Comparative Analysis of Deep Learning Techniques for the Classification of Hate Speech

*Iorliam, A., Agber, S., Dzungwe, M.P., Kwaghtyo, D.K., and Bum, S.
Department of Mathematics & Computer Science,
Benue State University, Makurdi, Nigeria
Corresponding author:miorliam@yahoo.com.

doi: https//doi.org/10.46912/napas.227

Abstract
Social media provides opportunities for individuals to anonymously communicate and express hateful feelings and opinions at the comfort of their rooms. This anonymity has become a shield for many individuals or groups who use social media to express deep hatred for other individuals or groups, tribes or race, religion, gender, as well as belief systems. In this study, a comparative analysis is performed using Long Short-Term Memory and Convolutional Neural Network deep learning techniques for Hate Speech classification. This analysis demonstrates that the Long Short-Term Memory classifier achieved an accuracy of 92.47%, while the Convolutional Neural Network classifier achieved an accuracy of 92.74%. These results showed that deep learning techniques can effectively classify hate speech from normal speech.

Keywords: Hate Speech, Deep Learning, Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN).

Introduction

The era of information age and instant global communication using the Internet has drastically moved us further away from physical face-to-face communications. Often taken into account is the fact that users of social media are allowed to be anonymous. Therefore, we have a situation where some persons may make offensive and hateful remarks about others without fear of repercussions. Harmful speech or speech that disparages a person or a group, otherwise known as hate speech has been analysed and debated by different researchers in different fields due to the rapid growth in the use of the Internet by people of all cultures and educational backgrounds.

Recently, hate speech gained dominance on social media particularly Twitter and Facebook. Individuals or group create posts on social media that demean or belittle other individuals or group of people. Due to the global effect of hate speech on social media, different approaches have been developed to curb this great challenge. In that regard, Twitter enforced new guidelines to remove hateful conduct and user-initiated hate speech capable of initiating or stirring violence or promoting hatred among users on its site (Twitter, 2020).

This paper has analysed data from Twitter, one of the leading social media channels. Data from Twitter was chosen because it is an internationally recognized real-time public microblogging site and it produces concise data sources for researchers, characterized by its short message limit of 280 characters per tweet. It has a frequency of 500 million tweets per day as at May, 2020 (Sayce, 2020).

Therefore, datasets from Twitter are used as inputs into the LSTM and CNN for classification of hate speech from normal speech.

Related Work

Several deep learning algorithms have proven to classify text datasets efficiently.

For example, Djuric, Zhou, Morris, Grbovic, Radosavljevic and Bhamidipati (2015), proposed to learn a distributed low-dimensional representations of comments using neural language models which can be used as inputs to a binary classifier. The proposed method achieved an AUC of 0.8007.

Ma, Huang, Xiang and Zhou (2015), proposed a framework tagged dependency-based Convolution Neural Networks (DCNN). They used the tree-based n-gram approach based on non-local interactions between words. Experimental results demonstrate that the model achieved a performance of 95.6% accuracy when the model was tested on the TREC dataset.

Nobata, Tetreault, Thomas, Mehdad and Chang (2016), proposed a supervised classification model with Natural Language Processing (NLP) features to surpass deep learning approaches. The features of the model provided a corpus of user comments annotated for abusive language. Results of the experiments showed that the model performed better than other similar approaches as at the time of the research.

Zhao and Wu (2016), leveraged on the traditional Convolutional Neural Network (CNN) to develop an Attention-based Convolutional Neural Network (ATT-CNN). With the attention-based strategy, the model gets hold of the long term contextual information and correlation between non-consecutive words independent of external information or features. Evaluation results using various datasets showed that the ATT-CNN model performed better than the original CNN with a performance accuracy of 94.7% and 96.0% on in-house data and public data, respectively.

Leveraging on morpho-syntactical features, sentiment polarity and word embedding lexicons, Del-Vigna, Cimino, Dell'Orletta, Petrocchi and Tesconi (2017), proposed a framework using Support Vector Machine (SVM) algorithm and Long Short Term Memory (LSTM) algorithm.

The evaluation results of the two algorithms illustrate that they can both classify hate speech effectively.

Badjatiya, Gupta, Gupta and Varma (2017), leveraged on CNN for hate speech detection while using LSTM to process arbitrary sequences of inputs and for capturing long-range dependencies in tweets. The similarity of words was handled with the help of Deep Neural Network (DNN). Random Embedding and Gradient Boosted Decision Tree (GBDT) was used for Fast-Text optimizer. Experimental results proved that a combination of LSTM, Random Embedding and GBDT methods outperformed individual techniques with an F1-score of 93%.

Another study by Zimmerman, Kruschwitz and Fox (2018), developed an ensemble method with neural networks to classify hate speech. The framework utilized public embedded models tested against a hate speech corpus from Twitter. Experimental results

illustrated that the performance of the ensemble model achieved an F-measure of 2% improvement more than the non-ensemble techniques. While a comparative analysis with handcrafted methods from authors of the hate speech dataset achieved a 5% increase.

Georgakopoulos, Tasoulis, Vrahatis and Plagianakos (2018), employed the use of Convolutional Neural Networks (CNN) to distinguish toxic statements in a large pool of text. Experimental results showed that the model outperformed the traditional bag-of-words method of text analysis. The model recorded a prediction performance accuracy of 90% higher than other approaches that achieved 65 to 85 per cent accuracies.

Wang, Li, Cao, Chen and Wang (2019), proposed a hybrid framework called Convolutional Recurrent Neural Network using the Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) classification algorithms. They tested their proposed model on two Chinese datasets and five English datasets and achieved very high classification algorithm that surpassed similar classification algorithms.

Recently, Nistor, Moca, Moldovan, Oprean, and Nistor (2021) proposed a Twitter sentiment analysis using Recurrent Neural Networks. They tested their proposed method using Twitter sentiment analysis training corpus and achieved an accuracy of 80.74%. Even though, several authors have proposed different methods to curb hate speech, the authors are still motivated to perform a comparative analysis on LSTM and CNN to understand which of the algorithms performs better especially when dealing with hate speech classification.

Materials and Methods

For the experiment, the Kaggle dataset (Kaggle, 2020) was used. This dataset contains:

- Serial Number: label for the number of rows for the dataset.
- Count: Number of CrowdFlower users who coded each tweet.
- Hate Speech: Number of CrowdFlower users who judged the tweet to be hate speech.
- Offensive Language: Number of CrowdFlower users who judged the tweet to be offensive speech.
- Neither: Number of CrowdFlower users who judged the tweet to be neither offensive nor offensive.
- Class: Class label for majority of Crowd Flower users. In this study, we coded "1" for hate speech and "0" for neither offensive nor offensive (non-hate speech).
- Tweet: Text tweet.

The columns used in this experiment are in the CSV file format with a total of 24783 rows and 2 columns (Tweet and Class). The dataset is split into 70% and 30% training and testing data set, respectively.

Our model employed the use of two deep learning techniques, precisely the Long- and Short-Term Memory (LSTM) and Convolutional Neural Network (CNN). This is to effectively carry out a comparative analysis of their performance for Hate Speech classification. These algorithms classify speeches on social media posts (Twitter) as to whether they are hate speech or not using Keras framework. The study focuses on comparing the classification performance of the two selected classifiers. Figure 1 represents the architectural flow of the proposed system.

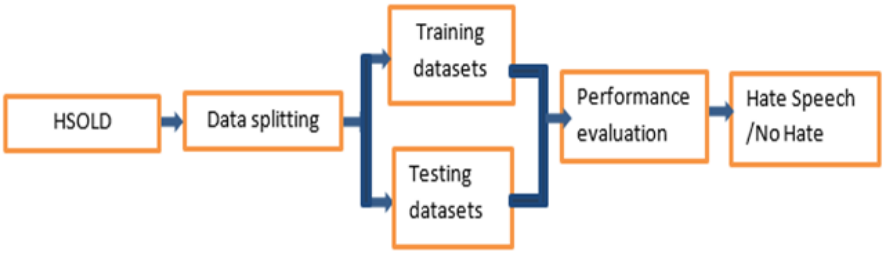

Figure 1: The Architectural Flow of the Proposed Model.

The framework accepts Hate Speech and Offensive Language Datasets (HSOLD). This data set is split into training and testing sets. After training and testing, a performance evaluation is performed to classify the datasets into Hate Speech or No Hate Speech as the case may be.

The purpose of the model is achieved by applying the algorithm shown below:

i. Start
ii. Load the Hate Speech and Offensive Language Datasets (HSOLD)
iii. Split the dataset into training and testing sets
iv. Train and Test the dataset with the LSTM technique.
v. Train and Test the dataset with the CNN technique.
vi. Evaluate the outputs of LSTM and CNN results.
vii. End

Evaluation Metrics

It is import to use appropriate metrics to evaluate a model. This study adopts the use of Recall, Precision, Accuracy, F1-score and Confusion Matrix for evaluation. This is necessary to ascertain the effectiveness and efficiency of the proposed approach over other existing state-of-the-art approaches. The formulae for each of the evaluation measures are given as follows:

Recall: $R = \frac{TP}{TP+FN}$

Precision: $P = \frac{TP}{TP+FP}$

Accuracy: $A = \frac{TP+TN}{TP+FP+FN+TN}$

F_1 Score: $F_1 = 2 \text{ X } \frac{\text{Recall X Precision}}{\text{Recall + Precision}}$

Where:

i. TP (True Positive); is the test result that defines a given condition exists, when it does.
ii. TN (True Negative); defines that a condition does not take place when it does not.
iii. FP (False Positive): is a test result that a given condition exists, when it does not.
iv. FN (False Negative): defines that a condition does not take place when it does.

Results and Discussion

A) The Long and Short Term Memory (LSTM) Technique.

The results of the LSTM achieved a promising Accuracy, F1 Score, Precision and Recall as shown in Figure 2.

```
2000/2000 [==============================] - 1109s 554ms/step
LSTM Accuracy: 92.47%
LSTM F1 Score: 0.812544
LSTM Precision:  0.5757828312894392
LSTM Recall:  0.7940886699507389
```

Figure 2: The LSTM Performance Matrix

In Figure 3, the train and validation loss vs Epochs is illustrated to show at which epoch the LSTM can correctly classify Hate Speech.

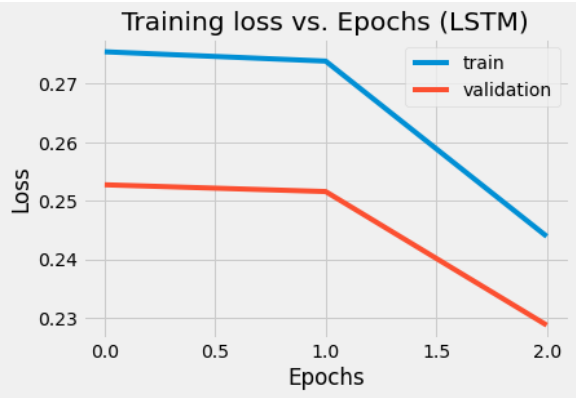

Figure 3: The LSTM Loss vs Epochs

Figure 3 demonstrates that as the Epochs increases especially at Epochs 2, the train and validation loss decreases. This shows that at Epochs 2, Hate Speech can be effectively classified.

The confusion matrix of the LSTM method is shown in Figure 4.

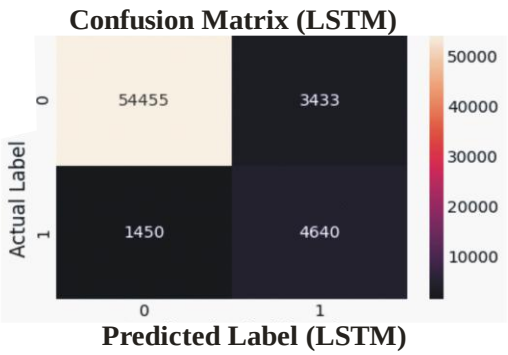

Figure 4: The LSTM Confusion Matrix

B) The Convolutional Neural Network (CNN) Technique.

Figure 5 shows the performance results of the evaluation metrics in terms of the Accuracy, F1 score, Precision and Recall of the convolutional neural network algorithm.

```
2000/2000 [==============================] - 344s 172ms/step
CNN Accuracy: 92.74%
CNN F1 Score: 0.809018
CNN Precision:  0.5957757704569607
CNN Recall:  0.7364532019704434
```

Figure 5: The CNN Performance Matrix

Figure 6 depicts the train and validation loss vs Epochs of the CNN classifier.

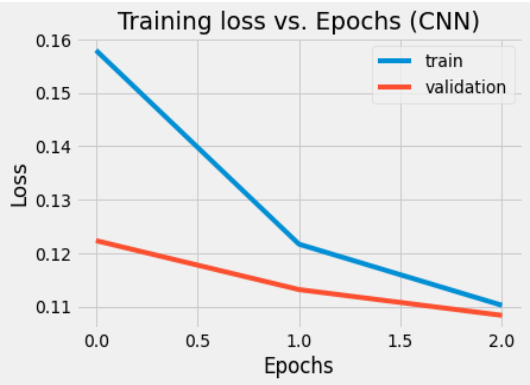

Figure 6: The CNN Loss Vs Epochs

From Figure 6, the Epochs increases as the train and validation loss of the model decrease. This proves the model can effectively classify Hate Speech at Epochs 2.

Figure 7 shows the confusion matrix of the CNN classifier.

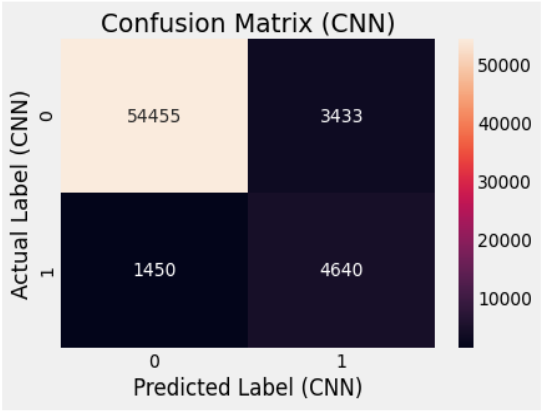

Figure 7: The CNN Confusion Matrix

Comparison of Results from the Two Deep Learning Techniques

Comparatively, the results in sections A and B shows that the CNN achieved an outstanding performance as compared to the LSTM. This is demonstrated in Figures 2 and 5.

Table 1 shows a summary of the results obtained in the two deep learning techniques.

Table 1: Results Summary of Deep Learning Techniques

Deep Learning Techniques	LSTM	CNN
Accuracy	92.47%	92.74%
F1 score	0.81	0.81
Precision	0.58	0.59
Recall	0.79	0.74

Conclusion and Future Work

The inherent complexity of the natural language constructs different forms of hatred, different kinds of targets, and different ways of representing the same meaning. This study however emphasized on the classification of tweets as to whether they are Hate Speech or No Hate Speech. Hence, this study carried out a comparative analysis of deep learning approaches. The LSTM and CNN deep learning algorithms have proven their efficiency in hate speech classification. Results have shown that the

CNN technique is the best and most suitable technique for classifying hate speech due to its outstanding performance of 92.74%. Future work shall consider the contextual usage or meaning of words to avoid classification error due to contextual misuse and misinterpretation of words or statements using the LSTM and CNN deep learning algorithms.

References

Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. (2017). Deep learning for hate speech detection in tweets. In Proceedings of the 26th international conference on World Wide Web companion (pp. 759-760).

Del-Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., and Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on Facebook. In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17) (pp. 86-95).

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., and Bhamidipati, N. (2015). Hate speech detection with comment embeddings. In Proceedings of the 24th international conference on World Wide Web (pp. 29-30).

Georgakopoulos, S. V., Tasoulis, S. K., Vrahatis, A. G., and Plagianakos, V. P. (2018). Convolutional neural networks for toxic comment classification. In Proceedings of the 10th Hellenic conference on artificial intelligence (pp. 1-6).

Ma, M., Huang, L., Xiang, B., and Zhou, B. (2015). Dependency-based convolutional neural networks for sentence embedding. arXiv preprint arXiv:1507.01839.

Nistor, S. C., Moca, M., Moldovan, D., Oprean, D. B., and Nistor, R. L. (2021). Building a Twitter Sentiment Analysis System with Recurrent Neural Networks. Sensors, 21(7), 2266.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content. In Proceedings of the 25th international conference on World Wide Web (pp. 145-153).

Kaggle (2020). Hate Speech and Offensive Language Dataset. Research hate-speech detection. [Online]. Available: https://www.kaggle.com/mrmorj/hate-speech-and-offensive-language-dataset?select=labeled_data.csv.

Sayce, D. (2020). The Number of tweets per day in 2020. [Online]. Available: https://www.dsayce.com/social-media/tweets-day/.

Twitter. (2020). Hateful conduct policy. [Online]. Available: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

Wang, R., Li, Z., Cao, J., Chen, T., and Wang, L. (2019). Convolutional recurrent neural networks for text classification. In 2019 International Joint Conference on Neural Networks (IJCNN) (pp. 1-6). IEEE.

Zhao, Z., and Wu, Y. (2016). Attention-Based Convolutional Neural Networks for Sentence Classification. In INTERSPEECH (pp. 705-709).

Zimmerman, S., Kruschwitz, U., and Fox, C. (2018). Improving hate speech detection with deep learning ensembles. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).