Hate Speech Detection in Twitter using Recurrent Neural Networks

L.Santhosh Kumar Undergraduate Student Department of Computer ScienceIFET College of Engineering Villupuram District, Tamilnadu, India

K.Vigneshwari Assistant Professor
Department of Computer Science IFET College of Engineering Villupuram District, Tamilnadu, India

***Abstract*— The exponential increase in Internet users brought up undesirable cyber concerns, such as cyberbullying, hate speech, and many others. The issues with hate speech on Twitter are covered in this article. The protected characteristics that are the target of the hate speech include gender, religion, race, and disability. This article discusses an experiment on the identification of tweets connected to hate speech on Twitter using a Gated Recurrent Unit model and features retrieved using the TF-IDF technique. With the aid of a convolution technique, this model uses the tweet text to capture the semantics of the tweets and obtained an accuracy of 95.69%.**

Keywords—Hatespeech, Twitter, Recurrent Neural Network, Long Short Term Memory, Gated Recurrent Unit

I. INTRODUCTION

Online bullying and harassment are referred to as cyberbullying and cyberharassment. The technological world has grown and improved, becoming more and more widespread, especially among young people. Cyberbullying is the practise of abusing or assaulting someone online, especially on social media sites. Harmful bullying activity (i.e., hate speech) includes a variety of actions, such as posting rumours, blackmail, sexual insults, a victim's personal information, or labels that are negative. Bullying and harassment are indicated by persistent actions and a desire to injure [1]. OSN platforms like Twitter, Instagram, Facebook, WhatsApp, WeChat, and others are widely used by users to communicate with one another [2]. The fact that information is available in a multitude of media, including audio, video, and photos, is one of the factors contributing to these platforms' appeal. These facts span a variety of topics, including politics, technology, science, music, and nature and space. There is barely any aspect of life that these OSNs haven't touched; they satisfy everyone's needs, hence users prefer to spend their time there [3], [4]. Cyberbully includes:

- Accessing a person's gaming or social networking account.
- Spreading offensive messages online while impersonating as someone else.
- Using abusive language in emails, texts, or instant messages (IMs) sent to another individual.
- Sharing private or embarrassing images online.
- Verbally abusing other players in online multiplayer games, console games with Internet access, and virtual worlds.
- Using someone else's password to sign in as them online
- Spreading rumours, secrets or smears about someone else that will harm their reputation.

Twitter is a unique sort of OSN where messages are only allowed to be 280 characters in length. Additionally, Twitter offers a follower and followee based interface in place of friend requests. This implies that you must follow other users in order to see their updates (friends, family, or favourite actors, for example). Similar to that, [5] requires that someone follow you in order to read your posts. Twitter is typically used by users to view the most recent international news as well as updates from their personal and professional contacts. Users are free to share anything on Twitter since there are no restrictions on what may be said; as a result, it is simple to publish abusive messages and unpleasant comments [6]–[8]. These issues motivated us to create a blueprint that will include as many postings connected to hate speech as we could.

The neural network architecture for NLP most frequently employed is the recurrent neural network. It has shown to be reasonably accurate and effective when used for text mining and language model development.

The main objectives of this article include:

- To enhance Twitter's ability to recognise hate speech in posts.
- To reduce the overhead of the manual feature extraction process.
- To outperform current models and achieve high accuracy in prediction on an unbalanced dataset.

The rest of the paper is structured as follows. Section 2 contains an overview of the research using deep learning approaches. Section 3 provides existing approach. Section 4 provides in-depth information about deep learning and various recurrent neural network models and performance

measures were discussed. Finally, Section 5 brings the article to a close.

II. Literature Review

In order to forecast an individual's personality, V. Balakrishnan et al. [9] proposed the identification of Big Five personality traits by means of IBM Watson's Personality Insights API. The study looked at the connections between the Big Five and the Dark Triad to further bring the darker qualities into the detection model. A total of 9484 tweets are included in the dataset, and they are divided into four categories: spammers (N=3208; %=33.8), aggressors (N=336;%=3.5), bullies (N=528;%=5.6), and normal (N=5608; %=59%). The detection of cyberbullying is substantially improved by taking the consumer's nature into account. Extroversion, conciliatory behaviour, neurotic traits, and psychopathic behaviour (Dark Triad) were discovered to be particularly essential in spotting bullies, with up to 95% recall and 96% precision.

To identify text and image-based cyberbullying, Kumari K et al. [10] suggested a hybrid model of Genetic Algorithm-Convolutional Neural Network-VGG-16 (GA-CNN-VGG16). To train and evaluate the suggested model, datasets from a range of social networks, including Twitter, Instagram, Facebook, etc. were gathered. This dataset consists of 2100 posts, of which 1481 are bullying-related and 619 are not. Each post includes an image and a comment. The weighted Precision, Recall, and F1 scores of (GA-CNN-VGG16) were found to be 80%, 79%, and 78%, respectively, according to experimental data.

S. Sadiq et al. [11] addressed the issue of automatically identifying aggression detection using a dataset of 20001 tweets from cybertrolls, of which 7822 (39%) are cyber-aggressive and 12,179 (61%) are not. They successfully implemented Multilayer Perceptron (MLP) with TF-IDF. The most accurate at 92% for detecting hostile behaviour was MLP.

The performance of the Dolphin Echolocation Algorithm-Elman Recurrent Neural Network (DEA-RNN) was carefully assessed by B.A.H. Murshed et al. [12] using a dataset of 10,000 tweets, of which 6,508 (65%) are non-cyberbullying tweets and 3492 (35%) are cyberbullying tweets. The results of the experiments proved that the DEA-RNN has superior accuracy, precision, recall, F1-score, and specificity (90.45%, 89.52%, 88.98%, and 90.94%).

So as to determine the features relating to online governmental and assets data on social media, A Valliappan et al. [18] stated an algorithm-based technique based on the LSTM, RNN, and BERT algorithms will be used. Our LSTM architecture has strong accuracy rates and performs well with datasets. Nonetheless, it moves a little more slowly than BERT architecture. When adopting the BERT architecture, better accuracy rates are obtained in earlier epochs.

S Abarna et al. [19] propose K-BERT (Knowledge-enabled BERT), a Deep Learning technique for categorising idioms and literals that incorporates accepting graphs into the phrases as domain information. Additionally, the ensemble will be constructed by stacking foundational systems like BERT and RoBERTa. The dataset such as Trofi Metaphor was used to train the model in this study, and a newly created internal dataset was used to test it.

A Kumar et al. [20], For the purpose of detecting cyberbullying in the text of social media, a hybrid model called Bi-GRU-Attention-CapsNet (Bi-GAC) has been presented. It gains advantage from learning sequential representations of semantics and geographical information about a location using a Bi-GRU along with self-attention subsequent to CapsNet. F1-score and the ROC-AUC curve are used as performance indicators to assess the proposed Bi-GAC model. On the standard Formspring.me and MySpace datasets, the results demonstrate an improvement in performance over the previous approaches.

III. Existing Work

The most popular artificial neural network for evaluating visual imagery is the convolutional neural network (CNN). CNNs are also known as Shift Invariant or Space Invariant Artificial Neural Networks (SIANN), and they are based on the shared-weight design of the convolution kernels or filters that move along input features and produce translation-equivariant outputs known as feature maps. The bulk of convolutional neural networks do not understand independent since they downsample the input. In addition to picture and video recognition, systems for recommendation, image classification, image segmentation, and medical image analysis are some of the tasks they can be utilised for. They can also be utilised for image categorization, financial time series, brain-computer interfaces, natural language processing, and financial time series [13].

IV. Proposed Work

The sequential learning issue in the standard neural network can be solved by using recurrent neural networks (RNN). They are known as "stateful" due to their unique ability to store information while reading the input sequence at each time step. A key characteristic of RNNs is constraint sharing, which allows the model to be applied to input arrangements of various lengths. With differences in the connections between the neurons, RNNs and feed-forward neural networks have a similar basic structure. The neurons in the network are allowed to have directed rounds in place of unidirectional linkages, which allow data to move from one layer to the next. They may have linkages or a self-loop.

The output from some nodes may influence subsequent input to the same nodes in a recurrent neural network (RNN), a collection of artificial neural networks in which connections between nodes can constitute a loop. Recurrent neural networks can be used to represent problems involving

time- and sequence-dependent data, such as text production, artificial intelligence, and stock market predictions. The Recurrent Neural Network includes models like the Gated Recurrent Unit (GRU) and Long Short Term Memory (LSTM).

A. Long Short Term Memory (LSTM)

In the field of deep learning, a type of recurrent neural network (RNN) architecture called Long Short Term Memory (LSTM) is employed. LSTM has a link to feedback that is distinct from feed-forward neural networks in general. It can process complete information sequences as well as single data pieces. A cell, an input gate, an output gate, and a forget gate make up an LSTM unit. The three gates regulate the data flow into and out of the cell since the cell recollects values over unpredictable time gaps. Because LSTM networks are effective at classifying, managing, and generating hypotheses based on statistical data, time gaps between the major events in time sequences are frequently present. When training conventional RNNs, discharging and disappearing gradient issues would arise. LSTMs were developed to address these issues [14].

The LSTM unit houses a memory cell with the following three gates:

Input gate: The total of input that is permitted to flow through the input gate is computed by the formula:

$$i = (xt\ Ui + st\text{-}1\ Wi) \quad (1)$$

The weight vector (Ui) is multiplied by the input value before the sigmoid function draws it between [0, 1]. This aids the gate in controlling the flow of input via input gate.

Forget gate (f): It aids the network in determining what information from the preceding level to send and in what quantity. This function's value is mapped by the sigmoid function between 0 and 1. It is denoted by the formula:

$$f = (xt\ Uf + st\text{-}1\ Wf) \quad (2)$$

When the preceding memory is multiplied by the zero vector, the input value zero is produced if no input is desired to be sent to the following level. The memory at level st-1 is multiplied by one vector if it needs to move up. The output vector is multiplied by the input vector if only a portion of the input will be given.

Output gate (o): It specifies the output that is passed at each network stage. It is calculated as follows:

$$o = (xt\ Uo + st\text{-}1\ Wo) \quad (3)$$

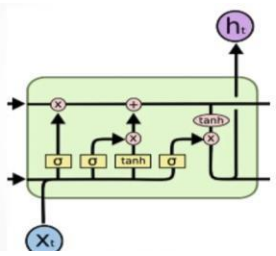

Fig. 1. LSTM cell

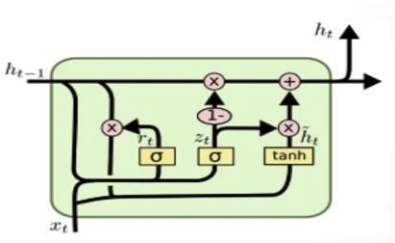

Fig. 2. GRU cell

B. Gated Recurrent Unit

By utilising update gates and reset gates, the GRU deep learning algorithm improves upon the LSTM technique to reduce algorithm complexity. The concealed state volume that is forwarded to the following state is controlled by the update gate. the result of the preceding concealed state information is defined using the reset gate [15].

Update Gate (z): It establishes how much previous knowledge should be transmitted into the forthcoming. It is comparable to the LSTM's Output Gate.

$$z = (xt\ Uz + st\text{-}1\ Wz) \quad (4)$$

Reset Gate (r): The amount of prior knowledge that should be forgotten is specified. It is comparable to how the Input Gate and Forget Gate work together in an LSTM model.

$$r = (xt\ Ui + st\text{-}1\ Wr) \quad (5)$$

V. EXPERIMENTS

A. Dataset Information

Totally 31,962 tweets were included in the dataset, of which 29,720 (92.98%) are Non-Hate Speech (NHS) and 2,242 (7.02%) are related to Hate Speech (HS).

B. Experiments

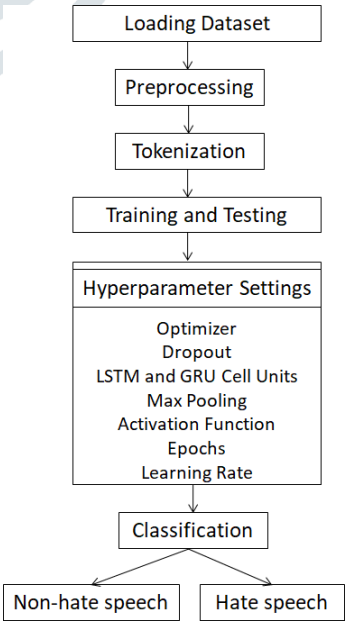

Fig. 3. Architecture diagram of Proposed System

Preprocessing

Preprocessing is the process of renovating raw data into something that can be used to create and train machine learning and deep learning models. The preprocessing process includes cleaning and preparing data for use is a necessary task that also improves the effectiveness of a machine/deep learning model. The non-word characters, whitespace, punctuations should be removed. The uppercase characters has replaced into lowercase characters. The stemming has applied to remove stop words. Finally, the preprocessed text is converted into TF-IDF vectorizer. The output is categorized and can be labelled numerically using label encoder.

Normalization

The process of normalisation involves translating the numerical columns' values to a standard scale. The most widely employed normalising methods includes Standard Scaler and MinMax Scaler.

TF-IDF Vectorizer

The TfidfVectorizer uses an in-memory lexicon to create a sparse word frequency matrix by mapping the most widely used words to feature indices.

Dividing the dataset into the test set and the training set

While preparing data for machine/deep learning, dataset can be separated into a training set and a test set.

Training set: A subcategory of the dataset that is used to fine-tune the deep learning model's performance and whose outcomes are already known.
Test set: A portion of the dataset used to gauge the deep learning model's performance that predicts the outcomes based on the test set.

Hyperparameter Settings

Activation Function

The activation function returns the top feature from the word vectors for the convolved features. It is put into practise by concatenating features from different windows with different filters. The vector values between the defined ranges are nonlinear function like tanh, ReLU or sigmoid.

Optimizer

The optimizer's job is to decrease the error rate of the model in order to increase its accuracy. Adam, SGD and Adagrad are commonly used optimizers.

Dropout

Dropout is a user-dependent variable with an input range of 0 to 1, and it is typically used to lessen the complexity between the links in the fully linked dense layer.

Epoch

The number of repetitions of a training approach is referred to as the epoch. The quantity of epochs is taken into consideration depending on the training data.

Max Pooling

The "max pooling" method of pooling chooses the biggest component from the feature map area that the filter was covered. Therefore, the salient features from the preceding feature map would be included in the feature map created by the max-pooling layer.

LSTM and GRU units

Memory cells in the LSTM and GRU network are represented by the number of LSTM and GRU units that indicates the ability to recall the details and compare them to prior evidence. For continued training, the information in the memory units is transferred forward in the subsequent time step.

Demonstration

The LSTM and GRU models were developed in Python 3.6.5 on 64-bit Windows 10 operating system utilising the Keras 2.0 API and Tensorflow backend. The data is preprocessed by removing numbers, special characters, replacing upper case to lower case and tokenized the preprocessed text. The HS is identified using the LSTM and GRU model, with the dataset split into 25% testing and 75% training sets. To pass the input, embedding is used. By guaranteeing that the model experiences the least amount of loss during training, the sigmoid optimizer enhances performance. The model was tested using pooling windows and containing maps of various sizes. The size of the pooling window that produced the greatest results was 5. An optimal dropout value for the model was 0.1. Additional testing on the model was done with a batch size of 64 and 10 epochs.

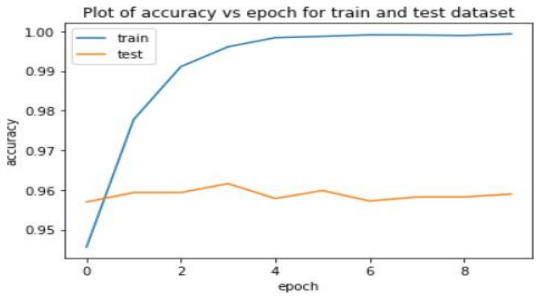

Fig. 4. GRU model's accuracy graph in relation to the number of epochs

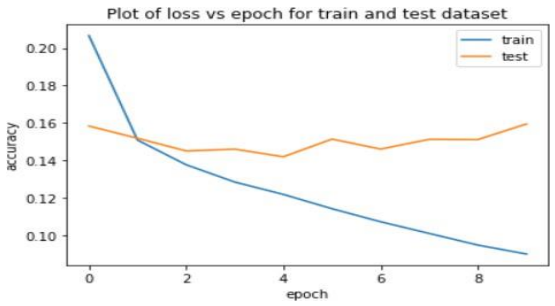

Fig. 5. GRU model's loss graph in relation to the number of epochs

C. *Performance Evaluation Metrics*

A confusion matrix is a tool for assessing how well a classification system is performing.

- True positive (TP): Both the real texts and the predicted texts are non-hate speech.
- True Negative (TN): Both the real texts and the predicted texts are hate speech.
- False Positive (FP): Real texts are hate speech but the predicted texts are non-hate speech.
- False Negative (FN): Real texts are non-hate speech but the predicted texts are hate speech.

TABLE I. CONFUSION MATRIX

	Normal	**Attack**
Normal	TP	FN
Attack	FP	TN

The quality of the deep learning algorithms can be scaled using metrics such as accuracy, precision, recall and F_1-Score etc. Some of the metrics are discussed below:

Accuracy

Accuracy means the amount of correctly categorized cases.

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn} \quad (6)$$

Precision

Precision means the true positive to the sum of predicted positive.

$$\text{Precision} = \frac{tp}{tp + fp} \quad (7)$$

Recall

Recall refers to the true positive to the sum of actual positive.

$$\text{Recall} = \frac{tp}{tp + fn} \quad (8)$$

F_1-score

The harmonic mean of recall and precision is known as the F1-score.

$$F_1 \text{ - score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (9)$$

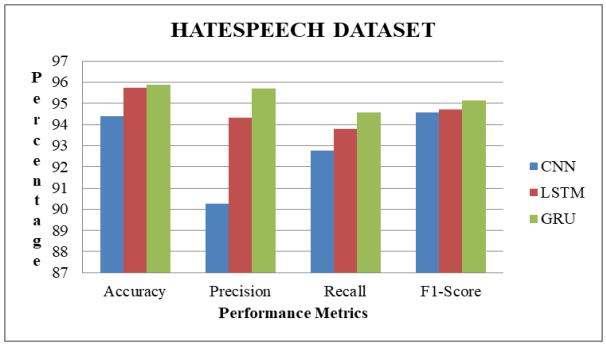

Fig. 6. Performance Metrics of three models using hatespeech dataset

VI. CONCLUSION

A Deep Convolutional Neural Network (DCNN), Long Short Term Memory (LSTM), and Gated Recurrent Unit (GRU) were applied to detect HS associated messages on Twitter with the characteristics collected using TF-IDF approach. This paper discusses the experiment on hate speech recognition on Twitter. The accuracy achieved by the Gated Recurrent Unit model, which separates hate speech text from tweets' semantics, was 95.89%.

In future, grid search optimization can be used for choosing values of optimizing hyperparameters like batch size, dropout, pooling size to achieve better performance. Hatespeech detection can further applied in different datasets in other social media boundaries such as facebook, instagram etc. and can be used in other languages.

REFERENCES

[1] https://en.wikipedia.org/wiki/Cyberbullying
[2] T. K. Das, D. P. Acharjya, and M. R. Patra, ``Opinion mining about a product by analyzing public tweets in Twitter," in Proc. Int. Conf. Comput. Commun. Informat., Jan. 2014, pp. 1_4.
[3] 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.
[4] O. Oriola and E. Kotze, ``Evaluating machine learning techniques for detecting offensive and hate speech in South African tweets," IEEE Access, vol. 8, pp. 21496_21509, 2020.
[5] K.Weller, A. Bruns, J. Burgess, M. Mahrt, and C. Puschmann, Twitter and Society, vol. 89, P. Lang, Ed. Cham, Switzerland: Peter Lang Publishing, 2014.
[6] F. Del Vigna12, A. Cimino23, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, ``Hate me, hate me not: Hate speech detection on Facebook," in Proc. 1st Italian Conf. Cybersecur. (ITASEC), 2017, pp. 86_95.
[7] C. Stokel-Walker, ``Alt-right's `Twitter' is hate-speech hub," New Scientist, vol. 237, no. 3167, p. 15, Mar. 2018.
[8] P. Charitidis, S. Doropoulos, S. Vologiannidis, I. Papastergiou, and S. Karakeva, ``Towards countering hate speech against journalists on social media," Online Social Netw. Media, vol. 17, pp. 1_10, May 2020.
[9] V. Balakrishnan et al., "Cyberbullying detection on twitter using Big Five and Dark Triad features", Personality and Individual Differences 141 (2019) pp.252–257.
[10] Kumari K et al., "Identification of cyberbullying on multi-modal social media posts using genetic algorithm", Trans Emerging Tel Tech. 2020; e3907.
[11] S. Sadiq et al., "Aggression detection through deep neural model on Twitter", Future Generation Computer Systems 114 (2021), pp.120–129.

[12] B.A.H. Murshed et al., "DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform", IEEE Access Journal, Vol. 10, 2022, pp. 25857-71.
[13] https://en.wikipedia.org/wiki/Convolutional_neural_network
[14] https://en.wikipedia.org/wiki/Long_short-term_memory
[15] https://en.wikipedia.org/wiki/Gated_recurrent_unit
[16] Ashwin Valliappan S, Ramya G R, "Identifying Fake Reviews in Relation with Property and Political Data Using Deep Learning", Procedia Computer Science 218 (2023) 1742–1751.
[17] S. Abarna, J.I. Sheeba and S. Pradeep Devaneyan, "An ensemble model for idioms and literal text classification using knowledge-enabled BERT in deep learning", Measurement: Sensors 24 (2022) 100434.
[18] Ashwin Valliappan S, Ramya G R, "Identifying Fake Reviews in Relation with Property and Political Data Using Deep Learning", Procedia Computer Science 218 (2023) 1742–1751.
[19] S. Abarna, J.I. Sheeba and S. Pradeep Devaneyan, "An ensemble model for idioms and literal text classification using knowledge-enabled BERT in deep learning", Measurement: Sensors 24 (2022) 100434.
[20] Akshi Kumar and Nitin Sachdeva, "A Bi-GRU with attention and CapsNet hybrid model for cyberbullying detection on social media", World Wide Web, Springer Volume 25, pages 1537–1550 (2022).