Hate speech detection using RNN from social media posts

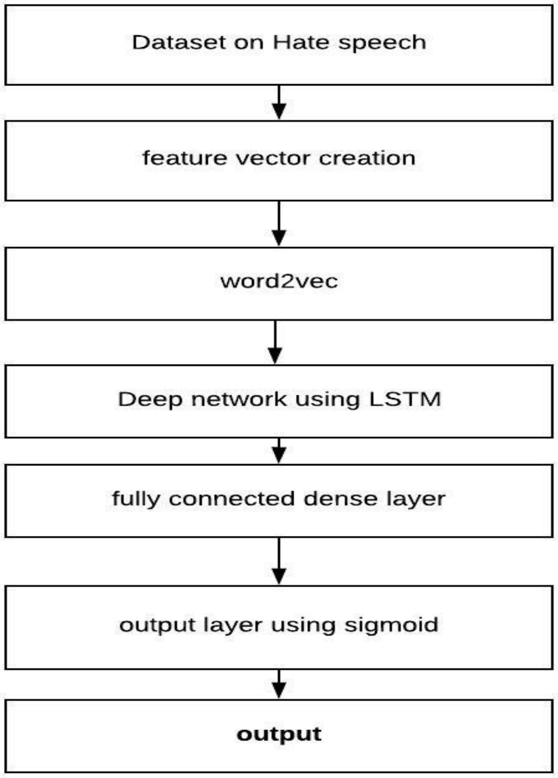
Abstract

As the increasing number of social media users 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 favor 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. The purpose of this paper is to use Natural Language processing to detect hate speech. We propose a Recurrent 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. We will evaluate our methods on the largest collection of hate speech dataset from twitter. 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.

Literature Review

From [1], we have found that they mainly used convolutional neural network (CNN), bag of words and word2vec for feature vectorization. They preferred CNN over recurrent neural networks. But we prefer RNN over CNN because CNN is very time consuming. Moreover, 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. One more thing is that CNN cannot handle sequential data while RNN can. In [2], 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. But we prefer word2vec for our model as it is more reliable word embedding technique than glove. It learns words from a large corpus more quickly than glove.

Working diagram



Dataset on Hate speech

feature vector creation

word2vec

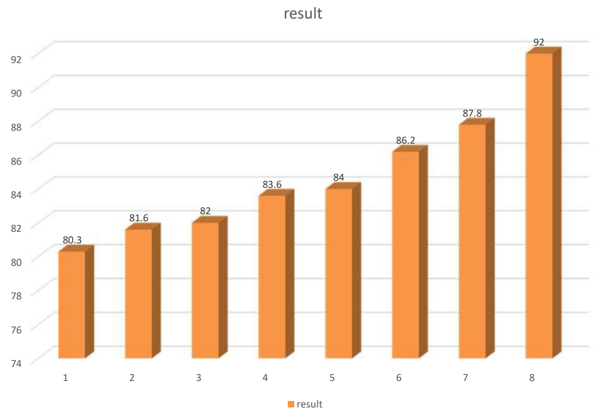
Deep network using LSTM

fully connected dense layer

output layer using sigmoid

output

Expected result



Reference

[1] Nockleby, John T. (2000), "Hate Speech" in Encyclopedia of the American Constitution, ed. Leonard W. Levy and Kenneth L. Karst, vol. 3. (2nd ed.), Detroit: Macmillan Reference US, pp. 1277–79. Cited in "Library 2.0 and the Problem of Hate Speech," by Margaret Brown-Sica and Jeffrey Beall, Electronic Journal of Academic and Special Librarianship, vol. 9 no. 2 (Summer 2008).

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

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

Supervisor

Dr. Md. Golam Rabiul Alam

Associate Professor, CSE, BRAC University

Authors:

Ashraf Bin Shahadat(16101199)

MD Mizanur Rahman Rony(16101184)

Eialid Ahmed Joy(16101182)

Adnanul Anwar (16101005)