Bima Journal of Science and Technology, Vol. 6 (2) Aug, 2022 ISSN: 2536-6041

A MULTI-PLATFORM APPROACH USING HYBRID DEEP LEARNING MODELS FOR AUTOMATIC DETECTION OF HATE SPEECH ON SOCIAL MEDIA

[1*]HYELLAMADA SIMON, [2]BENSON YUSUF BAHA and [3]ETEMI JOSHUA GARBA

[1]Department of Computer Science, Federal Polytechnic Mubi, Adamawa, Nigeria
[2]Department of Information Technology, Modibbo Adama University Yola, Adamawa, Nigeria
[3]Department of Computer Science, Modibbo Adama University Yola, Adamawa, Nigeria

Corresponding Author: hyellasimon@gmail.com

ABSTRACT

Hate speech on online social networks is a general problem across social media platforms that has the potential of causing physical harm to the society. The growing number of hateful comments on the Internet and the rate at which tweets and posts are published per second on social media make it a challenging task to manually identify and remove the hateful comments from such posts. Although numerous publications have proposed machine learning approaches to detect hate speech and other antisocial online behaviours without concentrating on blocking the hate speech from being published on social media. Similarly, prior publications on deep learning and multi-platform approaches did not work on the topic of detecting hate speech in English language comments on Twitter and Facebook. This paper proposed a deep learning approach based on a hybrid of convolutional neural network (CNN) and long short-term memory (LSTM) with pre-trained GloVe words embedding to automatically detect and block hate speech on multiple social media platforms including Twitter and Facebook. Thus, datasets were collected from Twitter and Facebook which were annotated as hateful and non-hateful. A set of features were extracted from the datasets based on word embedding mechanism, and the word embeddings were fed into our deep learning framework. The experiment was carried out as a three independent tasks approach. The results show that our hybrid CNN-LSTM approach in Task 1 achieved an f1-score of 0.91, Task 2 obtained an f1-score of 0.92, and Task 3 achieved an f1-score of 0.87. Thus, there is outstanding performance in classifying text as Hate speech or non-hate speech in all the considered metrics. Based on the findings, we conclude that hate speech can be detected and blocked on social media before it can reach the public.

Keywords: Hate Speech, Deep Learning, Classification, Word Embedding, Social Media,

INTRODUCTION

The social networking platforms such as Twitter and Facebook are the 21st century's media for information sharing. These platforms encourage social interactions among their users, where individuals composed and post all sorts of information to the public without a system to check the genuineness of such information. These lapses from the social networking platforms elevated the propagation of hate speech, cyberbullying, fake news, etc. which posed a threat to society at large. The perpetrators of such behaviour have the invisibility impression (Aroyehun & Gelbukh, 2018), thinking they cannot be physically arrested and prosecuted for the propagation of all these forms of antisocial behaviour.

Facebook and Twitter have made it compulsory for users to create a profile and make a list of friends and followers to interact with and share the content of interest to them. Generally, most discussions on social networks are political which posed a question of when an expression of one's opinion

becomes offensive, illegal, or immoral and how to deal with it (Jaki & De Smedt, 2018).

Though the Internet and social media, in particular, have greatly enhanced interaction, collaboration, and communication among individuals in different parts of the world. But there is no doubt that those media are sometimes used for spreading fake news and hate speech that targeted religious, ethnic groups, sex, or political groups. The proliferation of cyberbullying and cyber-aggression has caused serious social tension and unrest among different political parties, and religious/cultural beliefs.

Schmidt and Wiegand (2017) considered the term "hate speech" as a broad umbrella term for numerous kinds of insulting user-created content. Bullying and aggressive expressions on the Internet are serious issues affecting most Internet users (Sahay et al., 2018). Hate speech is a more personal and directed speech, and it's mostly informal, angrier, and frequently an open attack on its target via name-calling, using some logical words. In addition, most general hate speeches are dominated by religious hate and characterized by the use of deadly words such as destroy, kill, murder, etc., also extent words like many or million (ElSherief et al., 2018).

Considering the growing number of hateful comments on the Internet and the high volume of tweets and posts that are published on social media every second, manually identifying and removing hate speech from such content is challenging and time-consuming. Although many publications (Agrawal & Awekar, 2018; Salminen et al., 2020; Ridenhour et al., 2020) have proposed numerous models for identifying hate speech and other antisocial online behaviours without blocking the hate speech from reaching the public. Also, most of the hybrid deep learning approaches (Yenala et al., 2018; Asim et al., 2019; Faris et al., 2020) and multiplatform approaches (Bosco et al., 2018; Raiyani et al., 2018; Sigurbergsson & Derczynski, 2019; Salminen et al., 2020) have not concentrated on the issue of detecting and blocking the hate speech it from being posted on the social media, such as Twitter and Facebook.

This motivates the researchers to present a hybrid deep learning model that can automatically identify and block instances of hate speech as they appear across Twitter and Facebook. This will protect the posters and any prospective targets of hate speech from any risks or harm that such comments might cause.

Related works

Deep learning-based models have achieved great success in numerous NLP tasks, such as hate speech detection, cyberbullying, and various sentiment analysis problems. Simon, Baha, & Garba (2022) presented a systematic survey on trends in the use of machine learning algorithms to propose solutions to hate speech propagation on social media. The review shows that despite numerous works on the use of artificial intelligence techniques to detect hate speech on social media, but are still limited in some ways as there is no general standard as to what constitutes hate speech from one region to the other, and the freedom of speech laws are causing drawbacks. Some related works of literature reviewed for this work include:

A deep learning model for the detection of cyberbullying on social media platforms was proposed by (Agrawal & Awekar, 2018). The researchers broadly experimented with three real-world datasets (Twitter 16,000, Formspring 12,000, and Wikipedia 100,000 posts). The work systematically studied cyberbullying detection on various topics across multiple social media platforms (SMP) using deep learning-based models and transfer learning. They experimented with some

machine learning models such as Support Vector Machine (SVM), Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), and deep neural network models (Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), Bi-directional Long Short-Term Memory (BLSTM), and BLSTM with Attention), and used diverse words representation methods (such as bag of word unigram, bag of character n-gram, and GloVe embeddings). Their work revealed that the deep neural network model achieved the best result for detecting cyberbullying on social media platforms.

A similar approach to detect abusive language on Twitter in English, German and Hindi languages was proposed by (Mujadia et al., 2019), the authors experimented with machine learning and neural network-based models. The approach was an ensemble model of SVM, RF, and Adaboost classifiers with majority voting. The experiment was carried out in English, German and Hindi languages and with various combinations of word and character level n-grams for text classification by performing grid-search. The authors observed the combination of word unigrams and character n-grams (where n = 2, 3, 4, 5), TF-IDF vectors, and the combination of character and word level n-grams. The experiment yielded good results with a high level of accuracy.

Salminen et al. (2018) used a machine learning model to develop a model to identify and classify hate speech in online news media. They manually annotate 5,143 hateful comments posted on Facebook and YouTube videos from a given dataset of 137,098 comments on online news media. Using the dataset, they created a granular classification of different types and targets of online hate, and trained machine learning models to automatically detect and classify comments that portray hate. The authors used numerous machine learning techniques, such as LR, RF, DT, Adaboost, and Linear SVM. They generated a multiclass and multilabel classification model to automatically detect and categorize hateful comments in online news media. The results from the experiment revealed that SVM performed the best for the given dataset with an average F1 score of 0.79 using TF-IDF features. The work further revealed that the media and the authorities are the highly targeted group based on the comments in the dataset.

In Chopra et al. (2020) a model was proposed that focused on profanity, author profiling, and deep graph embedding to detect hate speech in Hinglish (Hindi-English) code-switched language on social media platforms. Using two real-world datasets from Twitter, they have shown how targeted hate embedding jointed with social network-based features outperformed the state of the art, both qualitatively and quantitatively. Lastly, they presented an expert-in-the-loop algorithm to eliminate bias in the proposed model pipeline.

A weakly-supervised deep learning model was proposed to reveal hateful users and indirect hateful conversations based on quantitative analysis (Ridenhour et al., 2020). The model considered a content level of interaction and used it for the classification of users who frequently participated in hateful conversations. The experiment revealed that a weakly-supervised model outperforms the baseline models to detect indirect hateful interactions which were evaluated on 19.2M posts. They used the multilayer network embedding techniques to generate features for the prediction task which showed efficient performance.

To classify inappropriate query suggestions, Yenala et al. (2018) developed a hybrid deep learning-based model to classify users' conversations in messengers. They proposed a deep learning architecture called

Convolutional Bi-Directional LSTM (C-BLSTM) which integrates the strengths of both CNN and BLSTM. They used LSTM and BLSTM sequential models to detect inappropriate conversations. The proposed models were trained end-to-end as a single model, and it effectively captured both local features and the global semantics. The C-BLSTM, LSTM, and BLSTM models were evaluated on real-world search queries and conversations, the results of the model significantly outperform both pattern-based and other hand-crafted feature-based baselines.

Similarly, Asim et al. (2019) proposed a hybrid deep learning text-based classification model. The approach proposed two-stage text classification methods which include a filter-based feature selection algorithm, and a deep CNN. They made use of the two most commonly used public datasets (20 Newsgroups data and BBC news data) to evaluate the techniques. The results revealed that the proposed technique outperformed the state-of-the-art of both machine learning and deep learning text-based classification methods, with a margin of 7.7% on 20 Newsgroups and 6.6% on BBC news datasets.

Another hybrid approach to detect hate speech on Twitter in the Arabic comments was proposed by Faris et al. (2020). The authors collected a dataset that contains hateful expressions on Twitter on various subjects of discussions from the Arabic region. They used a word embeddings mechanism to extract a set of features from the dataset. The word embeddings were fed into a hybrid CNN and LSTM framework. The developed approach achieved a great result in classifying tweets as hate or none hate with an accuracy of 71.68%.

A multi-platform model was proposed by Bosco et al. (2018) to detect hate speech on social media (Facebook and Twitter) in the Italian language. The proposed model used NLP techniques for the automatic detection of hateful content on social media. They considered the linguistic and metalinguistic features that differentiate the Italian language on Twitter and Facebook posts due to the differences in ways of the use of the two platforms and the character limitations posed for the messages, especially on Twitter. The authors used two different datasets from the two online social platforms. They trained and tested the model in three different ways: hate speech detection on Twitter, hate speech detection on Facebook, and cross-hate speech detection on Twitter and Facebook. The system achieved the best result with an f1-score of 0.8288 for hate speech detection on Facebook, 0.7993 for hate speech detection on Twitter, 0.6541 for cross-hate speech detection on Facebook, and 0.6985 for cross-hate speech detection on Twitter.

Another multi-platform approach was proposed by Salminen et al. (2018) for the detection of hate speech on social media using SVM, LR, NB, XGBoost, and Neural Networks. The authors used Word2Vec, Bag-of-Words, TF-IDF, BERT, and their combinations as feature representations. They used multi-platform data with a total of 197,566 Arabic language comments from four different platforms (YouTube, Twitter, Reddit, and Wikipedia) with 20% comments annotated as hateful speech, and the remaining 80% as non-hateful. The experiment shows that XGBoost performed the best with an f1-score of 0.92 using all the features.

Similarly, Raiyani et al. (2018) used a dataset of 15,000 aggression-annotated Facebook posts and comments written in Hindi and English languages to develop models based on fully connected Neural networks with an advanced preprocessor to identify aggression over social media (Facebook and Twitter). The model was designed with a Dense

architecture that performs better than Facebook FastText and deep learning models on this dataset. They concluded that the Facebook and Twitter test dataset and the model trained over the Facebook dataset can be used for an unknown Twitter test set, but not vice-versa due to the length of sentences, amounts of hashtags, citation of users, and the amount of retweet.

Sigurbergsson and Derczynski (2019) developed four automatic classification systems, each made to work for both English and Danish language. The authors built a Danish dataset containing user-generated texts from Reddit and Facebook, the dataset was annotated to capture different types and targets of offensive comments. The result of the experiment for the detection of offensive language in English and Danish achieved a great macro averaged f1-score. The experiment performed greatly in capturing the type and targets of offensive languages such as hate speech and cyberbullying in both English and Danish languages. The model was based on the LR classifier, Learned-BLSTM classifier, Fast-BLSTM Classifier, and AUX-Fast-BLSTM Classifier.

The current studies (Salminen et al., 2020; Agrawal & Awekar, 2018; Ridenhour et al., 2020) focused on the detection of hate speeches without blocking the hate speech. Also, based on the reviewed literatures, none of the hybrid deep learning works (Yenala et al., 2018; Asim et al., 2019; Faris et al., 2020) proposed a model for detecting hate speech in English language text on Twitter-Facebook. Likewise, none of the reviewed multi-platform approaches (Salminen et al., 2020; Bosco et al., 2018; Raiyani et al., 2018; Sigurbergsson & Derczynski, 2019) concentrated on Twitter and Facebook platforms.

Therefore, this paper proposed a multi-platforms (Twitter-Facebook) hybrid deep learning model based on CNN and LSTM for automatic detection and blocking of hate speech on social media as it occurred.

MATERIALS AND METHODS

In this section, we explain the development of our proposed hybrid CNN-LSTM model using GloVe pre-trained word embeddings to detect hate speech on multi-platforms (Twitter-Facebook).

Firstly, we collected Twitter and Facebook data from the platforms using their Application Program Interface (API). Secondly, we experimented with some deep learning techniques such as CNN, LSTM, and our hybrid CNN-LSTM to automatically detect hate speech on Twitter-Facebook platforms. Thirdly, we evaluate the performance of our model based on precision, recall, accuracy, and F1 score. Figure 1 shows the overview of our proposed model development stages.

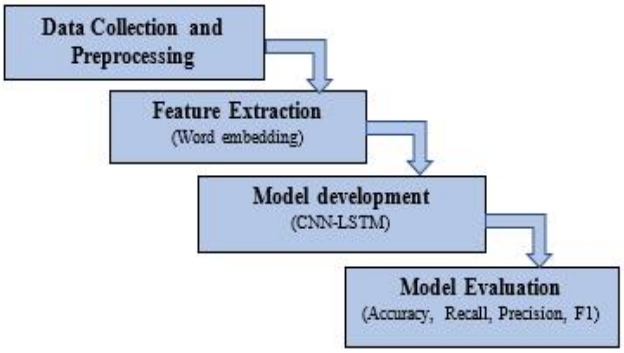

Figure 1: Overview of the model development stages

Data Collection and Preprocessing

We used the following steps for the collection of data using APIs and preprocessed the data using the NLTK library in python programming on a Jupiter notebook.

Step 1: Text data (tweets and posts) were collected from Twitter and Facebook using their APIs. The data collection covers

different critical and debatable areas such as religion, culture, sports, racism, and politics.

Step 2: The collected data was manually annotated as Hate speech (1) or non-Hate speech (0) using human annotators based on the definition of hate speech (Schmidt & Wiegand, 2019).

Step 3: All emoticons and emojis were converted to their words equivalent.

Step 4: All redundant or irrelevant text, numbers, punctuation, stop words, symbols, hashtags, and web addresses were cleaned out.

Step 5: All text was normalized by converting them to lower cases, and noisy data was removed.

Step 6: The tokenization task involved grouping texts into a sequence of discrete tokens (words).

A class-wise distribution of the collected dataset is presented in Table 1.

Table 1: Class-wise distribution of dataset

Class	Twitter-Facebook	Twitter	Facebook
Hate speech (1)	4082	2491	3294
Non-hate speech (0)	32658	17367	14081
Total	36740	19858	17375

Word Representation

Numerous feature representation techniques have been proposed in the literatures for text representation which is acceptable to deep learning algorithms. Bag-of-words-based feature representation techniques used n-grams or specific patterns as features, which are faced with data sparsity problems, and cannot capture the complete contextual information of data (Almeida & Xexeo, 2019). These problems are resolved by word embedding techniques which captured both syntactic and semantic information of text data.

Therefore, we used pre-trained GloVe word embeddings features for our proposed hybrid CNN-LSTM model development. Word embeddings are numerical representations of words that facilitate language understanding using mathematical operations, it relies on a vector space model that captures the relative similarity among individual word vectors, to provide information on the underlying meaning of words (Le & Mikolov, 2014).

Word embeddings have been generally used by some researchers (Altin et al., 2019; Umar et al., 2019) for online offensive and abusive text classification and were found effective. The various pre-trained word embeddings models are GloVe, Word2Vec, and FastText, and all of the models were proven effective for text classification (Salminen et al. 2020; Sigurbergsson & Derczynski, 2019).

The main advantage of GloVe over others is that it does not rely on local statistics, but on the global statistics (word co-occurrence from the entire corpus) to get the word vectors (Faris et al., 2020).

Model Development

The overall flow diagram for the development of our proposed hybrid CNN-LSTM model is presented in Figure 2.

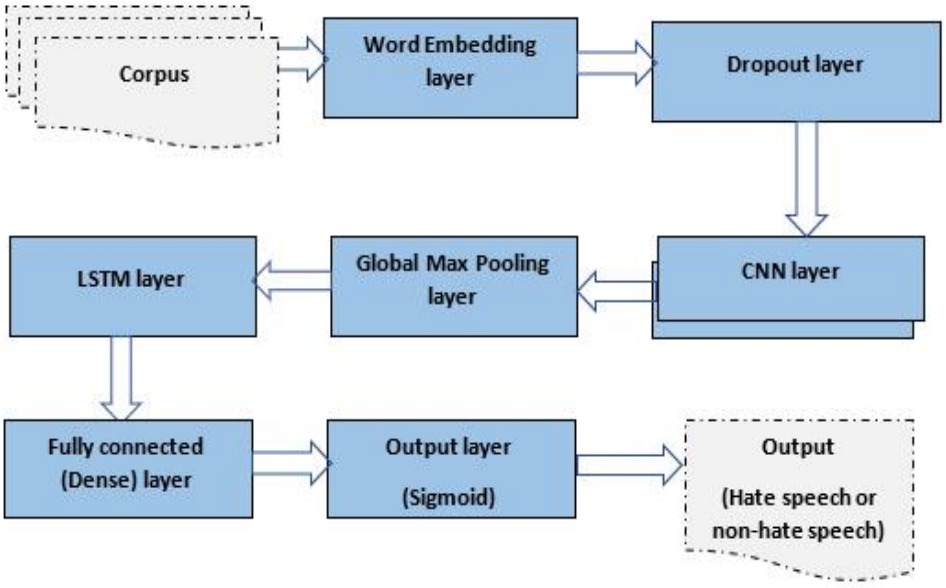

Figure 2: Flow diagram of the hybrid CNN-LSTM model for hate speech detection

Embedding Layer

Considering Figure 2 starts with the embedding layer which was constructed using GloVe pre-trained word embedding (Pennington et al., 2014). This model used a dropout layer to avoid the tendency of overfitting. The aim of using the pre-trained GloVe embedding is to convert each word into a unique vector and learn the words' relationship (Zha et al., 2019). Thus, the proposed approach concentrates on how a model can understand a textual sentence and decide whether hate speech or non-hate speech. The pre-trained GloVe word embeddings used is a package of embeddings with an 822Mb zipped file (Glove.6B.zip specifically Glove.6B.100d.txt). The GloVe model was pre-trained on a dataset of 1billion words (tokens) with a large vocabulary size of 400,000 words. We trained our model on embedding vector size of 100 dimensions.

CNN Layers

CNNs are very successful models in practice (Kresnakova & Sarnovsky, 2019). In this experiment, the convolutional layers receive a sequence of embedding vectors from the embedding layer and produce a tensor as its output. The convolutional layer applies 100 1-dimensional (Conv1D) CNN filters of size 5 over the embeddings, followed by global Max Pooling which was applied for the feature map obtained from each filter. This helps in recognizing the presence of the n-gram feature matching that feature in the text. This is also followed by a dense layer with 256 neurons, ReLU activation, and 0.20 dropout, which helps to compose multiple such features, thus taking a chance to learn a more diverse set of features.

LSTM Layers

The LSTM used in this paper is a kind of RNN that acts as a memory cell. The LSTM structure used the input gate, output gate, and the forget gate, also with additional layers (input and output activation layers). The model solves the problem of the vanishing gradient of the neural networks, and the problem of long-term dependencies (Alwehaibi & Roy, 2017). Thus, it's very efficient in handling sequence of words in a textual dataset (Collobert et al., 2011).

Dense Layer

The dense layer called the fully connected layer was used to take the output of the LSTM and convert it into class labels (probabilities) since we are dealing with a binary classification problem. The output goes through the output layer with 256, and 20 neurons, each having ReLU activations and a 0.20 dropout.

Hybrid CNN-LSTM Model

Advances in deep learning approaches have improved the development of predictive models, particularly CNNs and RNNs have achieved great results in text classification problems (Zhou et al., 2015; Tokala et al., 2018). In this paper, we combine the strengths of CNN and RNN architectures in a single structure for hate speech detection. For instance, CNN cannot capture long-distance dependencies in the input sentences due to the locality of the convolutional and pooling layers, but a single recurrent layer can efficiently overcome this limitation (Hassan & Mahmood, 2017).

Our proposed hybrid CNN-LSTM model employed the use of CNN to extract a sequence of higher-level phrase representations, which was fed into an LSTM structure to obtain the word representation. We used the word embeddings as the input to our CNN model and a 1-D matrix is applied to generate a series of feature maps. After the convolution and pooling operations, the encoded feature maps are taken as the input to the LSTM model. The long-term dependencies learned by LSTM are called sentence-level representations. The sentence-level representation is supplied to a fully connected network and the SoftMax output presents the classification result. Thus, our CNN-LSTM model controls the encoded local features extracted from the CNN model and the long-term dependencies captured by the LSTM model. The final output of the LSTM layer is merged into one matrix, then passed to a fully connected layer, which then produced the result as hate speech or non-hate speech.

Model Training and Testing

This section presents the experimental settings for the application of the collected hate speech datasets into a deep learning framework for training and testing. The dataset was divided into a training set (75%) and a test set (25%) (Salminen et al., 2020). The model training was based on supervised learning which produced a more capable text classifier. The classification module is composed of a fully connected dense layer. The proposed model was trained and tested on a laptop computer equipped with Microsoft Windows 10 operating system, Intel® Core™ i3-4005U CPU @ 1.70GHz & 1.70GHz with 4 GB RAM. The model's front end was developed using the Keras library (Chollet, 2015). and TensorFlow was used for the backend development (Abadi et al., 2016) in python programming which was trained and tested on the Jupiter notebook.

The experiment used 4 hidden layers with the Relu activation function at the first 3 layers, and the sigmoid activation function was used at the output layer. We make use of 2 neurons for the output layer which is the same as the number of classes in our dataset. The batch size used was 128, the verbose rate at 1 and 0.2 dropouts, the loss function was sparse categorical cross-entropy, and the SoftMax activation function using Adam optimizer. Due to computational power, the training was stopped at 10 epochs.

Task Organization

Task 1: Multi-platform: Twitter-Facebook dataset was used to train and test the model.

Task 2: Facebook-Cross-platform: The model was trained with the Twitter dataset and tested with a Facebook dataset.

Task 3: Twitter-Cross-platform: The model was trained with the Facebook dataset and tested with a Twitter dataset.

Evaluation Measures

We used the commonly used standard performance metrics to evaluate our model to ascertain its performance using eq. (1, 2, 3, and 4) based on Precision, Recall, Accuracy, and F1-score as follows (Faris et al., 2020; Sadiq et al., 2020):

Precision: This is the ratio of text that was correctly classified as Hate over the total number of hate texts in the corpus.

$$Precision = \frac{TP}{(TP+FP)} \qquad (1)$$

Recall: This calculates how much the classifier can recognize actual hate texts as hate.

$$Recall = \frac{TP}{(TP+FN)} \qquad (2)$$

Accuracy: This is the amount of correctly classified Hate and non-Hate texts against all the correct and the incorrect number of classified texts.

$$Accuracy = \frac{TP+TN}{(TP+TN+FP+FN)} \qquad (3)$$

F1-Score: This indicates the balance between precision and recall.

$$F1 = 2\ x\ \frac{Precision\ x\ Recall}{(Precision\ +Recall)} \qquad (4)$$

Where: TP – True Positive, TN – True Negative, FP – False Positive, and FN – False Negative.

RESULTS

Experimental Results

The result of Task 1 is presented in Table 2 showing the performance of our model using the Twitter-Facebook dataset in Table 1.

Table 2: Models test result for multi-platform task

Models	RNN	CNN	CNN-LSTM
Precision	0.90	0.91	0.91
Recall	0.85	0.89	0.91
F1-score	0.86	0.90	0.91
Accuracy	0.86	0.89	0.91

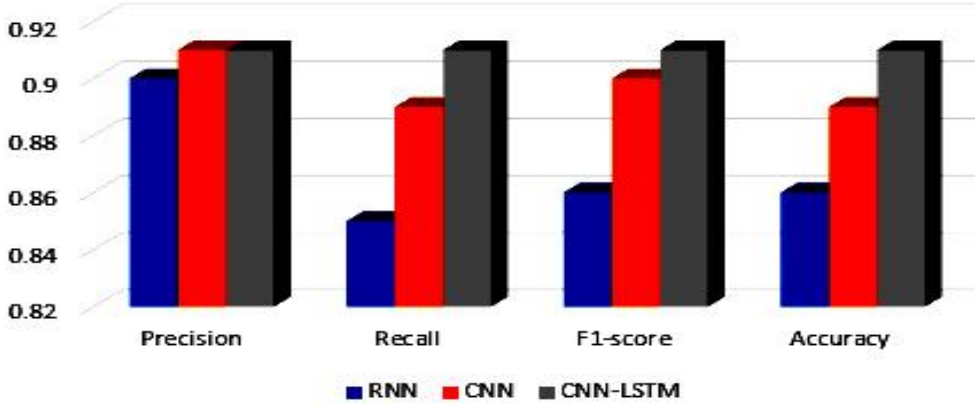

Figure 3: Graphical presentation of result for a multi-platform task

The result of Task 2 (Facebook-Cross-platform) experiment is presented in Table 3 and Figure 4 shows the model performance using Twitter-Train and Facebook-Test datasets in Table 1.

Table 3: Model test result for Facebook-Cross-platform task

Models	RNN	CNN	CNN-LSTM
Precision	0.83	0.85	0.92
Recall	0.80	0.86	0.93
F1-score	0.83	0.84	0.92
Accuracy	0.86	0.86	0.93

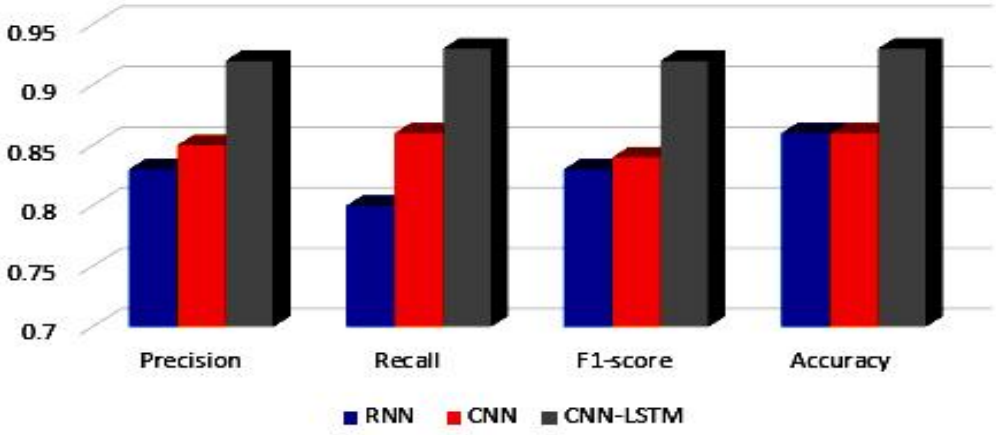

Figure 4: Graphical presentation of test result for Facebook-Cross-platform task

The result of Task 3 (Twitter-Cross-platform) experiment is presented in Table 4 and Figure 5 shows the model performance using Facebook-Train and Twitter-Test dataset in Table 1.

Table 4: Model test result for Twitter-Cross-platform task

Models	RNN	CNN	CNN-LSTM
Precision	0.83	0.84	0.87
Recall	0.81	0.86	0.88
F1-score	0.83	0.82	0.87
Accuracy	0.84	0.86	0.88

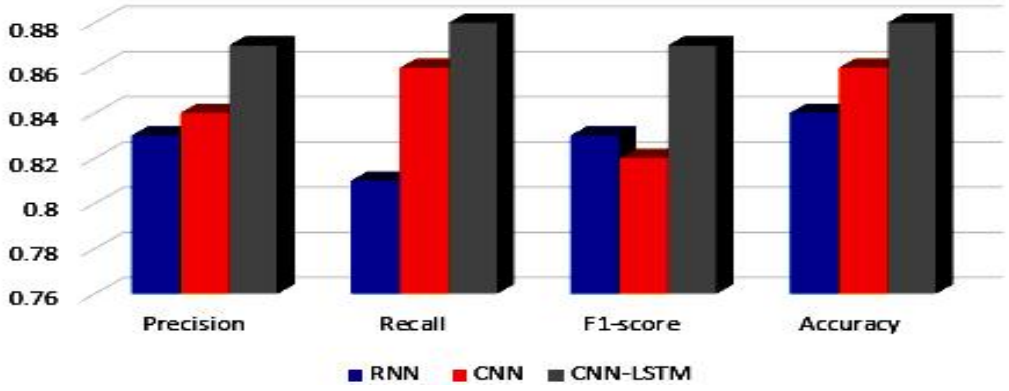

Figure 5: Graphical presentation of test result for Twitter-Cross-platform task

DISCUSSION

In this section, the researcher considers accuracy, precision, recall, and f1-score as standard metrics for the test results. Thus, F1-score was considered to determine our hybrid model's performance in the detection of hate speech.

The test result for Task 1 in Figure 3 shows how well our model has behaved for multi-platform hate speech detection. The result revealed that our proposed CNN-LSTM model with pre-trained Glove embeddings obtained an f1-score of 0.91 which outperformed RNN and CNN with 5% and CNN with 1% respectively, indicating that our model is efficient compared to similar work (Zhou et al., 2015). Regarding all metrics for multi-platforms classification, our proposed model yielded good metrics with precision, recall, and accuracy of 0.91 each.

Similarly, Figure 4 presented the result of the Task 2 experiment where the model was trained with the Twitter dataset and tested with a Facebook dataset for the task of Facebook-cross-platform detection of hate speech (Faris et al., 2020). The result in figure 4 revealed that our hybrid CNN-LSTM model having an f1-score of 0.92 outperformed the RNN model with 9% and the CNN model with 8%. Also, considering other metrics for Facebook-cross-platform classification, our proposed model produced a good result with a precision of 0.92, recall of 0.93, and accuracy of 0.93.

Another result for the Twitter-cross-platform approach was presented in Figure 5, showing the performance of our classifier (Salminen et al., 2020). Figure 5 shows the result of the Task 3 experiment where the Facebook dataset was used for model training and the test was done using the Twitter dataset in Table 1. Figure 5 revealed that our hybrid CNN-LSTM model with pre-trained Glove embeddings having an f1-score of 0.87 shows a significant increase in RNN with an f1-score of 0.83 and CNN with an f1-score of 0.82. This means that our hybrid CNN-LSTM model outperformed both RNN and CNN models by 4% and 5% respectively. Regarding all metrics, our model achieved 0.87, 0.88, and 0.88, for precision, recall, and accuracy respectively.

CONCLUSION

This paper examined the prevalence of hate speech propagation on social media and the various approaches presented in literatures to control such acts. Based on the gaps identified in the literatures, this research has successfully developed a hybrid deep learning model that can detect and block hate speech on social media platforms (Twitter and Facebook). In this paper, we experimented with various deep learning models (RNN, CNN, and hybrid CNN-LSTM) for online hate speech detection. The best performance for both Twitter and Facebook was found with hybrid CNN-LSTM as a classifier with GloVe pre-trained word embedding in all experiments. Our approach was based on three tasks (Task 1, Task 2, and Task 3), and the results revealed that our hybrid deep learning CNN-LSTM model outperformed other deep learning models in all the tasks. Based on the findings in this paper, we concluded that hate speech can be detected and blocked on social media platforms before it can reach the public.

Therefore, this paper recommends that the findings be considered by Twitter and Facebook to curtail the menace of online hate speech propagation.

This paper pinpoints challenges and areas for further study: the need for a large standard English text multi-platforms datasets for hate speech classification tasks, the need for a more comprehensive vocabulary resource of offensive, abusive, and hateful expressions for advanced improvement in hate speech detection tasks. Also developing a deep learning model that can detect hate speech in multimedia content on social media.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., & Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation (OSDI 2016),* Berkeley, CA, USA, (2016) 265–283.

Agrawal, S. & Awekar, A. (2018). Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms. *Springer, BBC News Article,* 141–153.

Aroyehun, T. S. & Gelbukh, A. (2018). Aggression Detection in Social Media: Using Deep Neural Networks, Data Augmentation, and Pseudo Labeling. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying.* Santa Fe, USA, 90–97.

Asim, M. N., Khan, M.G.U., Malik, M.I., Dengel, A., & Ahmed, S. (2019). A Robust Hybrid Approach for Textual Document Classification.

Almeida, F. & Xexeo, G. (2019). Word Embeddings: A Survey. *Computation and Language*.

Altin, L. S., Bravo, A., & Saggion, H. (2019). LaSTUS/TALN at SemEval-2019 Task 6: Identification and Categorization of Offensive Language in Social Media with Attention-based Bi-LSTM model. *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019),* Minneapolis, Minnesota, USA, 672–677.

Alwehaibi, A. & Roy, K. (2017). Comparison of Pre-Trained Word Vectors for Arabic Text Classification Using Deep Learning Approach. *17th IEEE International Conference on Machine Learning and Applications*, 1471-1474.

Bosco, C., Dell'Orletta, F., Poletto, F., Sanguinetti, M., & Tesconi, M. (2018). Overview of the EVALITA 2018 Hate Speech Detection Task. URL: http://ceur-sw.org/vol-2263/paper010.pdf

Chollet, F. (2015). Deep learning for humans. *Keras-team GitHub.* URL: https://github.com/fchollet/keras.

Chopra, S., Sawhney, R., Mathur, P., & Shah, R.R. (2020). Hindi-English Hate Speech Detection: Author Profiling, Debasing, and Practical Perspectives. *Association for the Advancement of Artificial Intelligence.*

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of machine learning research,* 12(08): 2493–2537.

ElSherief, M., Kulkarni, V., Nguyen, D, YangWang, W., & Belding, E. (2018). Hate Lingo: A Target-Based Linguistic Analysis of Hate Speech in Social Media. *Proceedings of the Twelfth International AAAI Conference on Web and social media,* 42-51.

Faris, H., Aljarah, I., Habib, M., & Castillo, P.A. (2020). Hate Speech Detection using Word Embedding and Deep Learning in the Arabic Language Context. *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods,* 453-460.

Hassan, A., & Mahmood, A. (2017). Deep Learning approach for sentiment analysis of short texts. 3rd International Conference on Control, Automation and Robotics, 705-710.

Jaki, S. & De Smedt, T. (2018). Right-wing German Hate Speech on Twitter: Analysis and Automatic Detection. Computation and Language (cs.CL). *Social and Information Networks.* DOI: arXiv:1910.07518 [cs.CL]

Kresnakova, M., Sarnovsky, M.V., & Butka, P. (2019). Deep learning methods for Fake News detection. *IEEE 19th International Symposium on Computational Intelligence and Informatics and 7th IEEE International Conference on Recent Achievements in Mechatronics, Automation, Computer Sciences and Robotics,* 000143-000148.

Le, V. Q. & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *Computation and Language*.

Mujadia, V., Mishra, P., & Sharma, D. M. (2019). IIIT-Hyderabad at HASOC 2019: Hate Speech Detection. URL: http://ceur-ws.org/Vol-2517/T3-12.pdf

Pennington, J., Socher, R., & Manning, C.D. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing,* 1532–1543.

Raiyani, K., Goncalves, T., Quaresma, P., & Nogueira, V.B. (2018). Fully Connected Neural Network with Advance Preprocessor to Identify Aggression over Facebook and Twitter. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullyin.* Santa Fe, USA, 28–41.

Ridenhour, M., Bagavathi, A., Raisi, E., & Krishnan, S. (2018). Detecting Online Hate Speech: Approaches Using Weak Supervision and Network Embedding Models. *Computer Science: Social and Information Networks*. DOI: arXiv:2007.12724 [cs.SI]

Sadiq, S., Mehmood, A., Ullah, S., Ahmad, M., Choi, G.S., & On, B.W. (2020). Aggression detection through deep neural model on Twitter. *Future Generation Computer Systems*, 114, 120–129.

Sahay, K., Khaira, H. S., Kukreja, P., & Shukla, N. (2018). Detecting Cyberbullying and Aggression in Social Commentary using NLP and Machine Learning. *International Journal of Engineering Technology Science and Research.,* 5(1): 1428-1435.

Salminen, J., Hopf, M., Chowdhury, S.A., Jung, S., Almerekhi, H., & Jansen, B.J. (2020). Developing an online hate classifier for multiple social media platforms. *Human-centric*

Computing and Information Sciences, 10(1).

Salminen, J., Almerekhi, H., Milenkovic, M., Jung, S., An, J., Kwak, H., & Jansen, B.J. (2018). Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. *Proceedings of the Twelfth International AAAI Conference on Web and Social Media,* 330-339.

Schmidt, A. & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media.* Valencia, Spain, 1–10.

Sigurbergsson, G.I. & Derczynski, L. (2019). Offensive language and hate speech detection For Danish. URL: https://arXiv.org/abs/1908.04531

Simon, H., Baha, B.Y., & Garba, E.J. (2022). Trends in machine learning on automatic detection of hate speech on social media platforms: A Systematic review. *FUW Trends in Science & Technology Journal,* 7(1): 001 – 016. e-ISSN:24085162

Tokala, S., Gambhir, V., & Mukherjee, A. (2018). Deep Learning for Social Media Health Text Classification. *In Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task,* Brussels, Belgium, 61–64.

Umar, A., Bashir, S., Ochei, L.C., & Adeyanju, I.A. (2019). Profiling Inappropriate Users' Tweets Using Deep Long Short-Term Memory (LSTM) Neural Network. *I-Manager's Journal on Pattern Recognition,* 5(4).

Yenala, H., Jhanwar, A., Chinnakotla, M.K., & Goyal, J. (2018). Deep learning for detecting inappropriate content in text. *International Journal of Data Science and Analytics,* 6, 273–286.

Zha, Z., Liu, J., Yang, T., & Zhang, Y. (2019). Spatiotemporal-Textual Co-Attention Network for Video Question Answering. *ACM Transactions on Multimedia Computing, Communications, and Applications,* 15(25): 1-18.

Zhou, C., Sun, C., Liu, Z., & Lau, F. (2015). A C-LSTM neural network for text classification. *arXiv preprint*.