Deep Learning for Hate Speech Detection in Tweets Using LSTM and Bi-LSTM

Ashif Mondal [*] [#1], Diya Neogi [*] [#2], Arpon Roy [*] [#3], Bidisha Saha [*] [#4]
[*] *Scholar, RCC Institute of Information Technology, Kolkata*
[#1] ashifmondal692@gmail.com, [#2] diya.neogi82@gmail.com, [#3] arponkolkata7@gmail.com, [#4] Sbidisha688@gmail.com

Mrs. Monika Singh [**] [#5]
[**] *Assistant Professor, RCC Institute of Information Technology, Kolkata*
[#5] singhmonika8@gmail.com

Abstract- With the massive availability of social networks to everyone and with the increase in social interactions on online social networks, we can observe a massive number of hateful activities in social networks. This scenario has influenced the researchers to deal with this problem programmatically as manual detection of such activities is not scalable. In this project we'll classify the tweets as racist, sexist or neither. This project is quite difficult because the tweets can have different form of hatred, different target but representing the same meaning. In such kind of topics deep learning methods has the most accuracy over large number of complex problems. In this project we'll be using SVM, Deep Neural Network (DNN). We'll be using deep learning architectures such as fast text, Convolutional Neural Network (CNN), Long Short-term Memory Network (LSTM). And we'll explore various tweet semantic embeddings such as char n grams, word Term Frequency Inverse Document Frequency (TF-IDF), Bag of Words Vectors (BoWV) over Global Vector (GloVe) for word representation.

Keywords- Hate Speech Detection; hate speech; deep learning; Deep Neural Network (DNN); Convolutional Neural Network (CNN); Recurrent Neural Network (RNN), FastText; Bi-directional Long Short-Term Memory (BiLSTM); Long Short-Term Memory (LSTM); text classification.

I. INTRODUCTION

Hate speech is one of the serious issues we see on social media platforms like Facebook and Twitter, mostly from people with political views. Now a days the usage of hate speech in social media is increasing massively. On Twitter hate speech are those which contains violence or abuse speech towards a person or group based on religion, gender, organization, country etc. Social media platforms such as twitter need to detect hate speech and prevent it from going viral or ban it at the right time. Now a days deep Learning based methods are used for such hate speech detection on Twitter, this task is critical for applications like controversial event extraction, building AI chatterbots, content recommendation, and sentiment analysis. In our work we would like to develop an AI chatterbox which will help us differentiate between messages that express explicit hate and messages that are just offensive. [10]

Detecting hate speech is a challenging task, however. First, there are disagreements in how hate speech should be defined. This means that some content can be considered hate speech to some and not to others, based on their respective definitions. We start by covering competing definitions, focusing on the different aspects that contribute to hate speech. We are by no means, nor can we be, comprehensive as new definitions appear regularly. Our aim is simply to illustrate variances highlighting difficulties that arise from such.

II. LITERATURE SURVEY

After going through some research paper on this topic we got basic ideas of the word classifiers such as Char n grams, TF-IDF, BoWV and the deep learning algorithm such as Convolutional Neural Network (CNN), Gated Recurrent Unit (GRU), FastText, Long Short-Term Memory Network (LSTM), Bi-Directional Long Short-Term Memory Networks (Bi-LSTM) etc.

A. Word classifiers:

A text classifier labels unstructured texts into predefined text categories. Instead of users having to review and analyze vast amounts of information to understand the context, text classification helps derive relevant insight.

The word classifiers used here are char n grams, TF-IDF, BoWV. [6] [7] [8] [9]

- Char n grams: Compared to word n-grams, which only capture the identity of a word and its possible neighbours, character n-grams are additionally capable of detecting the morphological makeup of a word.
- TF-IDF: It transforms text into a meaningful representation of numbers which is used to fit machine algorithm for prediction.
- BoWV: It uses the average of the word (GloVe) embeddings to represent a sentence. We experiment with multiple classifiers for both the TF-IDF and the BoWV approaches.

B. Deep Neural Network (DNN) methods:

Neural Networks replicate the way humans learn, inspired by how the neurons in our brains fire, only much simpler. The most common Neural Networks consist of three network layers: (i) An input layer, (ii) A hidden layer (this is the most important layer where feature extraction takes place, and adjustments are made to train faster and function better), and (iii) An output layer. Each sheet contains neurons called "nodes," performing various operations.

The Neural Networks we used here is CNN, GRU, FastText, LSTM, Bi-LSTM. [6] [7] [8] [9]

- CNN: A convolutional neural network is a feed-forward neural network that is generally used to analyze text by processing data with grid-like topology. Convolutional Neural Network provides accurate result on text classification such as sentiment analysis. A convolution neural network has multiple hidden layers that help in extracting information from a text. [11]
- FastText: It is also a feed-forward neural network but it uses backpropagation to update the word vectors. Backpropagation sends error information from the network's last layer to all of the weights within the network. Basically, it fine tunes the weights of a neural network based on the previous epoch's error rate. It improves the model by reducing the error. [13]
- GRU: Gated Recurrent Unit is a recurrent neural network. It has a memory unit and two gates such as update gate and reset gate. Update gate determines how much of the past knowledge needs to be passed along into the future. Reset gate determines how much of the past knowledge to forget. [14]
- LSTM: Like recurrent neural network, it uses internal memory to process the sequence of inputs. Recurrent Neural Networks were created to solve the sequential input data time-series problem. RNN has an internal memory that stores the information from previous samples' computations. The goal of AI in this case is to create a system that can understand human-spoken natural languages, such as natural language modeling, word embedding, and machine translation. LSTM has forget gate, input gate and output gate. Input gate determines the extent of information be written onto the Internal Cell State. Output gate determines what output to generate from the current Internal Cell State. And LSTM can save the data in its memory block until forget gate triggers. [12]
- Bi-LSTM: Unlike standard LSTM, the input flows in both directions, and it's capable of utilizing information from both sides which makes a Bi-LSTM different from the regular LSTM. The main reason is that every component of an input sequence has information from both the past and present. For this reason, Bi-LSTM can produce a more meaningful output, combining LSTM layers from both directions. [15]

By Ji Ho et al's paper over a dataset by Waseem and Hovy dataset, 2016 [3] Hybrid CNN Classifier (wordCNN + charCNN) gives F1 score of 0.827.

On the paper by Zhang et al, 2018 [5] we can see over 24k dataset CNN + GRU gives 0.919 F1 score.

By Ziqi Zhang, David Robinson, Jonathan Tepper [4] CNN + LSTM, emb-ggl2 gives 0.919 F1 score on 24k dataset.

According to the research paper "Deep Learning for Hate Speech Detection in Tweets" published on 2020 by Pinkesh Badjatiya, Shashank Gupta, Manish Gupta and Vasudeva Varma [1] we came to know that over a dataset of 16k tweets CNN + Random Embedding + GBDT gives F1 score of 86.4%, FastText + Random Embedding + GBDT gives F1 score of 88.6%, LSTM + Random Embedding + GBDT gives F1 score of 93%.

On the basis of another research paper "Detecting Hate Speech using Deep Learning Techniques" published on 2021 by Chayan Paul and Pronami Bora [2] we saw that on a dataset of 16k data LSTM gives F1 score of 97.85% where Bi-LSTM gives F1 score of 97.81%.

Reference	Dataset Size	Method	Result			
			Accuracy	Precision	Recall	F1
Ji Ho et al, Waseem and Hovy dataset, 2016	25k	Hybrid CNN Classifier (wordCNN + charCNN)	0.827	0.827	0.827	0.827
Zhang et al, 2018	24k	CNN + GRU	0.94	0.94	0.94	0.94
Ziqi Zhang, David Robinson, Jonathan Tepper	24k	CNN + LSTM, emb-ggl2	0.919	0.919	0.919	0.919
"Deep Learning for Hate Speech Detection in Tweets" published on 2020 by Pinkesh Badjatiya, Shashank Gupta, Manish Gupta and Vasudeva Varma	16k	CNN + Random Embedding + GBDT	0.864	0.864	0.864	0.864
		FastText + Random Embedding + GBDT	0.886	0.886	0.887	0.886
		LSTM + Random Embedding + GBDT	0.930	0.930	0.930	0.930
"Detecting Hate Speech using Deep Learning Techniques" published on 2021 by Chayan Paul and Pronami Bora	16k	LSTM	0.9785	0.9598	0.9986	0.9785
		Bi-LSTM	0.9781	0.9582	0.9990	0.9781

III. FINDINGS

From those papers we came to know convolutional neural networks (CNN), long short-term memory (LSTM) and bi-directional LSTM are three of the most popular deep neural network designs used for hate speech detection using deep learning models. There have been many word embeddings methods introduced over the years, such as word2vec, Glove, and FastText. This technique uses a combination of different models, such as LSTM, Bi-LSTM, and CNN. After reading those papers we found that LSTM and Bi-LSTM gives better performance among CNN, GRU, FastText, LSTM, Bi-LSTM. Mostly from last two paper we found LSTM and Bi-LSTM are the two latest methods and it shows superior performance across these datasets.

IV. TECHNOLOGIES AND METHODOLOGIES

In this section we discuss about the classifiers that we used to tokenize our dataset and the deep neural network methods for training the model. Here we proposed Char n grams, TF-IDF and BoWV. We experiment with multiple classifiers for both the TF-IDF and the BoWV approaches. The Neural Networks we used here is CNN, FastText, LSTM, Bi-LSTM.

CNN doesn't use backpropagation while FastText uses. FastText sends error information from the network's last layer to all of the weights within the network and fine tunes the weight of the current neural network based on the previous epochs error rate. That's the reason why FastText performs better than CNN.

On the other side LSTM has a memory block on its architecture. LSTM has forget gate, input gate and output gate. Input gate determines the extent of information be written onto the Internal Cell State. Output gate determines what output to generate from the current Internal Cell State. And LSTM can save the data in its memory block until forget gate triggers. For this features LSTM can memorize both short and long-time memory while FastText can only remember short time memory. So, LSTM is better than FastText.

In case of Bi-LSTM, it works in both directions, so it's capable to utilize information from both sides. So, it can predict a word on the basis of previous data like LSTM and also predict the previous word based on the next data. For this reason, Bi-LSTM is more superior than LSTM.

So, LSTM and Bi-LSTM is better methods than CNN and FastText. In this project we're using LSTM and Bi-LSTM.

Our dataset has 24,783 annotated tweets. This dataset was originally collected from Twitter and contains the following columns:

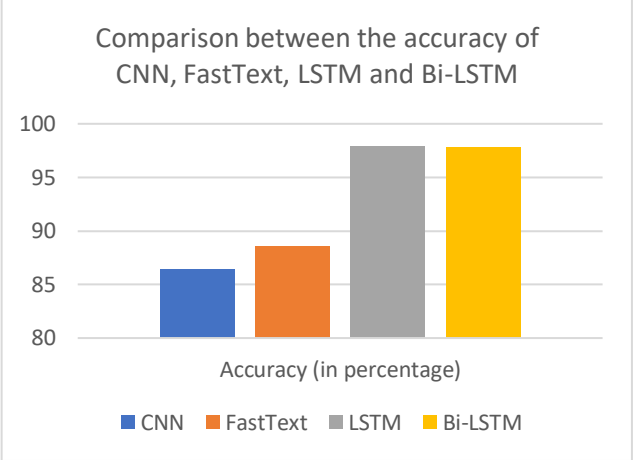
Index, count, hate_speech, offensive_language, neither, class and tweet

A	B	C	D	E	F	G
	count	hate_speech	offensive_language	neither	class	tweet

We classified the dataset into three categories:

- 0 - hate speech
- 1 - offensive language
- 2 - neither as positive or negative

After the network is learned completely, a new random tweet is tested against the network and it'll classify the tweet as racist, sexist or neither.



V. RESEARCH GAP

In twitter there are various kind of tweets like text, emojis, pictures, GIF, stickers, reacts etc. Even there are so many languages other than English. In this project we managed to work with the text which are tweeted in English language but there is not such algorithm to work with multilingual languages such as Hindi, Bengali, German, Spanish, Arabic, Chinese or Japanese which can translate those languages to English without changing the proper meaning. We've not used and algorithm for the emoji, picture, GIF, stickers which can properly identify the emoji and convert into text such as happiness, sorrow, ignore etc or analyse the meaning of picture, GIF or stickers if it is used in good motives or making fun of a person or some community. Another case is that if someone uses hate word as a meme or joke, the model cannot identify if it's a joke or hate word.

Method	Accuracy	Precision	Recall	F1	Missing
CNN	0.864	0.864	0.864	0.864	Multi-lingual comments other than English, emoticons, pictures, GIF, stickers, reacts of tweet etc.
FastText	0.886	0.886	0.887	0.886	
LSTM	0.9785	0.9598	0.9986	0.9785	
Bi-LSTM	0.9781	0.9582	0.9990	0.9781	

VI. CONCLUSION

We introduced a method for automatically classifying hate speech on twitter using deep neural network (DNN) such as CNN, FastText, LSTM, Bi-LSTM and we investigated the application of deep neural network architectures for the task of hate speech detection. With the massive availability of

social networks to everyone and with the increase in social interactions on online social networks, we can observe a massive number of hateful activities in social networks. This scenario has influenced the researchers to deal with this problem programmatically as manual detection of such activities is not scalable.

REFERENCES

[1] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta and Vasudeva Varma, “Deep Learning for Hate Speech Detection in Tweets”, 2020.

[2] Chayan Paul and Pronami Bora, “Detecting Hate Speech using Deep Learning Techniques”, 2021.

[3] Ji Ho et al, Waseem and Hovy, 2016.

[4] Ziqi Zhang, David Robinson, Jonathan Tepper, “Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter”, 2018.

[5] Zhang et al, 2018

[6] C. Paul, D. Sahoo and P. Bora, "Aggression in Social Media: Detection Using Machine Learning Algorithms," International Journal of Scientific and Technology Research, vol. 9, no. 4, pp. 114-117, 2020.

[7] P. Jadhav and B. V. Babu, "Detection of Community within Social Networks with Diverse Features of Network Analysis," Journal of Advanced Research in Dynamical and Control Systems, vol. 11, no. 12, pp. 366-371, 2019.

[8] S. P. Bhargav, G. N. Reddy, R. R. Chand, K. Pujitha and A. Mathur, "Sentiment Analysis for Hotel Rating using Machine Learning Algorithms," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 8, no. 6, pp. 1225-1228, 2019.

[9] L. P. Maguluri, I. Bhavitha, S. A. v. Reddy, T. N. Reddy and A. Chowdary, "An efficient method on supervised joint topic modeling approach by analyzing sentiments," Journal of Advanced Research in Dynamical and Control Systems, vol. 9, no. 18, pp. 3219-3230, 2017.

[10] Deep Learning for Hate Speech Detection: A Large-scale Empirical Evaluation | by Guansong Pang | Towards Data Science

[11] A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way | by Sumit Saha | Towards Data Science

[12] An Overview on Long Short Term Memory (LSTM) - Analytics Vidhya

[13] A hands-on intuitive approach to Deep Learning Methods for Text Data — Word2Vec, GloVe and FastText | by Dipanjan (DJ) Sarkar | Towards Data Science

[14] Understanding GRU Networks. In this article, I will try to give a… | by Simeon Kostadinov | Towards Data Science

[15] Complete Guide To Bidirectional LSTM (With Python Codes) (analyticsindiamag.com)