2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)
University of Malaya, Kuala Lumpur, Malaysia. Sep 24-26, 2021

2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON) | 978-1-7281-9951-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/GUCON50781.2021.9573687

Deep Learning for Hate Speech Detection in social media

Ashwini Kumar
Department of CSE
Graphic Era Deemed to be University
Dehradun, India
ashwinipaul@gmail.com
ORCID: 0000-0001-9908-7430

Vishu Tyagi
Department of CSE
Graphic Era Deemed to be University
Dehradun, India
tyagi.vishi@gmail.com
ORCID: 0000-0001-9015-9178

Sanjoy Das
Department of Computer Science
Indira Gandhi National Tribal University, RCM
Imphal, India
sdas.jnu@gmail.com
ORCID: 0000-0001-8018-0870

***Abstract*—In recent years, many people on the internet write and post abusive language on online social media platforms such as Twitter, Facebook, etc. Detection of hate speech is very difficult to solve manually, especially in social media. Thus, we need to be automatic detection of hate speech in social media. We have used a benchmark dataset of approximately 25 thousand annotated tweets and proposed a model based on deep learning methods. We also compare the performance of our deep learning methods to the traditional machine learning classifier in terms of F1 Score and Accuracy. The results obtained through the proposed method is very promising.**

***Index Terms*—Hate Speech, Twitter, Deep Learning, Social Media**

I. Introduction

In the current era, most people want to connect on multiple social media platforms such as Twitter, Facebook, and any community forums to freely express themselves and share information [8,11,12,14]. Despite the popularity of social media and its convenience, many internet users express hate speech on Twitter. The term 'Hate Speech' was commonly defined as any expression that disparages a person or group based on some characteristics such as nationality, gender, sexual orientation, religion or other characteristics [7, 9, 10]. Nowadays, Hate or aggressive Speech is increasingly being used on Twitter, and many international organizations and countries have been initiated to develop countermeasures of this matter's problems. Many researchers are still working on classical machine learning. But for a quick and good performance, we must use deep learning techniques to detect automatic hate speech with the help of deep neural networks. The issue of text classification is crucial in NLP, but we can easily solve this problem with different deep learning methods. In this paper, we have proposed a method using a deep learning approach to detect Hate Speech of Twitter Dataset having approx. 25k tweets. The accuracy of our proposed model is 92.5 for classifying tweets as hate or non-hate on a benchmark dataset. Our new method uses a Recurrent Neural Networks based Model with features to attain better results and outperformed classical machine learning methods. The rest of this paper is structured as follows: related work discussed in Section II; Methodology and Proposed Method described in Section III; Data set, Experiment and Results discussed in Section IV; Finally, we conclude our paper and future work present in Section V.

II. Related Work

Classical machine learning techniques are a common approach to detect automatic hate speech [2,3,8]. Most researchers used deep learning algorithms to learn features automatically from inputs instead of manual sections of adopting features. MacAvaney, S. et al. [5], 2019, proposed a multi-view stacked SVM model to classify hate speech. Using TF-IDF unigram as a feature with a linear SVM classifier and tuned hyperparameters to produce meta-classifiers. They have used different datasets and applied TF-IDF and character N-gram approach to evaluate the performance of the proposed model using accuracy and macro-averaged F1 score. The Proposed mSVM approach outperformed better results from all other BERT models in terms of accuracy and F1 score. For some datasets, the neural ensemble approach performed better for Twitter data than mSVM – based solution. Rezvan, M. et al. [6], 2020, introduced a category of five types of harassment and presented experiments on each category of harassment with a type of aware tweets corpora. They have used classical machine learning algorithms (SVM, KNN, GBM, NB) to train a binary classifier for harassment detection. Comparative study of various features on the performance of machine learning classifiers using accuracy and specificity based on the category of harassment. Performance of each type of harassment is not performed better results. Salminen, J. et al. [7], 2020, used several machine learning models (Logistic Regression, NB, SVM, XGBoost and Neural Networks) for online hate speech detection using multi-platforms (Twitter and Wikipedia). The best performance is BERT as a feature with XGBoost classifier for impactful representation of hateful comments. Similarly, FFNN has the highest performance to detect hate speech. Further improvements can be possible using Deep learning models and sets of parameters.

III. Methodology

We have used LSTM [1]based deep neural architecture with extract features automatically in our given experiment. Our proposed method has been implemented using Tensorflow, Keras library and python programming language. The description of our model has been given in Section B. the annotated dataset was used by the author in [1] and containing 24783 English texts have been categorized into Hate, Offensive and Neutral class.

A. Preprocessing

In Social media, data are unstructured and noise characters that affect the performance of our method. Currently, we have Twitter text data and must remove some text or special characters that reduce the performance of our proposed method. Twitter dataset using the following process of Normalized Text as follows.

- Removal of users, date, email, Urls and re-tweets.
- Removing punctuation i.e &; — , : ?
- Remove stop words
- Removing of hastag symbol from words, '#Foreignersmustfall' becomes 'Foreigners must fall'.
- Apply stemming and convert all texts into lowercase.

B. Proposed Method

The proposed Deep Learning Architecture of hate speech detection is shown in Fig. 1. We have used Pre-word embedding Glove 6B 300 Dimensional as Layer added in our model. The first step, to learn all words from the Twitter Dataset after pre-processing in embedding layer 120 x 300 matrix, in which the length of each sentence should be not greater than 120. Second step, to pass features into dropout (0.2) to stop overfitting. The next step is, the output passing into Bidirectional Long Short-Term Memory (LSTM) layer having the size of 32 x2 (64), that capture long term dependencies to extract features from input and then fed into the fully connected Dense Layer (64 x 6) matrix with Relu Activation function. Finally, the last dense layer with 3 units (Hate, offensive and Neither) with Softmax functions to predict output depends on the task.

We used Spatial Categorical Cross Entropy as a loss function and Adam optimizer (0.001) using 10 epochs to train our proposed model. The details of the proposed deep learning model shown in fig.2.

IV. Experiment and Results

We have used Davidson [2] data set containing 24786 annotated tweets are labeled as Hate, Offensive and Neither. This data set is split into 80% for training using 20 epochs and the rest 20% for validation to measure performance. We used TensorFlow and Keras as a backend in our proposed method with LSTM deep neural networks.

We compared our performance of the proposed method on the validation data for Davidson data set against classical machine learning, deep learning methods and baseline methods are shown in Table 1.

In Table 1, our proposed method performs better than classical and baseline machine learning methods. The accuracy of our method is 92.5% to detect hate speech of our benchmark dataset. In this proposed method, we used the Bidirectional

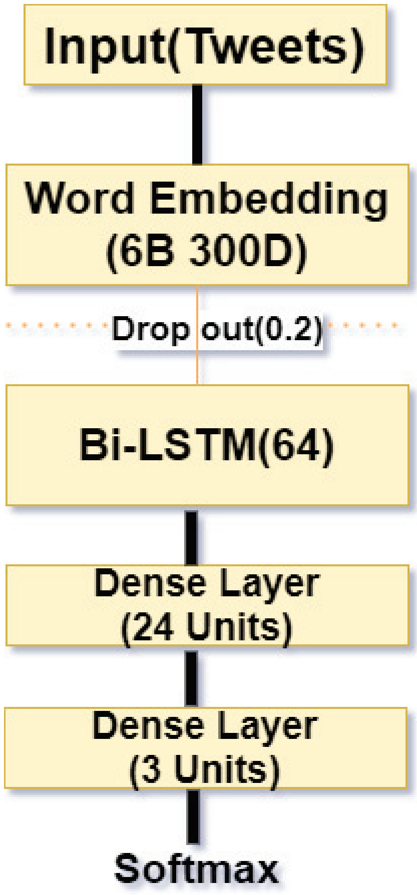
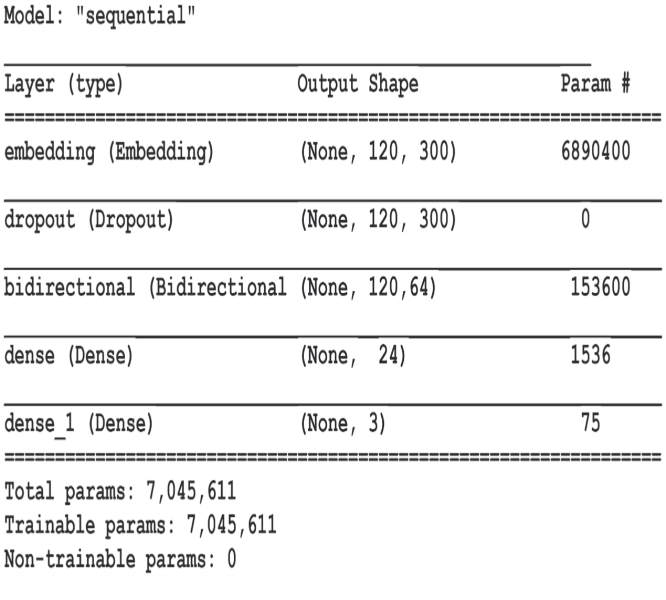


Fig. 1. Architecture of Hate Speech Detection Method



```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 120, 300)          6890400
_________________________________________________________________
dropout (Dropout)            (None, 120, 300)            0
_________________________________________________________________
bidirectional (Bidirectional (None, 120,64)             153600
_________________________________________________________________
dense (Dense)                (None,  24)                1536
_________________________________________________________________
dense_1 (Dense)              (None, 3)                   75
=================================================================
Total params: 7,045,611
Trainable params: 7,045,611
Non-trainable params: 0
```

Fig. 2. Details of our Proposed LSTM based Method

TABLE I
DETECTION ACCURACY OF MACHINE LEARNING METHODS AND OUR PROPOSED METHODS

Models	Accuracy (%)
SGDClassifier	75.87
LogisticRegression	85.21
LogisticRegressionCV	85.21
LinearSVC(SVM)	82
RandomForestClassifier	81.14
CNN	87
Proposed CNN-LSTM	**92.50**

LSTM (64) 32*2 stacked layers to handle long-term dependencies and extract features from fully connected dense layers to increase the performance of our method in terms of F1 score and accuracy. The behavior of our training and Validation of proposed method in the terms of performance and loss for the 10 epochs are shown in fig 3 and fig 4.

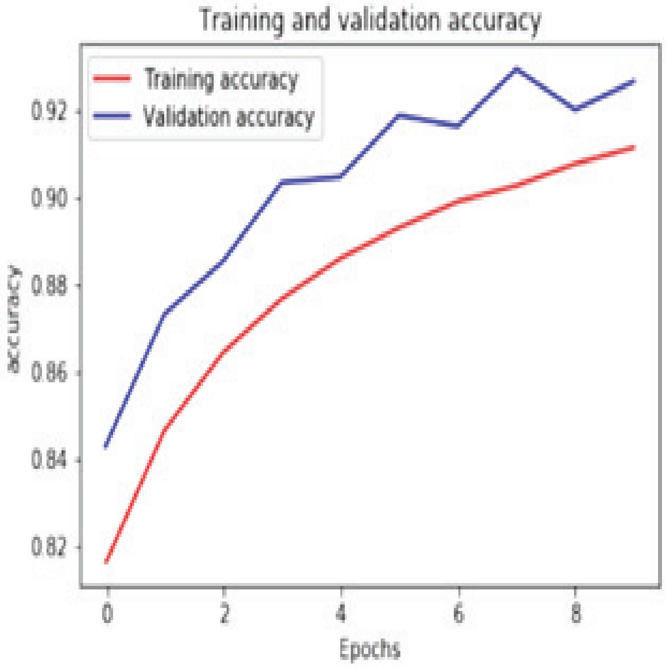


Fig. 3. Training and Validation Performance

V. CONCLUSION

In this paper, we proposed a novel approach for analyzing the class of our tweets as hate, offensive and neither from Davidson data set using LSTM based deep learning architecture. Our proposed method used Bidirectional LSTM (64) stacked layers with tuned hyperparameters to outperform better results against classical machine learning and baseline methods on our Davidson dataset. The overall accuracy of the proposed method is 92.5%. We can further extend and use advanced deep learning architecture to detect hate speech and explore more social media datasets for hate speech.

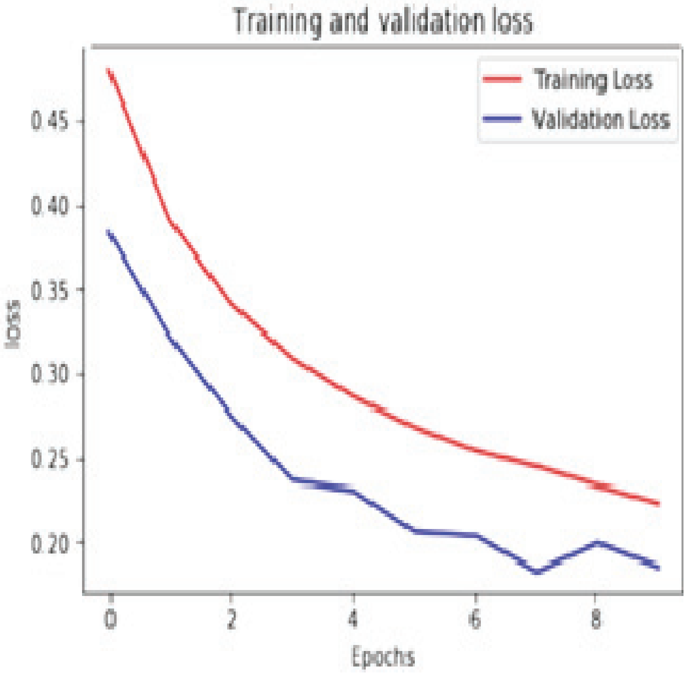


Fig. 4. Training and Validation Loss

REFERENCES

[1] Hochreiter, S., Schmidhuber, J., 1997. LSTM can solve hard long time lag problems. In Advances in Neural Information Processing Systems, (Neural information processing systems foundation), pp. 473–479

[2] Davidson, T. et al., 2017. Automated hate speech detection and the problem of offensive language, in Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017. AAAI Press, pp. 512–515.

[3] Burnap, P., Williams, M.L., 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. Policy and Internet, pp. 223–242.

[4] Malmasi, S., Zampieri, M., 2017. Detecting hate speech in social media, in: International Conference Recent Advances in Natural Language Pro cessing,RANLP.Association for Computational Linguistics (ACL), pp. 467–472.

[5] S. MacAvaney, H. R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, "Hate speech detection: Challenges and solutions," PLoS One, vol. 14, no. 8, 2019, doi: 10.1371/journal.pone.0221152.

[6] M. Rezvan, S. Shekarpour, F. Alshargi, K. Thirunarayan, V. L. Shalin, and A. Sheth, "Analyzing and learning the language for different types of harassment," PLoS One, 2020.

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] J. Salminen, M. Hopf, S. A. Chowdhury, S. gyo Jung, H. Almerekhi, and B. J. Jansen, "Developing an online hate classifier for multiple social media platforms," Human-centric Comput. Inf. Sci., vol. 10, no. 1, 2020, doi: 10.1186/s13673-019-0205-6.

[9] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," ACM Comput. Surv., vol. 51, no. 4, 2018, doi: 10.1145/3232676.

[10] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 2014, pp. 1532–1543, doi: 10.3115/v1/d14-1162.

[11] Tyagi V, Kumar A, Das S. Sentiment Analysis on Twitter Data Using Deep Learning approach. In2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) 2020 Dec 18 (pp. 187-190).

[12] Waseem, Z., Hovy, D., 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. Association for Computational Linguistics (ACL), pp. 88–93.

[13] Gambäck, B., Sikdar, U.K., 2017. Using Convolutional Neural Networks to Classify Hate -Speech. Association for Computational Linguistics (ACL), pp. 85–90.

[14] Kumar, A., Das, S., Tyagi, V., Shaw, R. N., Ghosh, A. (2021). Analysis of Classifier Algorithms to Detect Anti-Money Laundering. In Computationally Intelligent Systems and their Applications (pp. 143-152). Springer, Singapore.