Detecting Hate Speech using Deep Learning Techniques

Chayan Paul[1]
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, AP, India

Pronami Bora[2]
Department of Electronics and Communication Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, AP, India

***Abstract*—Social networking sites saw a steep rise in terms of number of users in last few years. As a result of this, the interaction among the users also increased considerably. Along with these posting racial comments based on cast, race, gender, religion, etc. also increased. This propagation of negative messages is collectively known as hate speeches. Often these posts containing negative comments in social networking sites create law and order situations in the society, leading to loss of human life and properties. Detecting hate speech is one of the major challenges faced in recent time. In recent past, there have been a considerable amount of research going on the field of detection of hate speech in the social networking sites. Researchers in the fields of Natural Language Processing and Machine Learning have done considerable amount research in in this area. This paper uses a simple up sampling method to make the data balanced and implements deep learning models like Long Short Term Memory (LSTM) and Bi-directional Long Short Term Memory (Bi-LSTM) for improved accuracy in detecting hate speech in social networking sites. LSTM was found to have better accuracy that Bi-LSTM for the data set considered. LSTM also had better values for precision and F1 score. Bi-LSTM only for higher values for recall.**

***Keywords*—Bi-directional Long Short Term Memory (Bi-LSTM); deep learning; hate speech; Long Short Term Memory (LSTM); text classification**

I. INTRODUCTION

Social Networking Sites (SNS) have provided us with easy ways to connect with various people or organization of our interest. Because of the evolution of various technologies like highspeed internet and handheld devices, these sites have reached to the large number of people in the society. Largest chunk of the users in these networks are young. Researchers have grabbed the large collection of data found in various social networking sites and conducted a considerable amount of research in different areas. Sentiment Analysis is one of the leading areas of research which involves a lot of data from social networks. There are a good number of researches done to find out the sentiment related to a specific product or service using data from social networking sites like Twitter [1] [2] [3] [4]. Apart from sentiment analysis there are other subsets of research done using the data from social networking sites; like detecting users with similar interest in a specific product or service [5] [6]; detection of abusive languages in social media [7] [8]. A good number of research works also have been done to improve the methodologies to analyse the data collected from social networking sites [9] [10].

One thing that these networks make possible now a days is direct interaction with various celebrities. An individual can directly interact with a celebrity and share their views. Similarly, various political parties and business houses utilise these networks for reaching out to their target audience. The problem arises when the users' opinion does not match for an issue. These issues can range from political affiliation to religious belief, opinions related to gender, cast and so on. These mismatch in opinion results in exchange of hate full contents in social networking sites. In fact, hate speech and abusive contents have become a current trend in social media sites and these often results to disturbance in the society. There are reports of riots breaking out in different cities where the main source of the spread of riots are found to be social media posts [11], [12]. Intuitively detection of hate speech in social networks become important.

Hate speech can be characterized as exchange of verbal or nonverbal information among the users with intolerance and aggression [13]. Hate speech can be in different forms, like interaction between users on social network which may contain unparliamentary languages. It could also be abusing a person or a certain group of people for their religious belief, their sexual orientation, their race, their political affiliation [14]. Often these exchange of abusive language lowers the self-esteem of the people and may lead to negative impact in the society [15]. Spread of hate speech has become a global phenomenon.

In this paper endeavors to build a deep learning model for classification of social media contents to either hateful or normal. Twitter was chosen as a platform where detection of hate speech was done. Open source dataset available publicly, was collected to train the models. This paper predominantly builds a Long Short Term Memory and a Bi Directional Long Short term Memory using the dataset.

This section of the paper is followed by a related works section, where the existing works in the related areas are discussed. The next section is methodology, where a discussion is presented on the different methodologies used in this paper. Next to methodology section, results obtained in this paper are discussed. The result section also has introductory discussion on different measures used in this paper for presenting the results. After results section, conclusion section presents the concluding remarks.

II. Related Works

The problem of detecting hate speech has been addressed by various researchers in different ways. In general, the problem can be addressed in different ways. One of the possible ways is to develop a pure Natural Language Processing model, which is generally an unsupervised model. So, the detection becomes comparatively easier as there is no need for a labelled data set. In this approach an NLP model can be designed which categorizes whether a sentence contains hate speech or not [16], [17]. In literature there are fewer works which were carried out totally based on pure NLP based concepts. One of the probable reasons is the models are comparatively slower than the models built using Machine Learning or Deep Learning Models.

The machine learning and deep learning models for detection of hate speech needs labelled data set which is used to train the model. A good number of researches has been carried out in this area where the researchers created their own dataset. The general procedure is to collect the data from a social networking site clean the data and then get them annotated by a team of experts who manually annotate if a text contains hateful message or not. Khan et al., conducted a comprehensive survey of machine learning models used extensively in NLP [18]. Ahmed et al. developed a dataset which consists of English and Bengali mixed texts and annotated the tweets as hate speech or non-hate speech [19]. Sahi et al. developed a supervised learning model to detect hate speech against women in Turkish language. They collected tweets mentioning clothing choices of women and used this data to train the machine learning models [20]. Waseem examined the influence of annotators' knowledge on classification model [21] Waseem et al. provided with a data set of 16,000 tweets and they also investigated which features provides the best performance when it comes to classification of hate speeches [22]. Also, there are a good number of works done where researchers take an open source data and try to develop models which are used to detect the hateful message in social networking sites [23] [24] [25].

The research works in some cases went beyond the binary classification of a message into hate speech and non-hate speech and make it multi class classification. Watanabe et al. conducted a study where they used twitter data to create a model which can classify tweets in three classes i.e., clean, offensive and hateful [26]. Kumar et al. developed a model using taking text messages from Facebook which could classify the messages into three different classes i.e., Aggressive, Covertly Aggressive, and Non-aggressive texts [27].

In this paper we collected a data set from Kaggle which contains tweets from American users. We built a deep learning model to classify the tweets into two categories, hate-speech and neutral.

III. Methodology

In this paper we proposed to classify the tweets using a Long Short Term Memory (LSTM) and a Bi Directional Long Short Term Memory (Bi-LSTM). Both LSTM and Bi-LSTM are versions of neural networks, with persistent memories [28].

A. Long Short-Term Memories (LSTM)

These are special types of neural networks which are designed to work well when one has sequence data set and there exists a long term dependency. These networks can be useful when one needs a network to remember information for a longer period. This feature makes LSTM suitable for processing textual data. Fig. 1 shows a typical architecture of an LSTM. As it can be seen in the diagram, an LSTM is a collection of similar cells, whereas each cell processes the input in a specific approach. Apart from the input from external sources, each cell also receives inputs from its earlier cell in the chain. This arrangement of cells, facilitates LSTM to remember earlier information for a longer time.

B. Bi-Directional Long Short-Term Memories (Bi-LSTM)

Normal form of LSTMs can remember or refer to the information which it has traversed till now. But it does not have any evidence about the information present after the point traversed till the point. This becomes a considerable drawback while dealing with sequence data, especially text. Bi-directional LSTM is another version of LSTM which can remember the information from both directions. In Bi-directional LSTM we basically do backpropagation in two ways. Once from the front and once from the back. This process makes Bi-LSTTM a powerful tool for analysing textual data.

C. Data Pre-Processing

We collected a dataset from Kaggle, an open source platform. The labelled data set contained two classes namely hate speech and non-hate speech. Hate speech is denoted as 1 and non-hate speech is denoted by 0. We removed the special symbols from the texts. Then we converted the texts in lower case. We also used stemming to convert the words into their basic words. We checked the dataset for number of data for hate speech and non-hate speech. We found the data set to be highly imbalanced. Fig. 2 represents the bar diagram for two classes. Table I also represents the number of tweets available in both the classes.

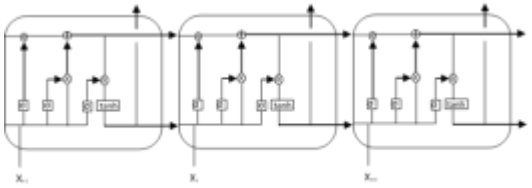

Fig. 1. Architecture of LSTM.

TABLE I. Number of Tweets in Classes

Class name	Number of tweets
Hate-speech (represented by 1)	2242
Non hate-speech (represented by 0)	29720

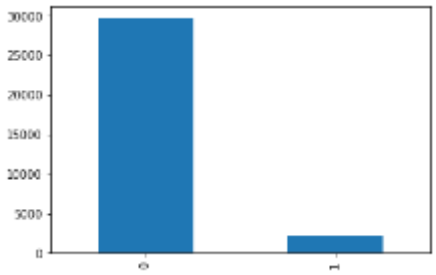

Fig. 2. Bar Diagram Representing Imbalanced Class.

With this state of the data set, if we apply classification algorithms, there is high chance of getting biased results. In this scenario, down sampling can be done to make the majority class equivalent to the minority class. But in this approach, we have risk of losing a large chunk of data which may affect the classification result. Finally, we went for up sampling the minority class, by randomly selecting from the class and adding them back to the data set. This approach provided us with a balanced data set, but the total number of tweets got increased drastically. Fig. 3 represents the balanced data set.

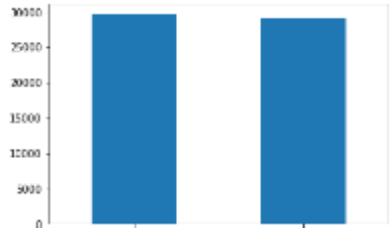

Fig. 3. Bar Diagram Representing Balanced Class.

We divided the data set into training and testing. We kept 67:33 ratio for training and testing. With the training data set we trained an LSTM and a Bi-LSTM. We applied one hot encoding to get the data ready for the algorithms. One hot encoding is a process which converts the text data into numerical data. Each of the words gets a unique numerical representation in one hot encoding. Then we applied padding. Padding is a process which adds zeros to either beginning or ending of sentences for making all the sentences of same length. Then we applied word embedding. Embedding is a process represents each of the words in a higher dimensional space. It is helpful in finding similarity and dissimilarity between the words effectively.

IV. Result

We first computed the confusion matrix for both the models. A confusion matrix presents four different values, namely true positive, true negative, false positive and false negative. True positive means the number of classes which were originally positive, and the model also classified them as positive. True negative means the classes were originally negative and the model also classified them as negative. False positive values are the number of classes which were originally negative, but predicted as positive by the model, and false negative means the classes were originally positive, but predicted negative by the models. Fig. 4 represents the idea of a confusion matrix.

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (FP)

Fig. 4. Confusion Matrix.

Fig. 5 and 6 present the confusion matrices for LSTM and Bi-LSTM respectively. In these representations, we presented the values in percentage instead of actual number of classes.

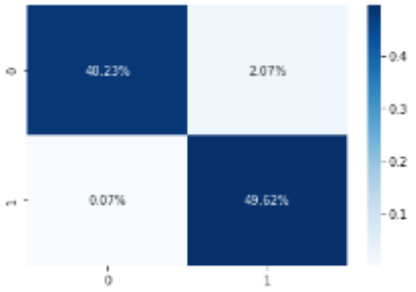

Fig. 5. Confusion Matrix for LSTM.

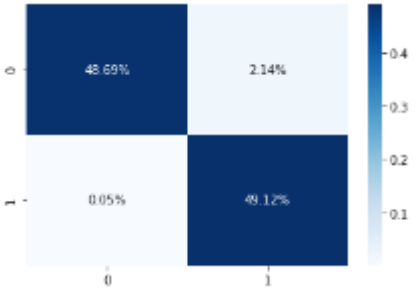

Fig. 6. Confusion Matrix for Bi-LSTM.

From the confusion matrices, we can see there is no considerable difference between the performances of these two models. LSTM has a bit higher false positive in comparison to Bi-LSTM, whereas Bi-LSTM has higher false positive. But it is evident that the differences between the values are very small. We also calculated the other performance measure values accuracy, precision, recall and F1 score. Below we discuss the values in very brief:

A. Accuracy

Accuracy is one of the most widely used performance measures and it is the ratio of total number of entries classified accurately to the total number of observations. For a balanced dataset Accuracy is the measure using which we can compare the performance of an algorithm. In this study, we got a slightly higher accuracy for LSTM, though the difference is very less.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

B. Precision

Precision is the ratio of entries that are correctly predicted positive to total positive entries. A higher value for precision means low false positive rates. As per the calculations in this study LSTM got slightly higher precision than Bi-LSTM.

$$precission = \frac{TP}{TP + FP}$$

C. Recall

Recall is the ratio of number of positive entries which were predicted correctly to total number of entries in the positive class. It basically reflects the proportion of positive observation which were correctly classified. In this study we can see that Bi-LSTM has better recall in comparison to LSTM.

$$recall = \frac{TP}{TP + FN}$$

D. F1 Score

F1 score is the weighted average of precision and recall, as a result it considers both false negative and false positive. For a problem where the classes are imbalanced, F1 score becomes better performance measure than accuracy. In this study we found the f1 score of LSTM also slightly higher than that of Bi-LSTM.

$$f1\ score = \frac{2 * recall * precission}{recall + precision}$$

We calculated the values for accuracy, precision, recall and F1 score for both the models. The calculated values are presented in Table II.

TABLE II. PERFORMANCE MEASURE SCORES FOR LSTM AND BI-LSTM

Model	Accuracy	Precision	Recall	F1 Score
LSTM	0.9785	0.9598	0.9986	0.9785
Bi-LSTM	0.9781	0.9582	0.9990	0.9781

V. CONCLUSION

The scores calculated for accuracy, precision, and f1 score suggest that LSTM has performed better than Bi-LSTM. But recall score is found to be better for Bi-LSTM than LSTM. Recall basically signifies the ratio of positive classification to total positive classification. Here in this study we considered hate speech as positive class. That means the model has less error in detecting the hate speech. In this context, Bi-LSTM has a slight edge over LSTM. Although, the difference between the scores are really very small to draw any comparison between the two models.

This study can be further extended for real world data set collected from twitter with context to some real events. It will be interesting to see how these models perform on new data set. Attention model is one area which has a good application in NLP, we plan to apply this model in our future works.

REFERENCES

[1] S. Muthukumaran, P. Suresh and J. Amudhavel, "Sentimental analysis on online product reviews using LS-SVM method," Journal of Advanced Research in Dynamical and Control Systems, vol. 9, no. 12, pp. 1342-1352, 2017.

[2] S. A. Devi, P. Sapkota and M. Obulesh, "Sentiment analysis on products using social media," Journal of Advanced Research in Dynamical and Control Systems, pp. 137-141, 2017.

[3] M. Bhargava and D. Rao , "Sentimental analysis on social media data using R programming," International Journal of Engineering and Technology(UAE), vol. 7, no. 2, pp. 80-84, 2018.

[4] C. G. Krishna , D. R. Meka, V. S. Vamsi and K. M. V. S. Ravi , "A survey on twitter sentimental analysis with machine learning techniques," International Journal of Engineering and Technology(UAE), vol. 7, no. 2.32, pp. 462-465, 2018.

[5] 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.

[6] 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.

[7] B. R. Rahin, K. K. Prem, N. Danapaquiameq, J. Arumugam and D. Saravanan, , "Blocking Abusive and Analysis of Tweets in Twitter Social Network Using NLP in Real-Time," Bioscience Biotechnology Research Communications, vol. 11, no. 1, pp. 94-103, 2018.

[8] 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.

[9] L. A. Deshpande, and M. R. Narasingarao, "ADDRESSING SOCIAL Popularity in Twitter Data using Drift Detection Technique," Journal of Engineering Science and Technology, vol. 14, no. 2, pp. 922-934, 2019.

[10] 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.

[11] H. Correspondent, "Facebook played a role in fuelling riots, says Delhi panel," 01 September 2020. [Online]. Available: https://www.hindustantimes.com/cities/facebook-complicit-in-aggravating-n-e-delhi-riots-says-delhi-assembly-panel/story-1HkXrGw4fWSOpoLUrVuCsO.html. [Accessed 03 September 2020].

[12] K. R. Balasubramanyam, "Bengaluru Riots: Karnataka to hold talks with social media giants on filtering fiery contents," 17 August 2020. [Online]. Available: https://economictimes.indiatimes.com/news/politics-and-nation/bengaluru-riots-karnataka-to-hold-talks-with-social-media-giants-on-filtering-fiery-contents/articleshow/77582323.cms. [Accessed 03 September 2020].

[13] K. Sreelakshmi , B. Premjith and K. P. Soman , "Detection of Hate Speech Text in Hindi-English Code-mixed Data," in Procedia Computer Science, Trivandrum, 2020.

[14] J. C. Pereira-Kohatsu, L. Quijano-Sánchez , F. Liberatore and M. Camacho-Collados , "Detecting and Monitoring Hate Speech in Twitter," Sensors (Basel), pp. 1-37, 2019.

[15] A. Gaydhani, V. Doma, S. Kendre and L. Bhagwat, "Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach," rXiv preprint arXiv:1809.08651., pp. 1-5, 2018.

[16] G. B. Herwanto, A. M. Ningtyas, K. E. Nugraha and P. T. I Nyoman , "Hate Speech and Abusive Language Classification using fastText," in International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Jetis, Indonesia, 2019.

[17] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in Proceedings of the Fifth International workshop on natural language processing for social media., Valencia, Spain, 2017.

[18] W. Khan, A. Daud, J. A. Nasir and T. Amjad, "A survey on the state-of-the-art machine learning models in the context of NLP," Kuwait Journal of Science, vol. 43, no. 4, pp. 95-113, 2016.

[19] S. Ahammed, M. Rahman, H. M. Niloy and S. M. H. Chowdhury, "Implementation of Machine Learning to Detect Hate Speech in Bangla Language," in International Conference on System Modeling & Advancement in Research Trends, Moradabad, India, 2019.

[20] H. Sahi, Y. Kilic and R. B. Saglam, "Automated Detection of Hate Speech Towards Women on Twitter," in 2018 International Conference on Computer Science and Engineering (UBMK), Turkey, 2018.

[21] Z. Waseem, "Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter," in Proceedings of 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science, , Austin, 2016.

[22] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," in Proceedings of NAACL-HLT 2016, San Diego, California, 2016.

[23] G. Koushik, K. Rajeswari and S. K. Muthusamy, "Automated Hate Speech Detection on Twitter," in 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), Pune, India, 2019.

[24] T. Davidson, D. Warmsley, M. Macy and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," in ICWSM, 2017.

[25] G. K. Pitsilis, H. Ramampiaro and H. Langseth , "Effective hate-speech detection in Twitter data using recurrent neural networks," Applied Intelligence, vol. 48, no. 12, p. 4730–4742, 2018.

[26] H. Watanabe, M. Bouazizi and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," IEEE Access, vol. 6, pp. 13825 - 13835, 2018.

[27] R. Kumar, A. K. Ojha, S. Malmasi and M. Zampieri, "Benchmarking Aggression Identification in Social Media," in Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Santa Fe, New Mexico, USA, 2018.

[28] C. Colah, "Understanding LSTM Networks," August 2015. [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs/. [Accessed 17 09 2020].