Hate-Speech Detection From Social Media Posts Using CNN

Ashraf Bin Shahadat, MD Mizanur Rahman Rony
Eialid Ahmed Joy,MD Adnanul Anwar
Supervisor: Dr. Md. Golam Rabiul Alam

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. With the increasing amount of hate speech online, methods that automatically detects hate speech is very much required. 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. 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%.

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[1] 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. . In [2] 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 [3] 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

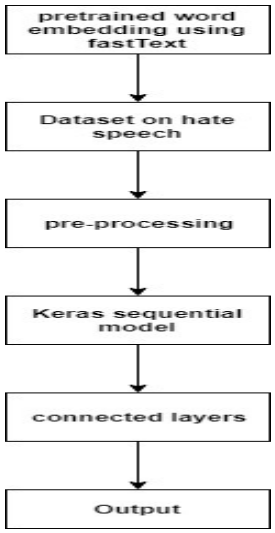

Figure 2: Work flow

System implementation

We have used twitter data from a dataset which contains 25000 of labeled twitter data. We took 17433 data as training data and 7350 data as testing data. . Below is a screenshot of the cleaned .csv data of our dataset. we used keras sequential model to train our CNN model through 9 layers. We added embedding_1,1Dconvulation,1DmaxPolling,1DGlobal MaxPolling etc layers and we made a dropout of 0.5.We used relu activation function as this is a multiclass model.

	index	count	hate_speech	offensive_language	neither	class	tweet
0	0	3	0	0	3	2	!!! RT @mayasolovely: As a woman you shouldn't...
1	1	3	0	3	0	1	!!!!! RT @mleew17: boy dats cold...tyga dwn ba...
2	2	3	0	3	0	1	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...
3	3	3	0	2	1	1	!!!!!!!!! RT @C_G_Anderson: @viva_based she lo...
4	4	6	0	6	0	1	!!!!!!!!!!!!! RT @ShenikaRoberts: The shit you...

Figure 1:Dataset of Hatespeech_detection

Data

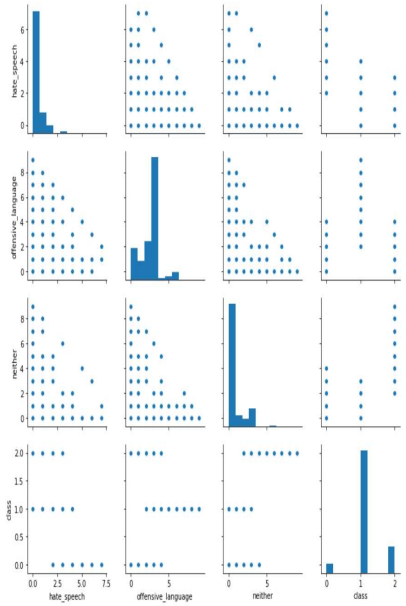

Figure 3: Pair plot of dataset

Result

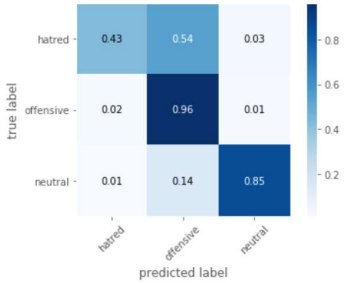

Figure 4: Confusion matrix

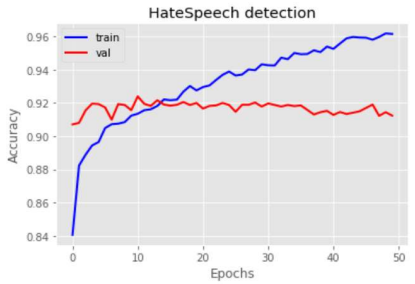

Figure 5: Accuracy vs epochs

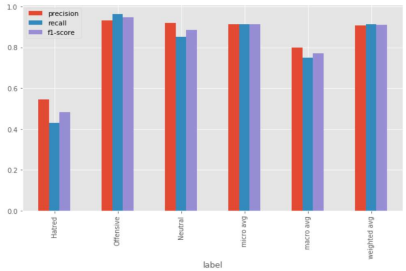

Figure 6: Label wise scores representation

Future working plan

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

[2] 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.

[3] 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.