Bangla Social Media Cyberbullying Detection Using Deep Learning

Anika Tasnim Rodela[1], Huu-Hoa Nguyen[2], Dewan Md. Farid[3](✉), and Mohammad Nurul Huda[3]

[1] Department of Information and Communication Technology, Bangladesh University of Professionals, Mirpur Cantonment, Dhaka, Bangladesh
[2] College of Information and Communication Technology, Can Tho University, 3/2 Street, Ninh Kieu District, Can Tho City, Vietnam
[3] Department of Computer Science and Engineering, United International University, United City, Madani Avenue, Badda, Dhaka 1212, Bangladesh
{anika,dewanfarid,mnh}@cse.uiu.ac.bd, nhhoa@ctu.edu.vn
https://cse.uiu.ac.bd/profiles/dewanfarid/

Abstract. The growth of social media over the past decade has been nothing short of phenomenal. An exponential increase has been seen on platforms like Facebook, Twitter, Instagram, LinkedIn, and YouTube, in their user base, accumulating billions of active users worldwide. Technology advancements, the ubiquitous use of smartphones, and the innate human desire for connection don't always contribute constructively; they might additionally end up in spreading violence in the form of bullying of others which is known as cyberbullying. As a result of cyberbullying, the victims will frequently experience anxiety, depression, and other mental diseases which can even result in suicide. Therefore the extensive need of detecting and controlling cyberbullying motivated us to automate this process. In this paper, we have introduced an approach to detect cyberbullying from social media data by using deep learning models. We have used Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU), and Convolutional Neural Network in the proposed hybrid approach along with word embedding technique called fastText. We achieved an accuracy of 91.63% with the proposed model by using a publicly available dataset containing 16,073 samples which outperformed all the state of the art models.

Keywords: Cyberbullying · Deep Learning · Embedding

1 Introduction

In the modern age of technology, cyberbullying, a widespread form of online harassment, has emerged as an alarming issue. It describes the practice of using social media platforms or any electronic media to harass, bully, or threaten others. Cyberbullying can have serious negative effects on the victims, including psychological harm, depression, anxiety, sometimes leading them to self-harm,

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024
N. Thai-Nghe et al. (Eds.): ISDS 2023, CCIS 1949, pp. 170–184, 2024.
https://doi.org/10.1007/978-981-99-7649-2_13

and even catastrophic outcomes like suicide. From 1997 people were using social media and now there are 4.26 billion users worldwide, which is more than half of the total world population [16]. Cyberbullying is continuously being proliferated with the growth of social media, necessitating the development of efficient detection and prevention techniques. According to a report in [31], 73% of students stated that they were bullied at least once in their lifetime and about 44% faced the same in the last 30 days. These concerning statistics highlight the pressing need for proactive measures to stop cyberbullying and defend vulnerable individuals.

While cyberbullying exists in all languages, there are a lot of bullying occurrences in the native languages as well. Bangla or Bengali is the sixth most spoken language in the world and currently 234 million native speakers are using it [21]. Bangladesh is a developing country and almost all the people of the country speak Bangla as their first language. With an annual growth rate of 10.1% in 2021–2022, Bangladesh had 50.3 million active social media users on a monthly basis as mentioned in [1]. Especially since the COVID 19 pandemic people are even more indulged in social media and this leverages different bullying situations or occurrences. Hence it has also become mandatory to introduce proper techniques to detect such cyberbullying occurrences in the native language like Bangla. It somewhat requires more strenuous efforts to detect cyberbullying in a native language such as Bangla rather than international languages such as English because the preprocessing, stop word removal, using proper stemmer etc. are different and more challenging for Bangla for the lack of proper resources. According to Kumar and Sachdeva [20] the linguistic challenges are also contributed by the cultural diversity, using hash-tags on trending topics which is often region specific, unconventional use of typographical resources including capital letters, punctuation, and emojis, as well as the accessibility of keyboards in one's native language. All this information points to the fact that detection of such harassment or bullying in the native language is demanding and at the same time more challenging.

Automatic detection and classification of cyberbullying from social media data is a task that requires both natural language understanding and generic text classification [20]. Recently deep learning (DL) models are being considered by researchers as an attempt to perform automated text classification and natural language processing. In particular, recurrent neural networks (RNNs) such as long short-term memory (LSTM) networks have been utilised to analyse the sequential connections or patterns and temporal dynamics in social media posts. To identify instances of cyberbullying, these models can effectively understand the proper context, linguistic sequences and sentiment. Convolutional neural networks (CNNs) have been used as well to extract useful information from text by taking into account the relationships and underlying structure found within sentences and phrases. The accuracy of cyberbullying detection systems has been improved by combining these deep learning models with approaches like word embeddings and attention mechanisms.

Machine learning and deep learning are being used frequently in classification tasks such as cyberbullying. There are several works on this topic, such as, toxic comments detection using supervised learning [6], abusive Bangla comments detection by using transformer-based models [5], cyberbullying detection from deep neural networks [3], hate speech detection [2], Classification Benchmarks based on Multichannel Convolutional-LSTM Network [19], Multi Labeled Bengali Toxic Comments Classification [7] etc. In spite of existing a good number of research works, most of them lack in some areas. Most of the research works lacked in introducing datasets with proper amount of data and the datasets being unbalanced. Moreover, some of them used very complex models which degraded the efficiency of the system. Addressing the issues, we have conducted our work on a large dataset of 16073 samples and we proposed a model which is comparatively easily and lightly implemented and resulted in much higher accuracy.

The rapid recognition of cyberbullying instances is essential in order to lessen its adverse effects. However, it is challenging and time-consuming to manually monitor and identify cyberbullying instances from the large volume of social media data. Hence, the development of automated processes utilising modern technology like deep learning has emerged as a possible remedy. In this regard we have performed classification of cyberbullying on a dataset containing 16073 instances where 8488 instances were cyberbullying and the other 7585 were non cyberbullying instances. We have used a hybrid approach of LSTM-GRU-CNN and corporate with three different embeddings: (1) fastText, (2) GloVe, and (3) Word2Vec. The proposed model achieved an accuracy of 91.63% which outperformned all the other models from our study and the recent studies.

The following parts of the paper go as follows: Sect. 2 represents the Literature Review stating and elaborating a comparative study on all the state of the art models. Section 3 discusses about the utilisation of deep learning, the dataset's description and our proposed model in details. Section 4 is the summary of the experimental results and Sect. 5 concludes the whole study.

2 Literature Review

There are several research works conducted related to this work of cyberbullying detection and classification. Various research works are focused on various factors such as, monolingual or multilingual content, binary or multi-class classification and machine learning or deep learning or hybrid models.

2.1 Monolingual and Multilingual

Based on the types of lingual contents, we can divide the research works into two categories such as, monolingual and multilingual. There have been many studies conducted on monolingual content of natural language processing such as, English language for hate speech or cyberbullying detection or any text classification tasks [17,23,25,32]. Other than English, Ptaszynski et al. conducted cyberbullying detection in Polish language [28], Gordeev [13] reported verbal aggression detection in Russian and English, Ibrohim et al. [18] and Pratiwi et al. [27] conducted hate speech and abusive language detection from Indonesian tweets. For multilingual contents, there are diverse existing research works as well. Haider et al. [14,15] introduced a multilingual cyberbullying detection system using machine learning and deep learning techniques and they performed the validation on Facebook and Twitter data in Arabic language. Another multilingual cyberbullying detection system was proposed by Pawar et al. [26] in two Indian languages namely: Hindi and Marathi. Arreerard et al. [4] proposed a machine learning model for classification of defamatory Facebook comments in the Thai language. Kumar and Sachdeva [20] developed a similar system to detect cyberbullying in Hindi-English code-mixing data using deep neural networks and transfer learning for cyberbullying detection.

2.2 Binary-Multilabel and Machine Learning-Deep Learning Classification

All the researchers in this domain formulated cyberbullying detection as a text classification problem. Based on the type of classification, in most papers we have found mostly binary classification while some incorporated multi-class classification as well. Chakraborty and Seddiqui [9] detected threat and abusive language in Bangla by using both machine learning and deep learning models while formulating the problem as a binary classification problem. They achieved an accuracy of 78% when using SVM with linear kernel. Malik et al. [24]performed toxic speech detection using machine learning and deep learning models along with word embedding techniques such as, BERT and fasttext. For machine learning models they used Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, XGBoost and for deep learning models they used Convolutional Neural Network (CNN), Multi Layer Perceptron, Long Short Term Memory (LSTM) They achieved 82% accuracy for binary classification using CNN. To categorise Bengali comments as toxic and non-toxic, Banik et al. [6] employed both supervised machine learning models such Naive Bayes, Support Vector Machine, Logistic Regression, and deep learning models like Convolutional Neural Network, Long Short Term Memory. For binary classification, Convolutional Neural Networks had the highest accuracy (95.30%). They utilised word2vec as a word embedding technique.

Das et al. [10] to detect Bengali hate speech employed Convolutional Neural Networks with Long Short Term Memory (CNN-LSTM), Convolutional Neural Networks with Gated Recurrent Unit (CNN-GRU), and attention-based Convolutional Neural Network while achieving 77% accuracy. To detect cyberbullying Ahmed et al. [3] used Convolutional Neural Network with Long Short Term Memory (CNN-LSTM) and gained 87.91% accuracy for binary classification and 85% accuracy for multi-class classification. Belal et al. [7] categorised Bangla toxic comments using a deep learning based pipeline. They used BERT as the word embedding technique and performed binary classification using Long

Short Term Memory (LSTM) achieving 89.42% accuracy. For multi-label classification, the authors used a combination of Convolutional Neural Network and Bi-directional Long Short Term Memory (CNNBiLSTM) with attention mechanism and achieved 78.92% accuracy. Using Multichannel Convolutional-Long Short Term Memory (MConv-LSTM) with BangFastText, Word2Vec, and GloVe for word embedding, Karim et al. [19] identified Bengali hate speech. Their model received a 92.30% F1-score using BangFastText embedding.

Based on the models used, we can divide the research works into three different categories such as, models based on machine learning algorithms, deep learning algorithms and a hybrid model where simultaneously multiple approaches are used. The authors in [8,9,29,30] used machine learning based approaches to detect cyberbullying or hate comments. Deep learning algorithms such as Convolutional Neural Networks (CNNs), Long Short Term Memory Networks (LSTMs), Recurrent Neural Networks (RNNs) have been used by many researchers [7,9,24]. Many researchers are also exploring hybrid models to get better performance in this text classification problem. The authors in [6,7,12,19] used Convolutional Neural Networks with Long Short Term Memory (CNN-LSTM), Convolutional Neural Networks with Gated Recurrent Unit (CNN-GRU), Convolutional Neural Network and Bi-directional Long Short Term Memory (CNNBiLSTM) with attention mechanism to build their models for this classification problem of cyberbullying detection.

3 Learning with Deep Learning

Deep learning has transformed artificial intelligence and data analysis through delivering capabilities that were unattainable for comprehending and analysing complex data. Deep learning algorithms have demonstrated outstanding results in applications such as image and speech recognition, natural language processing, and predictive analytics by applying artificial neural networks with multiple layers. Deep learning techniques enable machines to learn from enormous datasets and make intelligent choices by mimicking the complex structure and activity of the human brain. These algorithms utilise the strength of neural networks, which are constructed from interconnected "neurons" or nodes arranged in layers. Deep learning algorithms can derive increasingly abstract representations of data through multiple layers, enabling advanced and deeper understanding.

Most recently Natural language processing (NLP) tasks such as text classification activities have been significantly influenced by deep learning, evolving the field of study and resulting in advancements in numerous areas of language understanding, development, and analysis. Recurrent neural networks (RNNs) including Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), have emerged as prevalent deep learning models in natural language processing. These architectural frameworks are made to manage sequential data and recognise dependencies in text composed of natural language. For text categorisation problems, deep learning models like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) (including LSTM and GRU), and

Transformer-based models have been extensively used currently. Deep learning based word embedding techniques are largely used as well. Deep learning models often leverage word embeddings like Word2Vec or GloVe to represent words as dense vectors in a continuous space. The model can learn accurate representations of words due to these embeddings, which recognise semantic connections as well as contextual information. For sequential data analysis, Recurrent Neural Network (RNN) is widely used. Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) are the variants of RNN.

3.1 Long Short Term Memory (LSTM)

In sequential data to understand the context and long term dependencies LSTM [11] is largely applied. It addresses the vanishing gradient problem which allows it to learn better of long term dependencies in sequential data. Three gating mechanisms-the input gate, forget gate, and output gate-as well as memory cells are introduced by LSTM. By controlling the information flow, these gates allow the model to selectively recall or forget previous knowledge according to the present context. Due to this feature, LSTM performs particularly well in tasks involving sequential data, including speech recognition, sentiment analysis, and machine translation.

3.2 Gated Recurrent Unit (GRU)

GRU addresses some limitations of the traditional RNNs as well such as vanishing gradients and failure to capture long term dependencies. GRU follows gating mechanisms to control the flow of information- Update gate and reset gate. Update gate is for combining new information and determining to which context the previous hidden state should be updated. Reset gate is to control forgetting or resetting the previous hidden state. GRU keeps track of a hidden state that serves as the network's memory. Using the gating mechanisms, the current input and the prior hidden state are combined to update the hidden state at each time step. It collects essential information from previous time steps and encodes it into the hidden state that is currently in effect.

3.3 Convolutional Neural Network (CNN)

CNN is a deep learning architecture that is typically employed for signal and image processing tasks. Through convolutional and pooling layers, it is supposed to automatically generate hierarchical representations of data. CNNs excelled at tasks including image classification, object identification, and image segmentation because they use convolutional filters to identify local patterns and spatial dependencies. The spatial dimensions are reduced with the help of the pooling layers, while the key features are kept intact. By considering the input as a one-dimensional signal, CNNs have been further developed to handle sequential data, such as text categorisation and speech recognition.

3.4 Proposed Method

In this research, we have used Bangla social media text for cyberbullying detection. Therefore, following the most common and standard convention, we have pre-processed the raw text. The dataset of almost 16000 instances was used for this purpose. The properly annotated and labeled dataset was set for preprocessing and for this purpose at first, tokenization was employed to extract words from the texts. This step was completed by leveraging Natural Language Toolkit tokeniser. The next step was to remove all the punctuations from the text. Then we used a stemmer to convert the tokens to the basic form of each word by using Bangla Natural Language Processing Toolkit (BNLTK) stemmer. From the stem or the base words, it is convenient to convert each sentence into a numeric form. We have used a pool of 746 Bangla words as stop words which added some values to these texts. In the next procedure, all the texts were ensured to be of the same length which is 140 characters and this was our input sequence. At this point word embedding techniques such as GloVe, fastText and Word2Vec are used in order to convert the sentences into numeric vectors. These word embedding techniques work as vectorisers. For our Bangla dataset we used fastText which is trained by Wikipedia data. We have compared the performance with GloVe and Word2Vec as well (Fig. 1).

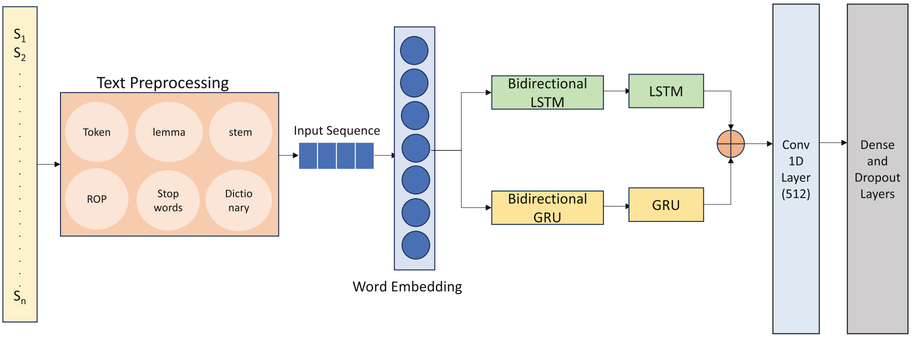

Fig. 1. Proposed model for cyberbullying detection.

In natural language processing (NLP), eliminating stop words and punctuation can alter the meaning of sentences or words, which can affect the classification tasks. Stop words like "and", "the", "is", etc. are often removed since they are common and don't have a specific or significant meaning, whereas punctuation can give significant clues about the structure and meaning of a sentence. To hold up the meaning of the sentences even after preprocessing we used lemmatization or stemming. Stemming can be used to reduce words to their root or base forms as opposed to fully eliminating them. This helps in retaining some of the semantic meaning while still reducing dimensionality. For example, "eating" and "ate" can be reduced to "eat". Again, we have used custom stop word lists; these lists are made to keep domain-specific or contextually appropriate stop words rather than eliminating all stop words. Word2Vec, GloVe, and other pre-trained word embeddings, are trained on large corpora and capture the semantic connections between words. Even when stop words and punctuation are removed, they can understand the complex meanings of words and their context. Moreover, as a result of using hybrid Approaches (combining traditional NLP techniques with neural network models), the problem can also be solved. At first texts are preprocessed by removing stop words and punctuation and then using a neural network model is used to learn to capture the context and meaning from the remaining words. Finally, we also assessed the effects of the preprocessing decisions. It's important to compare the performance of the model with and without these preprocessing procedures because sometimes eliminating stop words and punctuation may be deteriorating to the task at hand (Fig. 2).

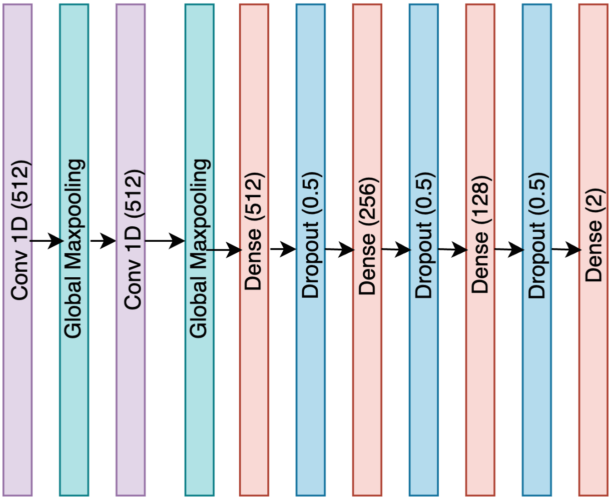

Fig. 2. CNN, Dense and Dropout Layers from methodological framework.

In our model, we have used a bidirectional LSTM layer and a bidirectional GRU layer of 200 neurons. Combining these two parallel layers, the output is fed into a conv1D layer which has 512 neurons. The combination of LSTM, GRU and CNN is used for capturing long term dependencies and extracting insightful features. After the CNN layer max pooling is used, another layer of Conv1D and maxpooling were embedded. Finally three dense layers having 512, 256 and 128 neurons were used where the dropout is 50%. ReLU activation function was used in every layer except for the last or the output layer. For the output layer Softmax activation function was used. A dense layer with two neurons was the output layer. After training for 30 epochs and setting the learning rate as 0.001 the best accuracy was achieved although at 3 epoch we achieved the highest accuracy after fine tuning our model. This proves the efficiency and lightness of our model in terms of both time and memory and other resources.

To train the model we selected 13000 instances among 16000 in the dataset by splitting the dataset in 80:20 ratio. For the test set the rest of the instances (3000 instances) were selected. The accuracy, sensitivity and positivity along with the correctly predicted values (both positive and negative) were calculated in order to evaluate the performance of the model. A heat map was used to show the distribution of the predicted results. The proposed model is compared to other state of the art models and a performance comparison is shown.

4 Experimental Results

We have evaluated the performance of the proposed model in the standard metrics as well as compared them with the state of the art models and embedding techniques.

4.1 Dataset

In this research, we have used social media text for the detection and classification of cyberbullying. For this purpose we used a publicly available dataset of Belal et al. [7] which was collected from three different sources [3,19,22] and was relabelled as the original ones were neither fully correct nor consistent. The authors labeled the data following a standard procedure with the guidance of three professional annotators. Two of them labeled each instance independently and the third expert provided his expertise to resolve any disagreements through proper discussions. Using the Cohen's kappa coefficient, we validated and evaluated the inter-annotator agreement to assess the quality of the annotations. By choosing 30 control samples and 100 randomly selected sentences that had already been expertly labeled, we were able to ensure the reliability of the annotators. Based on the evaluation, all three annotators had credibility ratings that were higher than 80%. The dataset contains 16, 073 social media texts in total of which 7, 585 are classified as non-cyberbullying and 8, 488 are classified as cyberbullying. Table 1 shows the distribution of cyberbullying instances by class.

4.2 Performance Comparison with Different Embeddings

Table 2 presents metrics and performance measures for the proposed models with three embedding techniques: fastText and Word2Vec and GloVe. These models were evaluated using our dataset, and their performance was assessed based on various metrics. The first metric, accuracy, measures the overall correctness of the models' predictions. In this case, fastText achieved an accuracy of 91.63%, higher than Word2Vec's accuracy of 90.42% and GloVe's 90.67%. These values indicate that all models were able to make accurate predictions on the given dataset. FastText achieved a sensitivity of 0.9423, much higher than Word2Vec's sensitivity of 0.9146 and GloVe's 0.9170. This suggests that fastText was slightly

Table 1. Various properties of dataset.

Properties	Values
Total instances	16073
Cyberbullying text	8488
Non cyberbullying text	7585
Training instances	13000(0:13000)
Testing instances	3000(0:3000)
Number of words before preprocessing	34340
Number of words after preprocessing	32009

better at correctly identifying positive instances. Specificity, or the true negative rate, measures the proportion of actual negative instances correctly identified by the models. The models had similar specificity values, with fastText at 0.8873, Word2Vec at 0.8926 and GloVe at 0.8952. These values indicate that the models were effective in correctly identifying negative instances. The table also provides two additional metrics, P_0 and P_1. It appears that the scores associated with the positive predictive value is (P_1) and negative predictive valus is (P_0). These findings provide insights into the models' performance and can inform the reader about the effectiveness of these approaches in the given research context.

Table 2. Performance evaluation with different embeddings.

Metrics	Proposed Model(fastText)	Word2Vec	GloVe
Accuracy	**0.9163**	0.9042	0.9067
True Positive	**1600**	1553	1557
False Positive	171	163	159
False Negative	98	145	141
True Negative	1346	1354	**1358**
Sensitivity	**0.9423**	0.9146	0.9170
Specificity	0.8873	0.8926	**0.8952**
P_0	**0.9321**	0.9033	0.9059
P_1	0.9034	0.9050	**0.9073**

Table 3. Performance evaluation with models.

Model	Accuracy	Precision	Recall	F1-score
LSTM+GRU+CNN+FastText(Proposed Model)	**91.63%**	**0.9170**	**0.9163**	**0.9162**
LSTM + BERT embedding	89.42%	0.89	0.89	0.89
MConv-LSTM + BERT embedding 8	87.87%	0.88	0.88	0.88
Bangla BERT fine-tune	88.57%	0.89	0.88	0.89

4.3 Performance Comparison with Different Related Models

The proposed model, which combines LSTM, GRU, and CNN with fastText embedding achieved the highest accuracy of 91.63% as mentioned in Table 3. The results of the three baseline methods have been obtained by rerunning those models on the mentioned dataset. We achieved the same results as mentioned in the paper publishing the original dataset. It demonstrated strong precision (0.9170), recall (0.9163), and F1-score (0.9162), indicating its effectiveness in accurately classifying text and outperformed the related models for the used dataset by Belal et al. [7]. The LSTM + BERT embedding model used by the author achieved an accuracy of 89.42% and exhibited consistent precision, recall, and F1-score values of 0.89, indicating its reliability in text classification tasks. The other two models had accuracy of 87.87% and 88.57% respectively. Our model outperformed and gained better score in all the metrics compared to these models (Fig. 3).

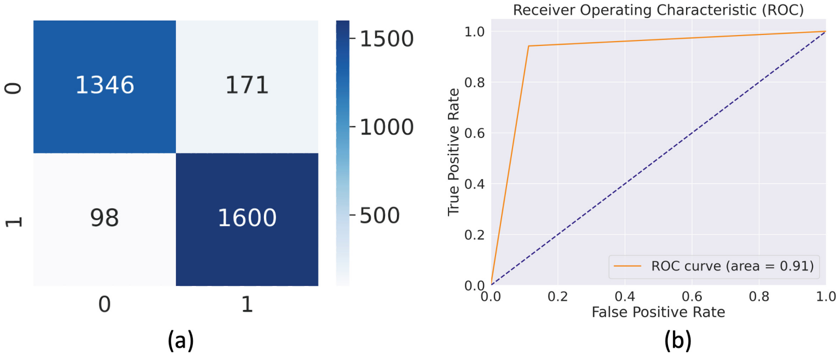

Fig. 3. (a) Confusion matrix (b) ROC curve of the proposed model. AUC score of the proposed model was 0.91.

4.4 Performance Comparison with Different Architectures of the Proposed Model

We have tweaked our proposed model with different combinations of the underlying architectures. We have used a sequence of LSTM, GRU, Convolutional layers in our proposed model. We have made some alterations to observe the change in the results by adding some layers and removing them as well as mentioned in Table 4. We have removed the two convolutional layers and the max pooling in a combination and this alteration produced a result of 91.56% accuracy while gaining sensitivity of 0.9187 and specificity of 0.9121. Again, including an additional LSTM layer, reducing a dense layer and adding an additional dense layer we analyzed the results. In this comparison, our proposed model's architecture outperformed the other ones.

Table 4. Performance Comparison with Different Architectures

Architecture	Accuracy	Sensitivity	Specificity
Proposed Combination	**91.63%**	**0.9423**	**0.8873**
Reducing Convolution layers	91.56%	0.9187	0.9121
Adding a LSTM Layer	90.05%	0.9079	0.9021
Adding a Dense Layer	91.58%	0.9164	0.9161
Removing a Dense Layer	91.50%	0.9321	0.8954

By adjusting various parameters, we developed our proposed model. The LSTM, GRU, CNN, and Dense layers were tried out in many ways before we figured out this model. In different units, filters, neuron ranges, we experimented with different combinations of 0 to 3 layers of BiLSTM, LSTM, BiGRU, GRU, Convolution, and dense layers. The results are displayed in Table 5.

Table 5. Model Tuning Range

Layer	No of Layers tried	Tuning Range
BiLSTM	0–3	100–300 units
LSTM	0–3	100–300 units
Convolutional Layer	0–2	50–500 filters
Dense layers (ReLU)	1–3	100–1000 neurons
Dense layers (Softmax Activation)	1	2 neurons

4.5 Text Analysis in the Feature Level

The reasons behind classifying a text as cyberbullying or non-cyberbullying depends on the analysis of the text. The percentage of each word carries an indication to whether the text will fall into cyberbullying category or not. It has been seen that the significant words that make the sentence a cyberbullying text, is found in greater percentage in the training set of all cyberbullying texts as in Fig. 4.

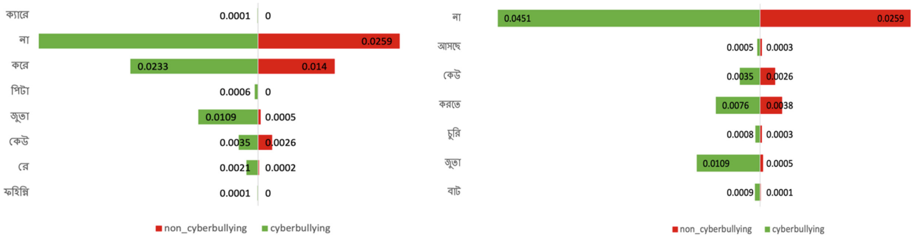

Fig. 4. The explanation of the predictions of a cyberbullying and a non-cyberbullying texts.

5 Conclusion

The rising rate of cyberbullying necessitates immediate action to safeguard people from its adverse effects. To minimise the consequences of cyberbullying, real-time cyberbullying detection is essential. By harnessing the strength of advanced neural networks and large quantities of social media data, deep learning algorithms provide a promising solution for automated cyberbullying identification. In order to provide a safer online environment for everyone, this research study intended to explore and assess current developments in deep learning models for cyberbullying detection and proposing a promising model with a satisfactory accuracy of 91.63%. To the best of our knowledge this work is unique and not published anywhere. There is still a lot of rooms for future studies and improvement such as building a much larger dataset and using better and more furnished tools like stemmers, embedding techniques, lemmatises etc. Therefore, in spite of providing a satisfactory result, this study opens up new opportunities for this research field by highlighting the potential in precisely recognising and classifying cases of cyberbullying.

References

1. Social Media in Bangladesh - 2023 Stats Platform Trends - OOSGA – oosga.com
2. Ahammed, S., Rahman, M., Niloy, M.H., Chowdhury, S.M.H.: Implementation of machine learning to detect hate speech in Bangla language. In: 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), pp. 317–320. IEEE (2019)
3. Ahmed, M.F., Mahmud, Z., Biash, Z.T., Ryen, A.A.N., Hossain, A., Ashraf, F.B.: Cyberbullying detection using deep neural network from social media comments in Bangla language. arXiv preprint arXiv:2106.04506 (2021)
4. Arreerard, R., Senivongse, T.: Thai defamatory text classification on social media. In: 2018 IEEE International Conference on Big Data, Cloud Computing, Data Science & Engineering (BCD), pp. 73–78. IEEE (2018)
5. Aurpa, T.T., Sadik, R., Ahmed, M.S.: Abusive Bangla comments detection on Facebook using transformer-based deep learning models. Soc. Netw. Anal. Min. **12**(1), 24 (2022)
6. Banik, N., Rahman, M.H.H.: Toxicity detection on Bengali social media comments using supervised models. In: 2019 2nd International Conference on Innovation in Engineering and Technology (ICIET), pp. 1–5. IEEE (2019)
7. Belal, T.A., Shahariar, G., Kabir, M.H.: Interpretable multi labeled Bengali toxic comments classification using deep learning. In: 2023 International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 1–6. IEEE (2023)
8. Cecillon, N., Labatut, V., Dufour, R., Linarès, G.: Abusive language detection in online conversations by combining content-and graph-based features. Front. Big Data **2**, 8 (2019)
9. Chakraborty, P., Seddiqui, M.H.: Threat and abusive language detection on social media in Bengali language. In: 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), pp. 1–6. IEEE (2019)

10. Das, A.K., Asif, A.A., Paul, A., Hossain, M.N.: Bangla hate speech detection on social media using attention-based recurrent neural network. J. Intell. Syst. **30**(1), 578–591 (2021). https://doi.org/10.1515/jisys-2020-0060
11. Gers, F.: Learning to forget: continual prediction with LSTM. In: 9th International Conference on Artificial Neural Networks: ICANN 1999 (1999). https://doi.org/10.1049/cp:19991218
12. Ghosh, T., Chowdhury, A.A.K., Banna, M.H.A., Nahian, M.J.A., Kaiser, M.S., Mahmud, M.: A hybrid deep learning approach to detect Bangla social media hate speech. In: Hossain, S., Hossain, M.S., Kaiser, M.S., Majumder, S.P., Ray, K. (eds.) Proceedings of International Conference on Fourth Industrial Revolution and Beyond 2021, pp. 711–722. Springer, Cham (2022). https://doi.org/10.1007/978-981-19-2445-3_50
13. Gordeev, D.: Automatic detection of verbal aggression for Russian and American imageboards. Procedia. Soc. Behav. Sci. **236**, 71–75 (2016)
14. Haidar, B., Chamoun, M., Serhrouchni, A.: Multilingual cyberbullying detection system: detecting cyberbullying in Arabic content. In: 2017 1st Cyber Security in Networking Conference (CSNet), pp. 1–8. IEEE (2017)
15. Haidar, B., Chamoun, M., Serhrouchni, A.: A multilingual system for cyberbullying detection: Arabic content detection using machine learning. Adv. Sci. Technol. Eng. Syst. J. **2**(6), 275–284 (2017)
16. Social Media User Statistics: How Many People Use Social Media? searchlogistics.com. https://www.facebook.com/mattwoodwarduk. https://www.searchlogistics.com/learn/statistics/social-media-user-statistics/. Accessed 12 July 2023
17. Huan, J.L., Sekh, A.A., Quek, C., Prasad, D.K.: Emotionally charged text classification with deep learning and sentiment semantic. Neural Comput. Appl. **34**(3), 2341–2351 (2021). https://doi.org/10.1007/s00521-021-06542-1
18. Ibrohim, M.O., Budi, I.: Multi-label hate speech and abusive language detection in Indonesian twitter. In: Proceedings of the Third Workshop on Abusive Language Online, pp. 46–57 (2019)
19. Karim, M.R., Chakravarthi, B.R., McCrae, J.P., Cochez, M.: Classification benchmarks for under-resourced Bengali language based on multichannel convolutional-LSTM network. In: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), pp. 390–399. IEEE (2020)
20. Kumar, A., Sachdeva, N.: Multi-input integrative learning using deep neural networks and transfer learning for cyberbullying detection in real-time code-mix data. Multimedia Syst. **28**(6), 2027–2041 (2022)
21. Lane, J.: The 10 most spoken languages in the world (2023). https://www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world
22. Luan, Y., Lin, S.: Research on text classification based on CNN and LSTM. In: 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), pp. 352–355. IEEE (2019)
23. Luo, X.: Efficient English text classification using selected machine learning techniques. Alex. Eng. J. **60**(3), 3401–3409 (2021)
24. Malik, P., Aggrawal, A., Vishwakarma, D.K.: Toxic speech detection using traditional machine learning models and BERT and fasttext embedding with deep neural networks. In: 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), pp. 1254–1259. IEEE (2021)
25. Mohammed, A., Kora, R.: An effective ensemble deep learning framework for text classification. J. King Saud Univ.-Comput. Inf. Sci. **34**(10), 8825–8837 (2022)

26. Pawar, R., Raje, R.R.: Multilingual cyberbullying detection system. In: 2019 IEEE International Conference on Electro Information Technology (EIT), pp. 040–044. IEEE (2019)
27. Pratiwi, N.I., Budi, I., Jiwanggi, M.A.: Hate speech identification using the hate codes for Indonesian tweets. In: Proceedings of the 2019 2nd International Conference on Data Science and Information Technology, pp. 128–133 (2019)
28. Ptaszynski, M., Pieciukiewicz, A., Dybała, P.: Results of the poleval 2019 shared task 6: first dataset and open shared task for automatic cyberbullying detection in polish twitter (2019)
29. Ritu, S.S., Mondal, J., Mia, M.M., Al Marouf, A.: Bangla abusive language detection using machine learning on radio message gateway. In: 2021 6th International Conference on Communication and Electronics Systems (ICCES), pp. 1725–1729. IEEE (2021)
30. Sazzed, S.: Abusive content detection in transliterated Bengali-English social media corpus. In: Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching, pp. 125–130 (2021)
31. Team, B.: All the latest cyberbullying statistics for 2023 (2023). https://www.broadbandsearch.net/blog/cyber-bullying-statistics
32. Yuvaraj, N., et al.: Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification. Comput. Electr. Eng. **92**, 107186 (2021)