English Offensive Text Detection Using CNN Based Bi-GRU Model

Conference Paper · October 2024

DOI: 10.1109/ICICT64387.2024.10839709

6 authors, including:

Tonmoy Roy
Utah State University
10 PUBLICATIONS **7** CITATIONS
SEE PROFILE

Md. Robiul Islam
Rajshahi University of Engineering and Technology
24 PUBLICATIONS **748** CITATIONS
SEE PROFILE

Asif Ahammad Miazee
Maharishi International University
11 PUBLICATIONS **186** CITATIONS
SEE PROFILE

Anika Antara
BRAC University
2 PUBLICATIONS **6** CITATIONS
SEE PROFILE

2024 2nd International Conference on Information and Communication Technology (ICICT)
October 21-22, Dhaka, Bangladesh

English offensive text detection using CNN based Bi-GRU model

Tonmoy Roy[1]
Data Analytics & Information Systems
Utah State University
Utah, United States
tonmoyroy1992@gmail.com

Md Robiul Islam[2]
Computer Science
William & Mary
Virginia, United States
robiul.cse.uu@gmail.com

Asif Ahmed Miazi[3]
Computer Science
Maharishi International University
Iowa, United States
asifahammad7@gmail.com

Anika Antara[4]
Electrical and Electronics Engineering
Brac University
Dhaka, Bangladesh
anikaantara1000@gmail.com

Al Amin[5]
Computer Science
Uttara University
Dhaka, Bangladesh
alaminbhuyan321@gmail.com

Sunjim Hossain[6]
Computer Science
Northern University
Dhaka, Bangladesh
hossainsunjim@gmail.com

***Abstract*—Over the years, the number of users of social media has increased drastically. People frequently share their thoughts through social platforms, and this leads to an increase in hate content. In this virtual community, individuals share their views, express their feelings, and post photos, videos, blogs, and more. Social networking sites like Facebook and Twitter provide platforms to share vast amounts of content with a single click. However, these platforms do not impose restrictions on the uploaded content, which may include abusive language and explicit images unsuitable for social media. To resolve this issue, a new idea must be implemented to divide the inappropriate content. Numerous studies have been done to automate the process. In this paper, we propose a new Bi-GRU-CNN model to classify whether the text is offensive or not. The combination of the Bi-GRU and CNN models outperforms the existing models.**

***Index Terms*—hate content, social media, CNN, Bi-GRU**

I. Introduction

Hate speech is a distinct form of language that involves abusive behaviour. The targets of this phenomenon are specifically selected based on their personal qualities or demographic background, including race, ethnicity, religion, colour, sexual orientation, or other comparable criteria [1]. A multitude of academics are diligently focused on addressing the issue of hate speech detection using natural language processing techniques. They are developing practical frameworks and creating automatic classifiers that rely on supervised machine learning models [2]. Sharing inappropriate content on social media platforms like Twitter and Facebook has become increasingly effortless, often targeting specific individuals or groups. Additionally, toxic language can manifest in various forms, including cyberbullying, which has played a key role in contributing to suicide. [3]. According to the United Nations strategy and plan of action on hate speech, there is no internationally agreed-upon legal definition, However, it involves incitement, which means intentionally encouraging discrimination, hostility, and violence [4].

Similarly, offensive speech refers to writing that includes abusive slurs or degrading terms [5], which are often mistaken for hate speech in various situations. Offensive language and hate speech detection are two specialised areas of study within the subject of natural language processing. The primary obstacle is in the fact that the majority of unsuitable content found online is presented in the form of natural language text. Consequently, it is imperative to develop efficient tools capable of extracting and analysing this content from unorganised textual data.

These technologies often utilise methods that are based on natural language processing (NLP), retrieval of data, machine learning, and deep learning. Research conducted by [6] focuses on safeguarding children from inappropriate content in mobile applications. The study suggests techniques for parents to assess the maturity level of smartphone apps, allowing them to select apps that are appropriate for their children's ages. Unfortunately, it is not practical to manually filter out harmful content on a large scale, and manually identifying or removing anything from the internet is a laborious undertaking. This serves as a driving force for researchers to develop automated techniques that can identify offensive information on social media platforms.

This study presents comprehensive experiments aimed at resolving the problem of categorising improper content through the utilisation of machine learning and neural network models. The methodology we employ involves optimising hyperparameters and utilising word embedding features on a dataset in the English language. Our main contribution in this paper as follows:

- We proposed a fine-tuned 1D Convolutional Neural Network (CNN) with Bi-GRU based offensive text detection model, which outperforms other benchmark models.
- The dataset contains more than 31 thousands tweets which is a big data to do work.

979-8-3315-0822-7/24/$31.00

The paper is organized as follows: In section II, we review the relevant literature on classifying offensive speech. Section III details our proposed methodology. Section IV covers our experiments and results. In Section V, we discuss the challenges encountered during implementation. Finally, in section VI, we conclude with a summary and outline future work.

II. Related Works

Social life has become part of life nowadays. According to the source, on average, a user spends 3 hours and 15 minutes on their phone each day, and that number increases if there is a holiday [7]. Individuals check their phones 58 times within 24 hours. Over the past few years, there has been a rapid development of social media, resulting in both benefits and drawbacks in its utilisation. Hate speech commonly targets specific races, religions, sexual orientations, genders, and castes, among other categories. This problem has prompted numerous academics to investigate an approach capable of identifying hate speech that is targeted towards specific individuals or groups [8].

Machine learning techniques for natural language processing (NLP) have traditionally relied on shallow models like Support Vector Machines (SVM) and Logistic Regression (LR). These models are trained using features that are both high-dimensional and sparse. Most techniques primarily concentrate on extracting textual aspects. Some scholars have employed lexical features, such as dictionaries [9] and bag-of-words [10]. Kumari et. al. propose a hybrid model for identifying aggressive posts that include both images and text on social media platforms [11]. In contrast, Kovacs et. al. employ various machine learning techniques and deep learning models to automatically detect speech that is both hateful and offensive [12]. Although machine learning (ML) has encountered problems, there have been significant advancements in developing ML-based systems for automatic detection, which have yielded promising outcomes. Some noteworthy examples of classifiers are the Naive Bayes (NB) [10], Logistic Regression (LR) [13], and Support Vector Machines (SVM) [13]. Offensive speech detection is not only used in English; it is also used in Hindi as well [14]. Also in Spanish and Italian [15]. Many researchers work on several languages to detect offensive text [16]. Corazza et al. [17] employ datasets in three distinct languages (English, Italian, and German) and train various models including LSTMs, GRUs, Bidirectional LSTMs, and others. Huang et al. [18] created a multilingual Twitter hate speech dataset by combining data from 5 languages. They also added demographic information to investigate the presence of demographic bias in hate speech classification. Aluru et al. [19] employed datasets from 8 different languages and produced embeddings using LASER [20] and MUSE [21]. These embeddings were inputted into various architectures, such as CNNs, GRUs, and different Transformer models. While the study achieved satisfactory performance across these languages, it had two main shortcomings. Firstly, the claim that the models are generalizable is questionable, as they did not leverage datasets from multiple other languages and lacked fine-tuning of parameters for better generalization. Consequently, these models are unlikely to perform well outside of the 8 languages studied. Secondly, the research focused primarily on models suited for low-resource settings without addressing their performance in resource-abundant environments. By aggregating datasets from 11 different languages, his study achieved superior performance [22].

III. Methodology

In this section, we present our proposed methodology for offensive text classification. Our proposed Bi-GRU and CNN models take input text and output the probability of it being in an inappropriate class. Our model consists of nine layers (a) Input Layer, (b) Embedding Layer, (c) Convolutional Layer(Conv1D), (d) MaxPooling Layer, (e) Bi-directional GRU_1(Bi-GRU), (f) Bi-directional GRU_2(Bi-GRU), (g) Dense Layer, (h) Dropout Layer, (i) Dense Layer.

A. Dataset Descriptions

In this subsection, we present our dataset that we have used in our experiments. The dataset contains 31,962 tweets. This dataset has been split into a training set and a test set. The ratio is 80:20 for training and testing respectively.

TABLE 1 Dataset

Classes	**Total Tweet**
Offensive	29720
Not offensive	2242

B. Pre-processing

In order to enhance the efficiency of our approach, it is important to carry out pre-processing steps to cleanse our textual data. Initially, the tweet underwent a process where all numerical values, punctuation marks, URLs (starting with http:// or www.), and symbols (like emojis, hashtags, and mentions) were eliminated. This was done because these elements do not contribute to the sentiment-related content of the tweet. The tweets were first broken down into individual words or phrases, a process called tokenization. This was done using a tool from the NLTK library. Then, all common words with little meaning on their own (like ”the”, ”a”, ”is”) were removed from the tokens. These common words are also provided by the NLTK library.

C. Bi-directional GRU (Bi-GRU)

Regular GRU models only consider information from previous words in a sequence. But to truly grasp the meaning of a word, it's important to also understand what comes after it. That's where Bidirectional GRUs (BiGRUs) come in. Our system uses BiGRUs, which are essentially two GRUs working together. One reads the text forward, the other backward. This allows the BiGRU to capture important details from the text and analyze each word in the context of both its past and future neighbors. This gives BiGRUs an edge over traditional

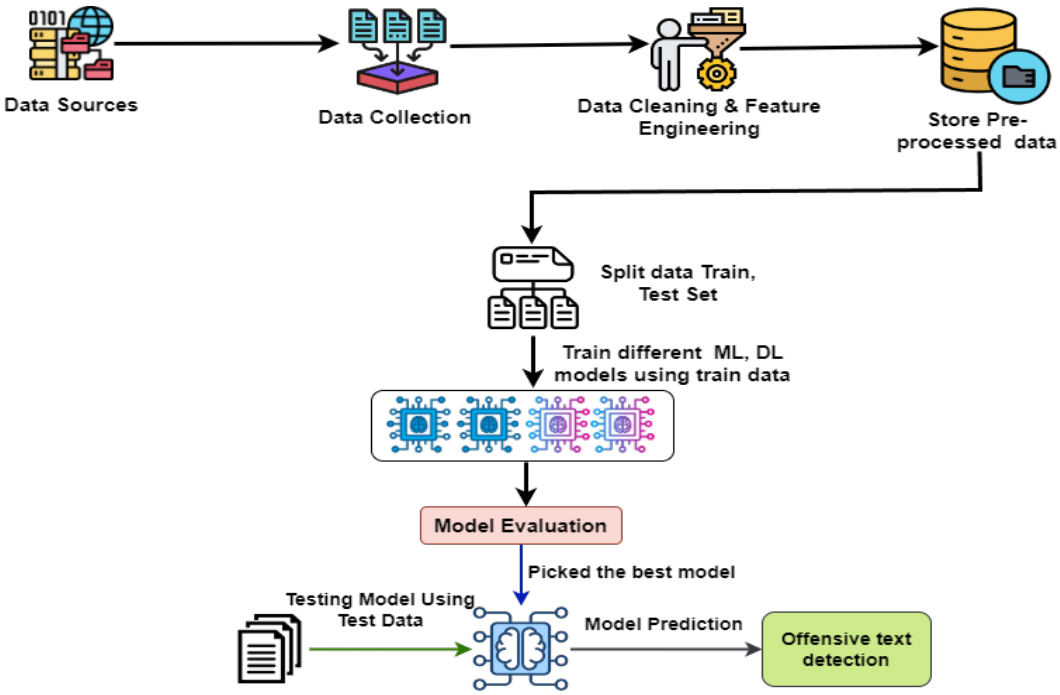

Figure 1: Proposed methodology

machine learning models when it comes to understanding the nuances of language and detecting hate speech.

D. Experimental setup

For these studies, we utilise Keras and Tensorflow as the underlying frameworks. We utilise Google Colab to execute the tests with the assistance of the Graphical Processing Unit (GPU). Regarding the training process, we employ a total of 100 epochs. Several traditional machine learning techniques were employed, including Naive Bayes, Support Vector Machine, Logistic Regression, Random Forest, and Adaboost. Furthermore, alongside these models, various techniques for word representation were employed, including count vectors as features and term frequency inverse document frequency (TF-IDF).

E. Evaluation Metrix

In this section, we'll first explain the dataset we used to evaluate our methods and compare them against existing models. Then, we'll detail the specific settings used in our experiments.Finally, we will showcase the performance results, which were assessed using metrics such as accuracy, F1-score, recall and precision.

Accuracy: is one crucial factor in evaluating the classifier's performance. This metric indicates the proportion of correct classifications among all predictions made by the classifier. Accuracy calculation is defined in Equation-1.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision: can be described as the count of accurate positive outputs generated by the model. The precision calculation is shown in Equation-2.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Recall: measures the effectiveness of a model in recognizing positive instances among all the genuine positive instances within the dataset. Recall calculation is defined in Equation-3.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

F1-Score: assists in assessing both recall and precision concurrently. The F1-score reaches its peak when recall matches precision. It can be calculated using Equation-4.

$$\text{F1-Score} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

IV. Experimentations and Results

In this section, we discuss the model's performance. We discuss about accuracy, precision and recall.

A. Results

Every model undergoes training using the training dataset, and its performance is assessed using standard classification

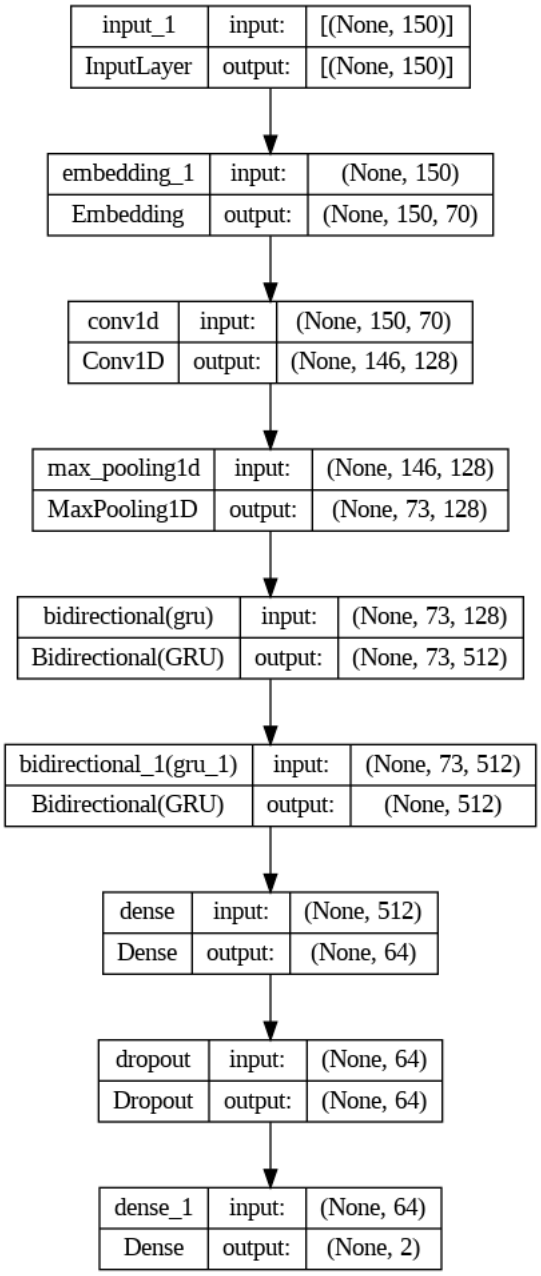

Figure 2: Model Flow Chart

metrics: Accuracy (Acc), F1-score (F1), Recall (R), and Precision (P). In order to improve the outcomes, different hyperparameters were experimented with and integrated [23]. This paper provides a comprehensive overview of all the findings, with a particular emphasis on the models that performed the best. Multiple preliminary evaluations were conducted prior to submitting the final results.

Fig. 3 shows the training and validation accuracy of the proposed model. Fig. 4 depicts the suggested model loss graph for 100 epochs on training and validation stage. Beside that Fig. 5 and Fig. 6 represents the training and validation recall and AUC graph respectively. Table 2 shows the evaluation results of different machine learning models and Table 3 shows the comparative analysis of some other existing research.

Figure 3: Training and Validation Accuracy

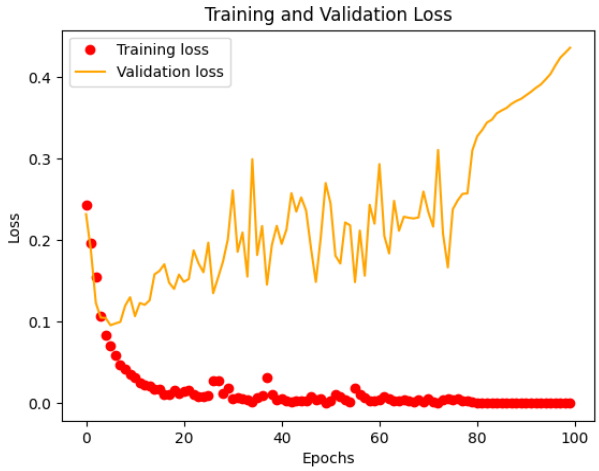

Figure 4: Training and Validation Loss per epoch

V. Conclusion and Future Work

This research describes a system that combines Convolutional Neural Networks (CNN) and Bidirectional Gated Recurrent Units (Bi-GRU). We were motivated to create contextual embeddings by the use of social networks, namely by utilising a Twitter dataset. We use the acquired information from this language model designed to detect offensive language and expressions of hate in written content. We conducted an assessment of various supervised machine learning classifiers to detect nasty and abusive content on Twitter, using a dataset specifically collected from Twitter. Deep learning enables the automatic acquisition of multi-level feature representations, while typical machine learning-based NLP systems rely extensively on manually engineered features. Creating these

Figure 5: Training and Validation Recall

Figure 6: Training and Validation AUC

artisanal elements requires a significant amount of time and effort and may not always be fully developed. The deep learning models achieved outstanding results in reducing false positives and showed promising outcomes in decreasing false negatives during the classification process. Our objective is to do more investigation into the application of deep neural network architectures for the purpose of detecting hate speech. We will conduct our inquiry by employing a range of word embedding approaches. In addition, we intend to broaden our research by incorporating additional datasets, particularly those in the Bangla language [24].

TABLE 2 Model Evaluation table

No.	Algorithm	Accuracy	Precision	Rcall	F1-Score
1	Logistic Regression	96.86	86.50	51.27	64.21
2	RandomForest Classifier	96.56	84.33	54.26	67.56
3	MultinomialNB	95.97	95.65	95.97	95.47
4	KNeighbors Classifier	95.97	81.41	81.41	81.41
5	Linear SupportVectorClassification	94.90	98.46	82.05	89.51

TABLE 3 Research comparison with others

No.	Authors	Algorithms	FE method	Accuracy
1	Vashistha et. al. [14]	LR	Word embedding	76.90%
2	Aluru et. al. [19]	LR	MUSE and LASER	76.90%
3	Deshpande et. al. [16]	TL	mBERT	76.90%
4	Proposed model	Bi-GRU and CNN	CountVectorizer	96.97%

REFERENCES

[1] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proceedings of the 25th international conference on world wide web*, pp. 145–153, 2016.

[2] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Computing Surveys (CSUR)*, vol. 51, no. 4, pp. 1–30, 2018.

[3] S. Hinduja and J. W. Patchin, "Bullying, cyberbullying, and suicide," *Archives of suicide research*, vol. 14, no. 3, pp. 206–221, 2010.

[4] "United nations." [Online]. Available: https://www.un.org/en/genocideprevention/documents/ [Accessed: 14/05/2024].

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

[6] B. Hu, B. Liu, N. Z. Gong, D. Kong, and H. Jin, "Protecting your children from inappropriate content in mobile apps: An automatic maturity rating framework," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pp. 1111–1120, 2015.

[7] "Time spend on smartphones." [Online]. Available: https:https://explodingtopics.com/blog/smartphone-usage-stats [Accessed: 15/05/2024].

[8] M. O. Ibrohim and I. Budi, "Hate speech and abusive language detection in indonesian social media: Progress and challenges," *Heliyon*, 2023.

[9] E. M. Alshari, A. Azman, S. Doraisamy, N. Mustapha, and M. Alkeshr, "Effective method for sentiment lexical dictionary enrichment based on word2vec for sentiment analysis," in *2018 fourth international conference on information retrieval and knowledge management (CAMP)*, pp. 1–5, IEEE, 2018.

[10] Y. Pandey, M. Sharma, M. K. Siddiqui, and S. S. Yadav, "Hate speech detection model using bag of words and naïve bayes," in *Advances in Data and Information Sciences: Proceedings of ICDIS 2021*, pp. 457–470, Springer, 2022.

[11] K. Kumari, J. P. Singh, Y. K. Dwivedi, and N. P. Rana, "Multi-modal aggression identification using convolutional neural network and binary particle swarm optimization," *Future Generation Computer Systems*, vol. 118, pp. 187–197, 2021.

[12] G. Kovács, P. Alonso, and R. Saini, "Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources," *SN Computer Science*, vol. 2, no. 2, p. 95, 2021.

[13] O. Oriola and E. Kotzé, "Evaluating machine learning techniques for detecting offensive and hate speech in south african tweets," *IEEE Access*, vol. 8, pp. 21496–21509, 2020.

[14] N. Vashistha and A. Zubiaga, "Online multilingual hate speech detection: experimenting with hindi and english social media," *Information*, vol. 12, no. 1, p. 5, 2020.

[15] A. Jiang and A. Zubiaga, "Cross-lingual capsule network for hate speech detection in social media," in *Proceedings of the 32nd ACM conference on hypertext and social media*, pp. 217–223, 2021.

[16] N. Deshpande, N. Farris, and V. Kumar, "Highly generalizable

models for multilingual hate speech detection," *arXiv preprint arXiv:2201.11294*, 2022.
[17] M. Corazza, S. Menini, E. Cabrio, S. Tonelli, and S. Villata, "A multilingual evaluation for online hate speech detection," *ACM Transactions on Internet Technology (TOIT)*, vol. 20, no. 2, pp. 1–22, 2020.
[18] X. Huang, L. Xing, F. Dernoncourt, and M. J. Paul, "Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition," *arXiv preprint arXiv:2002.10361*, 2020.
[19] S. S. Aluru, B. Mathew, P. Saha, and A. Mukherjee, "Deep learning models for multilingual hate speech detection," *arXiv preprint arXiv:2004.06465*, 2020.
[20] "Facebook research:laser: Language-agnostic sentence representations." [Online]. Available: https://github.com/facebookresearch/LASER [Accessed: 18/05/2024].
[21] "Facebook research:muse: Multilingual unsupervised and supervised embeddings." [Online]. Available: https://github.com/facebookresearch/MUSE [Accessed: 18/05/2024].
[22] M. R. Islam, A. Amin, and A. N. Zereen, "Enhancing bangla language next word prediction and sentence completion through extended rnn with bi-lstm model on n-gram language," *arXiv preprint arXiv:2405.01873*, 2024.
[23] M. R. Islam, M. M. Islam, M. Afrin, A. Antara, N. Tabassum, A. Amin, *et al.*, "Phishguard: A convolutional neural network based model for detecting phishing urls with explainability analysis," *arXiv preprint arXiv:2404.17960*, 2024.
[24] A. Amin, A. Sarkar, M. M. Islam, A. A. Miazee, M. R. Islam, and M. M. Hoque, "Sentiment polarity analysis of bangla food reviews using machine and deep learning algorithms," *arXiv preprint arXiv:2405.06667*, 2024.