperforms other classifiers for Sub-tasks A and B. It is worth to note that, Sub-task C is not applied for Germany. For Hindi language posts, MLP outperforms other classifiers for Subtask A, while SVM reported better results than MLP and linear classifier for subtask B.

There is a gab between M-f1 and W-f1 for all subtasks except for the subtask A for Hindi. This is due to the fact that the performance of the subtasks B and C depends on the performance of the subtask A. If the prediction of subtask A is wrong then by default subtasks B and C will also have wrong prediction.

Table 2. Statistics of the Training Corpus

Sub-task	Class Labels	English	German	Hindi
A	NOT	3591	3412	2196
	HOF	2261	407	2469
B	HATE	1143	111	556
	OFFN	451	210	676
	PRFN	667	86	1237
C	TIN	2041	-	1545
	UNT	220	-	924
Total Samples		5852	3819	4665

Table 3. M-f1 and W-f1 for all classifiers and subtasks for English, Germany and Hindi

Subtask	Classifier	English		Germany		Hindi	
		M-f1	W-f1	M-f1	W-f1	M-f1	W-f1
A	SVM	**66.12%**	**73.02%**	45.62%	76.64%	75.49%	75.52%
	Linear Classifier	44.32%	65.13%	45.65%	76.70%	74.13%	74.11%
	MLP	62.61%	69.05%	**46.37%**	**76.92%**	**75.94%**	**75.98%**
B	SVM	**42.36%**	**67.69%**	22.81%	76.64%	**47.20%**	**59.00%**
	Linear Classifier	23.02%	64.94%	22.83%	76.70%	44.16%	56.05%
	MLP	34.50%	62.12%	**23.46%**	**76.92%**	46.63%	58.37%
C	SVM	**42.23%**	**69.89%**	-	-	51.72%	**69.67%**
	Linear Classifier	29.44%	64.95%	-	-	47.86%	68.63%
	MLP	39.75%	65.93%	-	-	**52.38%**	68.03%

6 Conclusion

In this paper, a machine learning approaches have been used for creating a model for detecting hate speech and offensive language content. Proposed model achieved good results compared to its simplicity. Extension of our work includes using deep learning approach to build the classifier and test it on much bigger dataset.

References

1. Al-Hassan, A., Al-Dossari, H.: Detection of Hate Speech in Social Networks: A Survey on Multilingual Corpus. In: Proceedings of 6th International Conference on Computer Science and Information Technology (CoSIT 2019). pp. 83–100. Dubai, UAE (February 23-24 2019)
2. Davidson, T., Warmsley, D., Macy, M.W., Weber, I.: Automated Hate Speech Detection and the Problem of Offensive Language. In: Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017. pp. 512–515. AAAI Press (2017)

3. Garain, A., Basu, A.: The Titans at SemEval-2019 Task 6: Offensive Language Identification, Categorization and Target Identification. In: May, J., Shutova, E., Herbelot, A., Zhu, X., Apidianaki, M., Mohammad, S.M. (eds.) Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019. pp. 759–762. Association for Computational Linguistics (2019), https://www.aclweb.org/anthology/S19-2133/
4. Gaydhani, A., Doma, V., Kendre, S., Bhagwat, L.: Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach. CoRR (2018), http://arxiv.org/abs/1809.08651
5. de Gibert, O., Perez, N., García-Pablos, A., Cuadros, M.: Hate Speech Dataset from a White Supremacy Forum. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). pp. 11–20. Association for Computational Linguistics, Brussels, Belgium (Oct 2018). https://doi.org/10.18653/v1/W18-5102
6. MacAvaney, S., Yao, H.R., Yang, E., Russell, K., Goharian, N., Frieder, O.: Hate speech detection: Challenges and solutions. PLoS ONE (2019)
7. Mandl, T., Modha, S., Patel, D., Dave, M., Mandlia, C., Patel, A.: 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)
8. Nayel, H., Shashirekha, H.L.: Improving NER for Clinical Texts by Ensemble Approach using Segment Representations. In: Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017). pp. 197–204. NLP Association of India, Kolkata, India (December 2017), http://www.aclweb.org/anthology/W/W17/W17-7525
9. Nayel, H.A.: NAYEL@APDA: Machine Learning Approach for Author Profiling and Deception Detection in Arabic Texts. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
10. Nayel, H.A., Shashirekha, H.L.: Mangalore University INLI@FIRE2018: Artificial Neural Network and Ensemble based Models for INLI. In: Mehta, P., Rosso, P., Majumder, P., Mitra, M. (eds.) Working Notes of FIRE 2018 - Forum for Information Retrieval Evaluation, Gandhinagar, India, December 6-9, 2018. CEUR Workshop Proceedings, vol. 2266, pp. 110–118. CEUR-WS.org (2018), http://ceur-ws.org/Vol-2266/T2-10.pdf
11. Nayel, H.A., Shashirekha, H.L., Shindo, H., Matsumoto, Y.: Improving Multi-Word Entity Recognition for Biomedical Texts. CoRR **abs/1908.05691** (2019), http://arxiv.org/abs/1908.05691
12. Nockleby, J.T.: Hate Speech. In: Encyclopedia of the American Constitution. pp. 1277–1279 (2000)
13. Theodoridis, S., Koutroumbas, K.: Chapter 3 - Linear Classifiers. In: Pattern Recognition (Fourth Edition), pp. 91 – 150. Academic Press, Boston, fourth edition edn. (2009)
14. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 75–86 (2019)
15. Zhang, Z., Luo, L.: Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter. CoRR **abs/1803.03662** (2018), http://arxiv.org/abs/1803.03662