JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

HATE SPEECH ANALYSIS AND REPORT GENERATION ON TWITTER

Pooja Chaudhari, Shivdeep Chaudhari, Pratik Fegade, Anand Kulkarni, Prof. Rahul Patil

ABSTRACT: As social media has become one of the most used media among youngsters and people, social media analytics has become a very crucial and important factor these days. Our project aims at analyzing the social media i.e. Twitter. As many people are spreading hate speech on social media, it has become necessity to analyze and keep track of these hate speeches and the users who are frequently performing hate speech This paper will give you a real-time approach to analyze the tweets on twitter based on hate speech detection and an analysis report will be generated of the users which are categorized as hateful users. In this paper we are using CNN LSTM hybrid architecture [1,2]. Also some generalized analysis of Twitter such as trending hashtags, people's perspective about any trending topics, wordClouds on any topic and much such type of pictorial reports will be generated. This paper will provide you a systematic approach for user specific real-time hate speech report generation and analysis.

KEYWORD: Hate Speech, CNN, LSTM, Glove, user specific report, worldClouds.

I. INTRODUCTION :

Over the last 15 years, the internet has brought us to various social media platforms for expression of thoughts, which fulfils the basic meaning of freedom of speech for an individual who is a regular internet user. One of the best platforms to express one's views and perspectives on a topic, be it political, social or anything else, is Twitter. Studies have revealed that on an average there are 60 tweets being tweeted at each second on Twitter. However, every tweet tweeted on twitter is not necessarily a genuine tweet. Some users deliberately put hate speech tweets on some topics, especially the topics which are trending on top.

Wikipedia has defined the hate speech as "The speech that attacks a person or group of people by their race, religion, cast, gender, nationality, ethnicity, etc. is a hate speech." Basically, it is the text which might affect a particular group of people badly, which might lead to disturbance in overall peace or sometimes, violence too. Therefore, it is necessary to track such users who are deliberately and regularly, in order to identify them and create their hate speech report.

This system starts searching for hate speech suspects from top trending topics on twitter by classifying each tweet as either hate speech or non-hate speech. Users who are caught as suspects spreading hate speech are kept on record for further observation whether they continuously put hate speech on the social media or it was just once they did it in a while.

If the suspect doesn't cross the threshold limit of the hate speech in a given time, he is taken off the record of the suspects. If the suspect puts hate speech tweets more than the threshold of the hate speech count, the system generates the report of hate speech for the user. The system works on real time data, which is beneficial for more precise tracking of suspects.

Thus, the report generated using this tool might prove useful in identifying threats to peace of society by taking appropriate action against them on the basis of the report based on their hate speech tweets on Twitter.

II. PROPOSED SYSTEM :

To surmount the issue regarding user specific hate speech analysis , we are proposing our system which is based on real-time processing of tweets fetched through Twitter APIs and Glove embedding + CNN+ LSTM architecture [1,2] for hate speech classification and report generation based on frequency and other parameters of hateful user and their tweets. Our proposed system has 4 basic modules:

1. Modules:
    1.1. Authentication Module
    1.2. Classification Module
    1.3. Real-time user tracking Module
    1.4. Report Generation and analysis module
2. Dataset:

2.1.1.Authentication Module: As we are going to use twitter API in this process we need to authenticate our web application with twitter by hosting the web application through Twitter Developer Account. For this we need to provide authentication through the consumer key, consumer secret, access token and access token secret which we get while hosting the web application on the developer account. In order to keep these parameters secret we are going to use pickle in order to convert file into non readable format

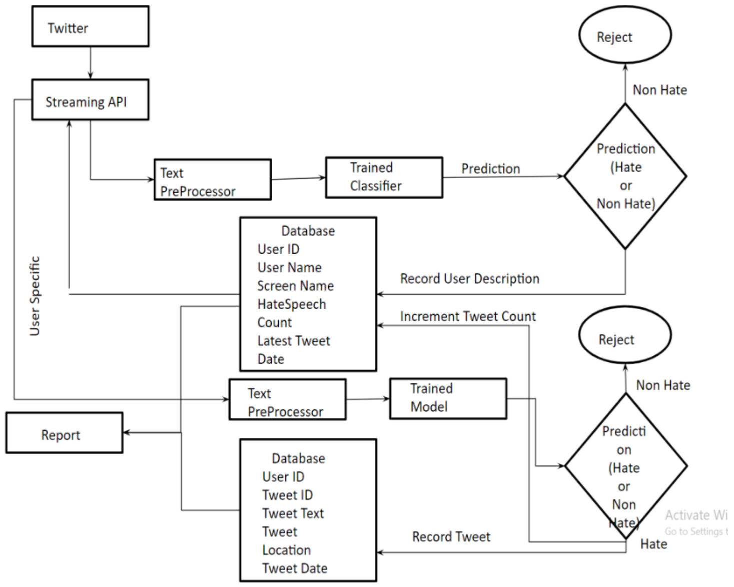

fig 1. System architecture diagram

2.1.2.Classification Module: In order to classify tweets fetched as hate or not we are using Deep Neural Networks [3]. As the accuracy of NPL classifier depends on how well the data is processed, we are using various preprocessing methods to remove URLs, numbers, punctuation marks, user tags etc. This helps to improve the accuracy of classification models. The preprocessed data is then given as input to the classification model but before that we need to perform embedding.

Classification model is trained on a dataset which is combination of multiple dataset used in [2,3,4,5].Classification model gives accuracy of 94 % indicates that it can correctly classify approximately 94 tuples out of 100.As we know that Deep Neural Network such as CNN does not work on textual data, they need real vector as input. Thus to convert text into numbers we are using the Glove model to map words to number vectors [3]. Here we are using Glove100 to map tweets into number vectors of size 100. A convolutional neural network and long short term memory RNN is used to extract the features from the input vector.

In this CNN acts as the best features extractor. So it will extract the feature by using different feature maps. We are using max pooling after CNN so that for the further processing, features with larger value will be considered. This will improve the processing speed as important features are only considered. Output of max pooling will be given as input to the LSTM layer. LSTM will then model the sequential information using its different gates like forget gate, keep gate etc. Finally the output of LSTM after performing the max pooling will be given as input to a dense layer with activation function as sigmoid. Sigmoid function calculates the probability of a particular tuple to be classified as of a particular class. This is how the classification module works. In order to improve the accuracy of the model a spatial dropout layer is added which purposely ignores or removes some words from the training data in order to circumvent the model from being biased and reliant to specific words. Also during validation hyperparameter tuning is done to improve the accuracy.

Before reaching to the decision of using Glove + CNN + LSTM architecture for classification, we have tried experiments on different methods such as CNN, LSTM,CNN+LSTM, Elmo + CNN + LSTM , Word2Vec +CNN+LSTM etc. A brief information about different approaches is given in experimental results.

2.1.3.Real-time user tracking Module: As we are going to track users who are frequently doing hate speech on social media, we need to first identify some suspect users. For that we are going to retrieve some tweets from some trending hashtags. These tweets are then classified into hate or not. For tweets which are classified as hate, we will store tweet_id, tweet_text, user_id etc into the database. After performing this process on several trending topics we will get a list of suspected users who are doing hate speech. Now we are going to track these suspected users in a real-time streaming process.

Twitter API provides us with an API for real-time streaming. Using this API we will track all the users simultaneously, thus we will get real-time access to all the tweets tweeted by suspected users. As soon as any tweet is created by any of the suspected users, our module will take that tweet and classify it as hate or non-hate. If the tweet is hate then we need to increment the count of the number of hateful tweets by 1. All of this process will continue to work unless the user stops it explicitly. Thus we can implement real-time user tracking.

2.1.4.Report Generation and analysis module : By tracking the users in a real-time streaming process, we would get a database containing hateful users, their user id, number of hateful tweets etc. By using these parameters in the database we can dig out some useful knowledge or patterns. In order to make this analysis comprehensible we are using some visualization tools like wordClouds, Pie-charts etc. This module will be an aid to understand the analysis done from social media.

Also an user specific report will be generated based on frequency of hateful comments or tweets which would be an aid to identify users which are

2.2. Dataset : Dataset used in our project is the dataset created by us by combining multiple dataset in order to generate unbiased and large dataset compared to other datasets. We have combined the Davidson dataset with the fox news commercial dataset by writing a python script and also combined the labels for comments/tweets.

The dimensionality of our dataset is 20499 x 2 . Two columns in dataset are “Tweets” and “Label”. In our dataset there are two classes “hate” and “noHate”. There are 14077 examples of class “noHate” and 6422 examples of class “hate” .

III. Experimental Results: To be comparable with results reported in our baseline model, we calculate macro-average Precision, Recall and F1 score over all classes in the dataset. The following tables compare our proposed model against the baseline model.
The highest figures are highlighted in bold.

Model	**Precision**	**Recall**	**F1**
CNN (word embedding)	93.0	92.8	92.9
CNN (Glove embedding)	93.1	93.2	93.2
LSTM (word embedding)	92.4	92.2	92.3
LSTM (Glove embedding)	93.0	93.0	93.0
CNN-LSTM (word embedding)	93.4	93.2	92.9
CNN-LSTM (Glove embedding)	**94.2**	92.8	**94.0**
CNN-LSTM$_{base}$ (embedding - learn)	93.4	92.9	93.1
CNN-LSTM$_{base}$ (embedding - ggl1)	94.2	93.9	92.1
CNN-LSTM$_{base}$ (embedding - ggl2)	94.0	94.1	94.0

tabel 3.1 Experimental results for classification

Also the Receiver’s Operating Characteristics is one of the performance measures to analyze the classification performance of the binary classifier. ROC is nothing but a probability curve and AUC i.e. Area under the curve which is nothing but the degree of separability between classes. High value of AUC indicates the model is good. For our Glove + CNN + LSTM is 97.3. Below are three AUC-ROC curves from our different experimental methods.

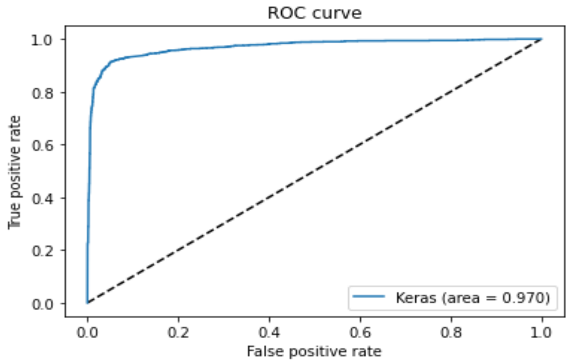

fig 2 .ROC curve for LSTM classifier(word embedding)

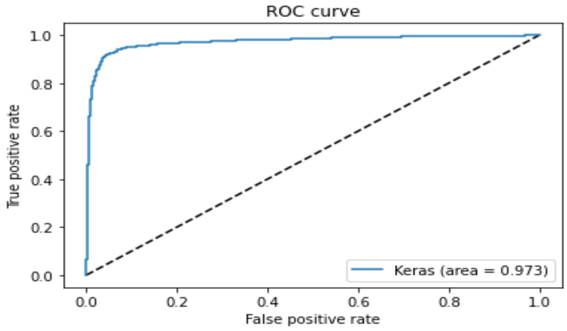

fig 3.ROC curve for LSTM classifier(Glove embedding)

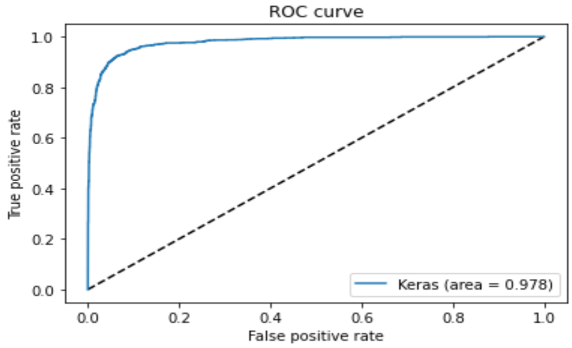

fig 4. ROC curve for Glove CNN LSTM classifier

IV. CONCLUSION AND FUTURE WORK :

We have introduced a general approach for learning the glove vector representations and have shown a large improvement when applying deep contextualized neural networks CNN-LSTM hybrid model. We have also confirmed that the model works more efficiently when used with embeddings from language modelling (ELMO) and improves the overall performance in accuracy but hampers the performance of the proposed system. Our model performs better than the pretrained embeddings used in the baseline model. We can finally conclude that deep convolutional neural networks utilizing glove vector models have good performance in the task of Twitter Hate Speech Detection. For future work, we tend to do hate speech detection on multilingual tweets.

V. REFERENCES :

[1]. Shivdeep Chaudhari, Pooja Chaudhari, Pratik Fegade, Anand Kulkarni, Prof. Rahul Patil , “Hate Speech And Abusive Language Suspect Identification And Report Generation” , International Journal Of Scientific & Technology Research Volume 8, Issue 11, November 2019

[2]. Zhang, David Robinson, Jonathan Tepper, “Hate Speech Detection Using CNN-LSTM Based Deep Neural Network”,Ziqi ACM The Web Conference WWW 2018, ACM New York.

[3]. Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Learning for Hate Speech Detection in Tweets. In Proceedings of the 26th International Conference on World Wide Web Companion (WWW ’17 Companion).

[4]. Manoen Horta Ribeiro, Pedro Calais, Yuri Santos "Characterizing and Detecting Hateful Users on Twitter"‖, 12th International AAAI Conference on Web and Social Media (ICWSM 2018).

[5]. Zhao Jian Qian, Gui Xiaolin "Deep Convolutional Neural Networks for Twitter Sentiment Analysis ", IEEE conference 2017

[6]. Zhang, Ziqi and Luo, Lei. "Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter". Semantic Web, vol. 10, 1 Jan. 2019 : 925 – 945.

[7]. Pitsilis, G.K., Ramampiaro, H. & Langseth, H. "Effective hate-speech detection in Twitter data using recurrent neural networks", *Appl Intell* **48,** 4730–4742 (2018).

[8]. Thomas Davidson, Dana Warmsley, Michael Macy, Ingmar Weber, "Automated Hate Speech Detection and the Problem of Offensive Language", Eleventh International AAAI Conference on Web and Social Media ,May 2017

[9]. Fabio Del Vigna, Andrea Cimino, Felice Dell'orletta , Marinella Petrocchi , and Maurizio Tesconi "Hate me, hate me not: Hate speech detection on Facebook", First Italian Conference on Cybersecurity (ITA SEC 17),2017