Engineering and Technology Journal e-ISSN: 2456-3358
Volume 06 Issue 12 December-2021, Page No.-1100-1104
DOI: 10.47191/etj/v6i12.05, I.F. – 6.39

Review Paper on Hate Speech Detection

Ramakrishna Hegde[1], Bharath G[2], Kiran Kumar[3], Sai Charan[4], Chandan Y M[5], Soumyasri S M[6]

[1]Assoc Prof. Dept. of CSE, Vidyavardhaka College of Engineering, Mysore, India
[2,3,4,5]Vidyavardhaka College of Engineering, Mysore, India
[6]Asst. Prof, Dept. of CSE, Sapient College of Commerce and Management, Mysore, India

ABSTRACT: This is a review paper on the topic "Hate Speech Detection". One of the main disadvantages of social media is the way it is used to spread hate. This hate can affect an individual or a group in different ways like, degrading their mental health leading to anxiety and depression. This can lead to suicides or homicide. So it is very important to control how a platform can be used in spreading a particular message. To do this we have to identify the hate speech content automatically, this can be done with the help of techniques in machine learning and deep learning. We have reviewed few papers that deal with the different methodologies of detecting hate speech in a given text.

KEYWORDS: Convolutional Neural Networks(CNN), Long Short Term Memory (LSTM), (Recurrent Neural Networks)RNN, (Multilayer Perceptron)MLP, Multiple Kernel Clustering(MKC), Langragian Support Vector Machines(LSVM).

I. INTRODUCTION

With growth of users in social networks, there was also an increase in the hateful activities that permeate these communicative structures. Hate speech can be defined as any communication that deprecates a person or a group based on some characteristics such as race, color, ethnicity, gender, nationality, religion or other features. And main motive that encourages users to spread hate on social networks is anonymity, so users can spread hate on a particular target. For this reason, the hatred propagated can generate the irreversible consequences, where young people who approach with cyberbullying and homophobia, mainly commit suicide.

Social media platform like Facebook and Twitter deals with lakhs of posts that may contain abusive and offensive hate speech. Such content may affect one's self esteem and can hinder their mental health, which might lead to problems such as anxiety and depression.

According to the survey 34% of students are experiencing the cyber bullying in their life time and 18% of them are reported to have harmed themselves, which includes 1 out of 4 girls and 1 out of 10 boys. The research suggests that the suicide rate has been doubled since 2008 and making it the second most reason for death in the age group of 10-34 years.

The problem statement of our project states that with increase in the influence of social media we need to make sure that the content flowing through this platform is carefully monitored and it should be free from any kind of hate speech. So we came up with the idea of, "Hate Speech Detector". Our main motto is "POST NO HATE". For this we have gone through few papers and related work and reviewed the same here.

While going through some of the paper we got some knowledge of how they implemented the concept and their methodology which made us to understand the concept very well and to come up with the newer methodology with better implementation keeping their idea as the base.

II. RELATED WORKS

We have reviewed few papers on this topic, where they used convolutional neural networks and deep learning for identifying the hate speech in tweets and social media. Few of them are mentioned in the next part.

III. METHODOLOGY

Convolution Neural Network[1] is applied detecting the hate speech against Immigrants and Women on social media platform Twitter. The dataset used here consists of 9000 tweets in English and 4469 tweets in Spanish. The data was structured as Id, Text, Hate speech, Target range and Aggressiveness

id	text	HS	TR	AG
93874	@username stupid wish you die.	1	1	1
18267	Leftwing filth Deport them all. #Sendthemback	1	0	1
18345	1,500 migrants have died in Mediterranean in 2018	0	0	0

they have performed two tasks in which, Task A was to classify the tweets into English and Spanish, and the Task B was to classify the tweets into hateful and non-hateful and identify their aggressiveness towards a individual or a group.

The evaluation was done on the following matrix on the basis of accuracy, precision, recall and F1 score.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1\text{-}score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

Here TP, TN, FP, FN refers to True Positive, True Negative, False Positive and False Negative respectively.

True positive refers to the examples which are hateful and model recognizes it as hateful.

True negative refers to the examples which are not hateful and model recognizes it as not hateful.

False positive refers to the examples which are not hateful and model recognizes it as hateful.

False negative refers to the examples which are hateful and model recognizes it as not hateful.

For Task B we calculate Exact Match Ratio (EMR)

$$F1\text{-}score = \frac{F_1(HS) + F_1(AG) + F_1(TR)}{3} \quad (5)$$

$$EMR = \frac{1}{n}\sum_{i=1}^{n} I(Yi, Zi) \quad (6)$$

Here, Yi indicates i^{th} instance and Zi indicates labels to be predicted.

They have performed preprocessing of data where they removed all links, numbers, special characters and stop words.

Then they have performed word embedding which is a supervised statistical language model trained using Deep Neural Network. The purpose of this is to predict next word given the previous word, for this task they have used GloVe (Pennington et al., 2014) and FastText (Joulin et al., 2016) model with 300 dimensions.

In CNN they have used two layers of convolutions and two layers of pooling and further four filters were used, 2 of 3 dimensions and 2 of 4 dimensions. In these two dense layers, the first one has 512 neurons with relu activation functions and dropout of 0.5, the second has sigmoid activation function for loss and optimization they have used binary cross entropy and RMSprop functions respectively.

In another application Deep Learning[2] approach is applied to identify the tweets which are hateful and can able to classify effectively. In this approach the dataset used were consists of tweets from twitter in English and Spanish. There are sets of tasks that has to be performed on this dataset, firstly a pre-processing work is performed where the data is cleaned i.e., emoticons, urls, other string etc., are removed and then the set of tasks are performed, they are

1. To detect hate speech
2. Other features of hate speech.

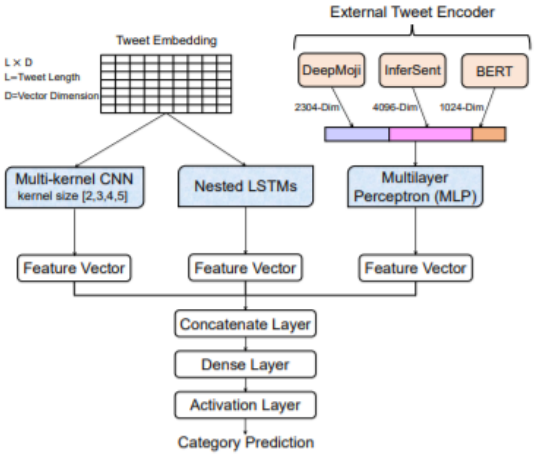
Task A was to predict whether given tweet is hateful or not. Here the input is given to a fully connected Neural Network of two dense layers with the relu activation function. The input text is hateful or not is obtained in the output layer with two units, relative to the number of classes. One with softmax activation function and another with relu activation function.

Task B was to classify how aggressive the tweet is and know the target individual. Here the aim is to analyze some features of hateful messages, as mentioned above.

For the detection of target of hate (TR), they have used the information of the part-of-speech by tagging process of the tweets. The sequence of labels was analyzed with a LSTM with RNN, obtaining a vector. Then, the obtained vector is concatenated with output of task A and it is used as input to a dense layer with the relu activation function.

Finally, the output layer will be having two neurons with the softmax activation function. In this way, the prediction corresponding to the offensive target in the tweets will be obtained.

The group of researchers are used Neural Network Model [3] for detecting hate contents on twitter. the dataset used is nearly 9000 tweets contents, but they have used only the English part of the Dataset. The English part is further divided into training, validation and testing sets which contains 9000, 1000 and 2971 annotated tweets respectively. The authors have identified two tasks to be performed, Task A is to classify the input tweet as hateful or non-hateful, and Task B is to identify the target affected by tweet like individual or a group, each of the task is treated as a binary classification problem.



The authors have used the above framework to achieve the desired task, the input tweets are converted into a word embedding matrix of dimension L * D where, L is the length of the tweet and D is vector dimension.

This word embeddings are fed as input to Multi kernel CNN and Nested LSTM models to extract feature vectors which are more informative on the tweets and helps in identifying unique words which make the tweet hateful or abusive.

Apart from this the authors have also used tweet encoder models like DeepMoji, BERT and InferSent to encode each tweet into multidimensional vectors.

Multi-kernel Convolutions is performed on tweet embeddings, the authors have used 4 different kernel sizes 2, 3, 4 and 5. Multi-kernel convolutions are basically applying convolutions with different filter sizes, which can improve the effectiveness of the model over a single kernel convolution.

Two layers of Nested LSTM (Long Short Term Memory) are also applied on the tweet embeddings to extract feature vectors. The LSTM cells acts as memory elements and can read or write relevant information to a particular cell based on the previously seen data.

The DeepMoji tweet encoder performs supervision on 1246 million tweets, it also uses two Bi-directional LSTM layer to encode tweets into 2304 dimensions vectors.

The BERT (Bidirectional Encoder Representations from Transformers) is used to encode a tweet into 1024 dimensions vector.

The InferSent model used with fastText to encode a tweet into 4096 dimensions vector.

The output of these three encoders are concatenated and fed as input to Multilayer Perceptron Network, these are neural network with feed forward mechanism and also uses backpropagation to get feature vectors for the model.

Now they have concatenated all the feature vectors generated by Multi-kernel CNN, Nested LSTMs and Multilayer Perceptron and feed it to a Neural Network with softmax function and cross entropy loss function and then train the Neural Network. The loss function used is,

$$E(x^{(i)}, y^{(i)}) = \sum_{j=1}^{k} 1\{y^{(i)} = j\} \log(y_j^{\sim(i)})$$

where x(i) is the training sample with its true label y(i) which can be 0 or 1, they have used stochastic gradient descent (SGD) and Adam optimizer with a learning rate of 0.001 and dropout of 0.02 at softmax layer and L2 regularization with a factor of 0.01.

In another case , the group of researchers are categorized offensive languages on social media [4] . The training dataset consists of 13,240 tweets, a trail dataset of 320, and a test dataset of 860 tweets. 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 is to identify a tweet as offensive or not.
Task B is to identify a tweet as targeted or not.
Task C is to identify whether a tweet targets an individual or a group or others like organization.

In sub task of B they have replaced the short form notations with the corresponding literals as 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

followed by stemming of words with snowball stemmer. LSVM is used for classify offensive language in social media, whether the tweet is a targeted or untargeted. For this, the words in the form of unigrams and bigrams that are most frequent are considered as features with 12 normalization.

IV. RESULTS

In paper 1 the result they obtained by this CNN model is as follows.

Task A				
English				
Model	**F1**	**P**	**R**	**Acc**
CNN-FastText	0.488	0.628	0.574	0.520
Spanish				
Model	**F1**	**P**	**R**	**Acc**
CNN-GloVe	0.696	0.708	0.712	0.696

Here they obtained a better F1 score for Spanish data than the F1 score for English data in Task A, and the confusion matric obtained for Task A is

English			
Class	**F1**	**P**	**R**
hateful	0.617	0.465	0.916
not hateful	0.359	0.792	0.232
Spanish			
Class	**F1**	**P**	**R**
hateful	0.685	0.598	0.802
not hateful	0.707	0.817	0.622

The result obtained for Task B by this model is,

Task B		
English		
Model	**F1**	**EMR**
CNN-FastText	0.577	0.297
Spanish		
Model	**F1**	**EMR**
CNN-FastText	0.609	0.430

In paper 2 to evaluate the model, different techniques are used. For the task A, systems are evaluated using standard methods like accuracy, precision, recall and F1-score and it is ranked by f1 score.

The following fig shows the results obtained for each of the languages in the task A. In addition, the results of the system in the first place of the ranking are shown for each of the metrics.

Language	Task A			
	Acc	P	R	F1-score
English	0.453	0.545	0.516	0.39
Best-English	0.653	0.69	0.679	0.651
Spanish	0.723	0.717	0.722	0.718
Best-Spanish	0.731	0.734	0.741	0.73

For task B, systems are evaluated with two parameters: partial match and exact match ratio (EMR). In partial match, each dimension which has to be predicted i.e. HS, TR and AG is evaluated separately with others using standard evaluation techniques including accuracy, precision, recall and F1-score, and then combined. In exact match, all the dimensions that have to be predicted are considered. The submissions are ranked by the EMR measure.

Language	Task B	
	F1-score	EMR
English	0.532	0.268
Best-English	0.467	0.57
Spanish	0.74	0.618
Best-Spanish	0.755	0.705

In paper 3 the evaluation criteria used is Accuracy, Precision, Recall and F1 Score, the authors have conducted training with each component at a time and removing other components, thereby calculating the F1 score for each component individually, then they have also calculated the F1 score of the model with the entire component working simultaneously. Here is the result of the performance of the models,

Method	Accuracy	Precision	Recall	F1-Score
MKC-NLSTMs-MLP	0.493	0.633	0.555	0.440
−MKC	0.441	0.507	0.503	0.381
−NLSTMs	0.483	0.630	0.548	0.423
−MLP	0.495	0.636	0.557	0.443
−MKC−NLSTMs	0.458	0.507	0.505	0.431
−MKC−MLP	0.475	0.592	0.537	0.419
−NLSTMs−MLP	0.485	0.640	0.549	0.423

As seen in the above table, the model without MLP has a good accuracy and F1 score than the entire model; this indicates that the pre-trained MLP model contributed negatively. The model without MKC or NLSTMs shows decrease in F1 score; this indicates that the model with them is working efficiently. So the author suggests using MKC upon NLSTMs with no MLP which has better performance than other combinations.

In paper 4 the metric used for evaluation of the model is F1-score and accuracy with 0.5282 and 0.8792 respectively. 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.

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

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

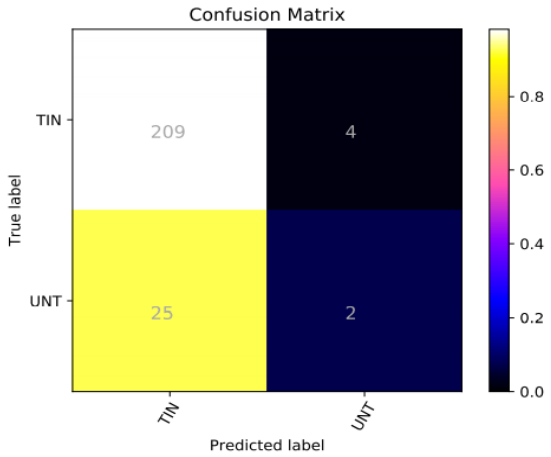


Figure 1: Confusion matrix.

V. CONCLUSION

All the referred papers share a common goal of identifying hate texts and the targets affected by such content. The main evaluation criteria used to judge the models are F1 score, Accuracy, Precision and Recall.

The model discussed in Convolution Neural Networks[1] related work, it is shows that relatively less complex and easy to construct with use of external tweet encoder and simple convolutions achieve a decent F1 score with GloVe model for Spanish than English language.

The model discussed in the Deep Learning Approach[2] is also uses both English and Spanish dataset and is built using LSTM with RNN, which turns out to be very effective for Spanish language with better accuracy and F1 score, but not much effective for English language.

The model discussed in Neural Network[3] concept is much more complex to build, it uses Multi kernel convolutions, NLSTMs, Multilayer perceptron and uses some external tweet encoders as well, the model is trained only on English language and it has achieved a decent accuracy and F1 score for various different combinations of the architecture, however the model with MKC and NLSTMs turns out to be the most efficient combination of all.

The model used in Categorizing the offensive language[4] contents on social media platforms uses LSVM to perform the desired task and operates only on English language, although it is a algorithm driven approach the model turns out to be decent in identifying the targets insulted by the hate speech with good accuracy and F1 score.

ACKNOWLEDGMENT

We would like to express our deep sense of gratitude towards all the scholars, our project mentor Dr Ramakrishna Hegde, Department of Computer Science and Engineering, VVCE, Mysore and Dr Pankaj Dwivedi, Central Institute of Indian Languages, Mysore for their guidance and assistance.

REFERENCES

1. Convolutional Neural Networks for Hate Speech Detection Against Women and Immigrants on Twitter. Alison P. Ribeiro and Nadia F. F. da Silva https://www.aclweb.org/anthology/S19-2074.pdf
2. Identifying Hateful Tweets with a Deep Learning Approach. Gretel Liz De la Pena Sarrac ˜ en https://www.aclweb.org/anthology/S19-2073.pdf
3. A Neural Network Model for Detecting Hate Speech in Twitter, Umme Aymun Siddiqua, Abu Nowshed Chy and Masaki Aono https://www.aclweb.org/anthology/S19-2064.pdf
4. Categorizing Offensive Language in social media, Ritesh Kumar, Guggilla Bhanodai, Rajendra Pamula and M. R. Chennuru https://www.aclweb.org/anthology/S19-2098.pdf
5. Datasets-http://hatespeechdata.com