Hate-speech detection from social media post using CNN

Submitted by:

Ashraf Bin Shahadat– 16101199

MD Adnanul Anwar - 16101005

Eialid Ahmed Joy – 16101182

MD Mizanur Rahman Rony - 16101184

Supervised by:

Dr. Md. Golam Rabiul Alam

Associate professor

Department of computer science and engineering

Abstract

As the increasing number of social media user emerging day by day from various backgrounds and different diverse moral codes to today's wildly popular platforms, a space for hate space has emerged. Although social media platforms favors communication and information sharing, these are also used to launch harmful campaigns against individuals or specific groups. We aim at containing and preventing such hate campaigns. With the increasing amount of hate speech online, methods that automatically detects hate speech is very much required. Therefore, our goal is to use Natural Language processing to detect hate speech. We propose a Convolution Neural Network structure that will serve as a feature extractors which will be explicitly effective for capturing the context and the semantics of hate speech. The dataset we are using is a twitter dataset of 24784 tweets collected by Thomas Davidson in his paper "Automated Hate speech Detection and Problem of Offensive Language". Our classifier will assign each tweet as one of following categories: hate, offensive and neutral. More specifically, it will distinguish hate speech form normal text and can achieve higher classification quality than current state-of-art algorithms. Our system has obtained the F1 score of 0.482, 0.946 and 0.885 for hatred, offensive and neutral respectively. The accuracy of our system is approximately 91%.

Introduction

As the technology is growing day by day the communication through internet is also increasing. Every remote parts of the world are connected now to each other through massive number of social sites. Every minute, there are 510,000 comments generated on Facebook [1]. Also 350,000 tweets generated on Twitter in every minute [2]. Usage of social media is increasing with a very high rate and it is enlightening us in many ways. On the other hand, due to the easy access to the social media in the name of freedom of speech some people are spreading hatred. Moreover, people are belonging different community; ethnicity, religion and financial status are all in the same social sites. As a result, if someone has intensions to hurt any group he/she can use the social platforms to do the evil works. According to American Constitution "Hate speech is speech that attacks a person or a group on the basis of attributes such as race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity. It is a matter of concern that people are involving themselves into hatred so much now a days by sharing controversial topics against different group of people. Therefore a detection process for hatred content is highly needed for all social sites. Our target is to make a hatred detection system which will automatically check contents before uploading for the social sites and classify into three classes (Hate, offensive, neutral). The task is really challenging because of the complexity of natural language, also people use different ways or many different words to represent same meaning. For this reason we will use deep learning methods to detect hate speech. To implement our system we will use dataset from a paper done by Thomas Davidson. The dataset consists 24784 tweets from tweeter. We have modified our dataset using nltk model. For feature extraction we have used the built in FastText vector generator. Our system classifies the sentences into three class which are hatred, offensive and neutral. We have labeled neutral as 2, offensive as 1 and hatred as 0.

Research objective

We are doing research on Hate-speech detection from social media post. We have some objectives for our research that we will try to focus on and try to fulfill. They are-

- Make new model which will detect hate speech from social media post
- Raise the accuracy as much as possible
- Implement new methods
- Make or collect dataset in Bangla and find out hate speech from that data set
- Make our model fast so that we can use our model in real-time.

Literature Review

Spreading hatred in social media has been attracting the researchers for the last few years. There are lots of works on emotion detection, aggression, cyberbullying detection and so on. In[6] they have studied that using machine learning techniques such as Random Forest (RF) and Convolutional Neural Network (CNN) combined with Part-of-Speech (PoS) information can produce good results while detecting emotion from text. From [3], we have found that they mainly used convolutional neural network (CNN), bag of words and word2vec for feature extraction. They preferred CNN over recurrent neural networks. They stayed it needs a huge amount of data to train any model whereas RNN needs less than CNN. In addition, CNN works on the current inputs only but RNN works on both current and previously received inputs. In [4], they used glove instead of word2vec. Though both works almost same but there is still a small difference between word2vec and glove. Word2vec is a predictive model whereas glove is a count based model. From [5], we got to know some difficulties that every researchers face. They noted from their reviewed papers that many researchers failed to publish their weight initialization methods. For model's training, they initially used keras library with tensor flow backend but later they switched to theano to solve the issue with reproducibility of weight initializations. In [7] they have first divided posts into neutral and abusive posts. They have used four classes of features set such as N-grams, linguistic, syntactic and distributional semantics. Then they have further classified the abusive posts into hate, derogatory and profanity. In [8] they discussed that lexical methods cannot perform well while identifying hate speech from texts. Considering all we have develop a CNN based architecture for hatred and offensive language detection in English language.

Methodology

Pre-processing

Preprocessing goes through to raw data just to clear everything unnecessary in a dataset. We used NLTK (Natural language toolkit) library for our preprocessing.

Stopwords are , . a an the for in etc. Any complete sentence always consists of these stopwords. We have to remove all of these stopwords because there is nothing to deal with these kind of words in machine learning. So we imported stopwords from nltk library and removed every stopwords that was available. In addition, we also removed punctuation marks. After that, We imported RegexpTokenizer from nltk library and did the tokenization part. RegexpTokenizer removes punctuations, stopwords all before and after any word. Then we used pad sequence so that every word has same lengths.

Work flow

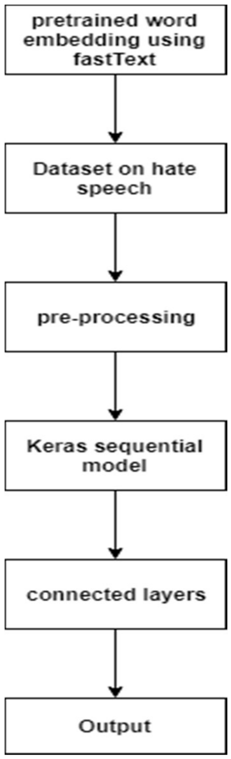

Fig1 : Workflow of our model

We managed our dataset from a github project. We pretrained word embedding using fastText library. Then we took train and test data from that github project and did the preprocessing of those data so that we can eradicate all kind of unnecessary data. we used keras sequential model to build our model. we feed embedded layer for training. Then sequentially feed conv1D, maxPooling, conv1D, conv1D, conv1D, globalMaxPooling and dense layer for output..

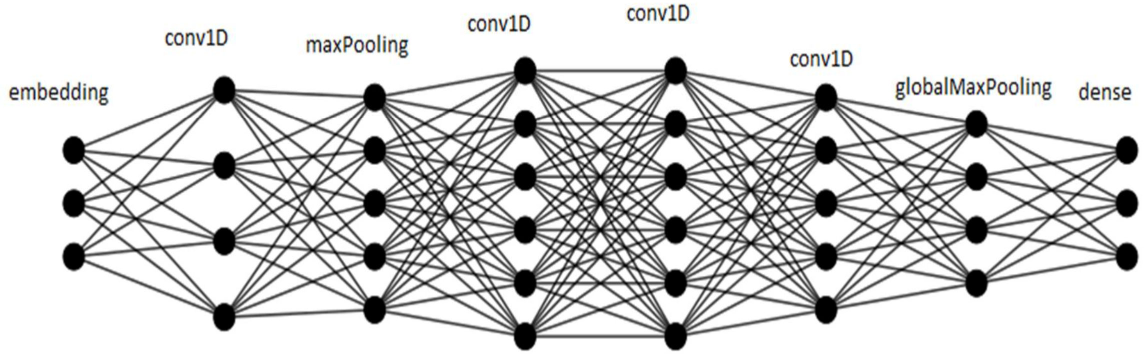

Fig2: Fully connected layer

In above figure, we explain basically how our model has been trained. Firstly, we feed our our vector file to embedding layer that we got from pretrained word embedding layer on first step of workflow. Then comes conv1D layer. After that, Maxpooling layer will take best accuracies from all and then ran conv1D for three times. In the next layer, GlobalMaxPooling takes all the best accuracies that is found as a whole. Dropout layer drops some of the accuracies that we found after globalmaxPooling so that we can get best accuracy possible. In the end, dense layer helps to create our genuine model which gives accuracy of 91%.

Implementation:

We used Python as our programming language and Jupyter Notebook as our IDE to run our program. We used CNN(Convolutional Neural Networks), FastText and Keras in our project. To run our project we need to set our environment first. We import –

- Numpy
- Pandas

- Keras
- tqdm
- NLTK
- Matplotlib.pyplot
- sklearn
- seaborn
- mlxtend.plotting

First of all in our project we used FastText for pre-trained word embedding. We have used FastText instead of Word2Vec as FastText makes vectors of both words and characters and assign values which will help us to increase the overall accuracy. After this we got 36084 word vectors.

Then we used pandas data frame and divide our whole dataset into training and testing. We took 17433 data as training data and 7350 data as testing data. We used Stopword and Tokenizer from NLTK to pre-process our data. Then we used keras sequential model to train our CNN model through 9 layers. We added embedding_1,1D convulation,1D maxPolling,1DGlobalMaxPolling etc layers and we made a dropout of 0.5.We used relu activation function as this is a multiclass model. Here we used Loss function = mean square error. We also used Adam optimizer.

Result:

During the analysis, we have seen that our overall testing accuracy is 0.9124 which is 91.24%. We used relu as activation function and 50 epoch to develop the accuracy. We used almost 25000 of labeled data. We get 0.482, 0.946 and 0.8850 f1 score for hateret, offensive_language,nither.

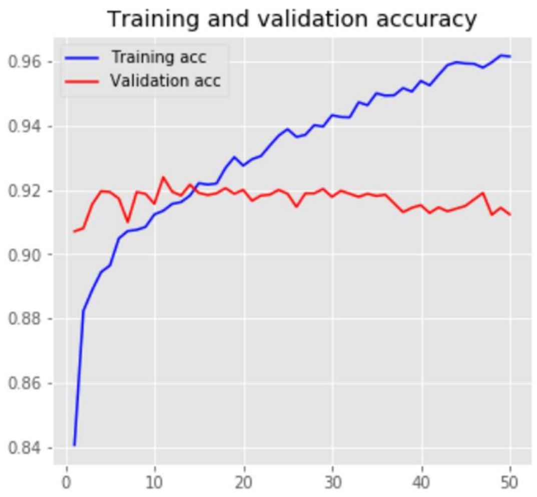

Fig3:Training and validation accuracy

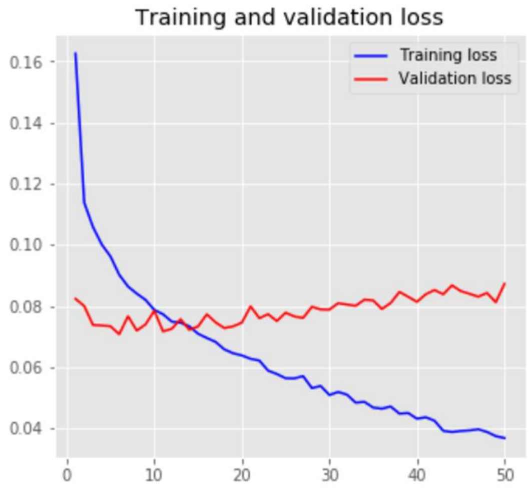

Fig4:Training and validation loss

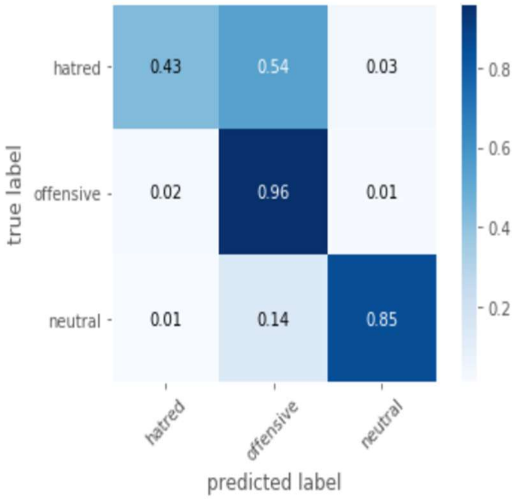

Fig5:confusion matrix

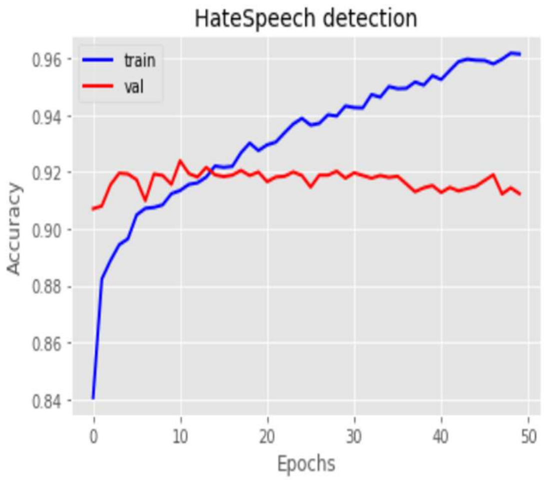

Fig6:Hatespeech detection

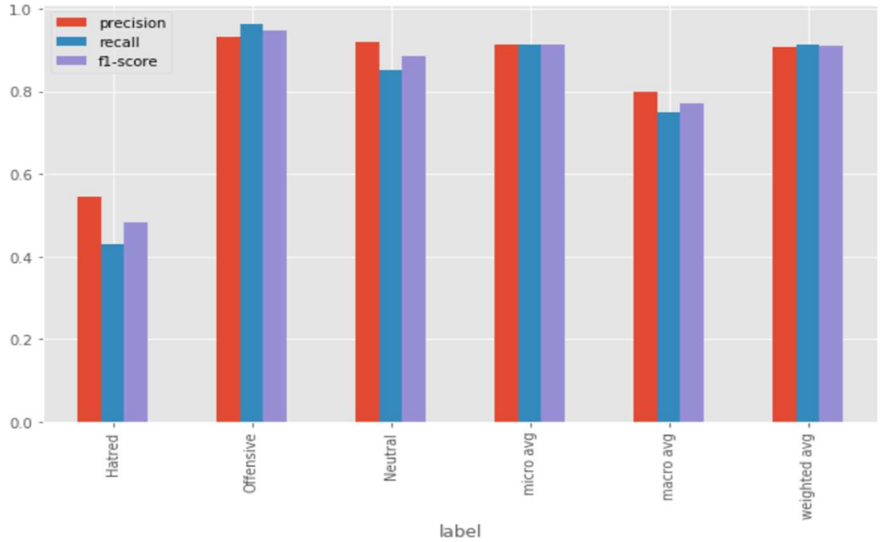

Fig7:Label wise scores representation

In figure 5 we can see f1-score, recall, precision scores for hatred, offensive, neutral, micro avg, macro avg and weighted avg.

Limitations and future expectations:

We got great help from our thesis supervisor and also tried to get help from various thesis papers and projects. But we always feel the absence of papers on our exact topics. We also faced problems for finding our datasets.

We used a dataset of 25000 data. In the future we will try to make our own dataset and also improve existing datasets. We will also try to develop our own model for feature extraction instead of using existing word embedding models.

References

[1] Zephoria.com, 2018. [Online]. Available: https://zephoria.com/top-15- valuable-facebook-statistics/. [Accessed: 22- Jun- 2018].

[2] “Twitter Usage Statistics - Internet Live Stats”, Internetlivestats.com, 2018. [Online]. Available: http://www.internetlivestats.com/twitterstatistics/. [Accessed: 22- Jun- 2018]

[3] Shanita Biere. 2018. Hate Speech Detection Using Natural Language Processing Techniques.

[4] Arjun Roy, Prashant Kapil, Kingshuk Basak, Asif Ekbal. 2017. An Ensemble approach for Aggression Identification in English and Hindi Text

[5] Steven Zimmerman,Chris Fox,Udo Kruschwitz. 2018. Improving Hate Speech Detection with Deep Learning Ensembles.

[6] Rodmonga Potapova and Denis Gordeev. 2016. Detecting state of aggression in sentences using cnn. arXiv preprint arXiv:1604.06650.

[7] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In Proceedings of the 25th international conference on world wide web, pages 145–153. International World Wide Web Conferences Steering Committee.

[8] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. arXiv preprint arXiv:1703.04009.