Bangla Hate Speech Detection System Using Transformer-based NLP and Deep Learning Techniques

Omar Faruqe
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
omar.faruqe15@northsouth.edu

Mubassir Jahan
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
mubassir.jahan@northsouth.edu

Md. Faisal
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
md.faisal3@northsouth.edu

Md. Shahidul Islam
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
shahidul.islam3@northsouth.edu

Riasat Khan
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
riasat.khan@northsouth.edu

2023 3rd Asian Conference on Innovation in Technology (ASIANCON) | 979-8-3503-0228-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ASIANCON58793.2023.10269919

***Abstract*—Hate speech is a form of discriminatory communication disrupts community standards and breaches the line of self-limitation, causing harm to others and occasionally leading to cyberbullying. Hate speech spreads hatred toward a person or a particular group based on various characteristics, e.g., race, religion, gender, and so on is referred to as bias. The offensive speech detection system is the frontier where researchers are battling to provide secure internet using natural language processing and machine learning approaches. In this research, we strive to create an automatic Bangla hate speech detection system using natural language processing (NLP) and deep learning approaches. We utilized our custom dataset and some labeled data from an open-access repository in this work. 4,978 data from both sources were merged and implemented in our proposed model. Different data preprocessing techniques, tokenization, stemming, and removal of stopwords have been applied. Four deep learning and NLP-based classifiers have been applied to detect Bangla hate speech. Google API has been employed to convert text from Bangla to English. The emojis were removed from the datasets and the data were translated into Bangla. The GRU and Attention techniques performed best with 98.87% and 98% accuracies, respectively.**

***Index Terms*—Attention Model, BERT Fine Tuning, Bi-LSTM, cross-validation, GRU, hate speech detection, transformer, word embedding.**

I. Introduction

Hate speech is an offensive expression that attacks or disparages a person or group based on race, ethnicity, national origin, political opinion, religion, sexual orientation, gender identity, or other characteristics [1]. Hate speech can be extremely harmful, as it can fuel discrimination, violence, and intolerance towards certain groups and can create a climate of fear and hostility [2]. It is essential to recognize and address hate speech in order to promote a more inclusive and tolerant society [3]. While freedom of speech is a fundamental right, it is important to understand that hate speech is not protected under the First Amendment in the United States and is also prohibited by law in many other countries. It is crucial that we respect the rights and dignity of all individuals and work towards creating a society that celebrates diversity and inclusivity.

Bangla is one of the world's most widely spoken languages. However, hate speech detection in Bangla has not been explored intensely. The target of the project is to detect hate speech in Bangla. Everyone has the right to free expression. However, in the guise of free speech, this fundamental right is used to physically or verbally discriminate against others and assault general people. Hate speech that displays hatred towards a person or group based on characteristics such as race, religion, gender, and so on is referred to as prejudice. Hundreds of incidents have happened due to hate speech and crimes, resulting in disputes, riots, and even murders. Despite the government and concerned stakeholders' efforts to combat or minimize the problem, finding effective answers for regulating or resolving it through legislation or harsh penalties against agitators is exceptionally challenging. Social media are increasing intelligent frameworks and utilizing machine learning approaches to improve hate speech identification.

In this paper, an automatic Bangla hate speech detection system has been developed using deep learning and NLP techniques. A combined dataset of approximately 5,000 samples from two authentic open-source repositories has been obtained. The dataset has been collected from various social media platforms. Google Translator, the Googletrans Python package, has been used in Bangla Emoji detection for data preprocessing. Next, tokenization, TF-IDF vectorization and word embedding techniques are applied. Transformer-based BERT, GRU, Attention, and LSTM models have been employed to detect Bangla hate speech. These automated, deep learning-based prediction systems are deployed in the Google Colab framework. The performance of the proposed system is evaluated in terms of mean average precision, frame rate, accuracy, etc.

II. RELATED WORK

Recently, extensive research has been done to investigate the automatic detection of hate speech, especially from social media and news websites. Some recent articles on Bangla hate speech recognition are discussed in the following paragraphs. For instance, in [4], the researchers created a dataset corpus that contained the Bangla comments from the social media site Facebook and annotated them for negative, positive, and neutral categories to detect hate speech. The authors explored, analyzed, and preprocessed the data, then performed the exploratory data analysis. They used the traditional machine learning algorithm with SVM models, BNLP tools for data cleaning, and Term Frequency Inverse Document Frequency vectorizer to extract the data into a matrix of features. They employed a vectorizer to measure the accuracy using uni-gram, bigram, and trigram classifiers. Finally, the authors concluded that the highest accuracy was obtained with the MNB classifier with unigram features, where 62.85% of cases were predicted correctly compared to the other classification methods. Junaid et al. [5] implemented various deep learning techniques to detect Bangla hate speech-related videos. The authors used their privately curated dataset, first converting video to audio and corresponding Bangla text format. Hyperparameter adjustment was applied to LSTM and GRU deep learning techniques. This work obtained 98.89% and 86.67% accuracy for the GRU and LSTM models, respectively.

Karim and his colleagues [6] implemented an automated abusive social media comment detection system for the Bangla alphabet and translated the Bangla text dataset obtained from various public Facebook page comments. The authors used an online scraper tool to avoid generating a biased dataset and collected more than 2,000 labeled comments. For dataset preprocessing, they used a profanity-based detection algorithm which helps to detect divisive comments. Random forest gives the best performance with 72.1% accuracy compared to other models. Jahan and team [7] proposed an explainable deep learning approach for different Bangla hate speech detection deep learning models. To achieve better performance, the researchers trained their custom dataset with Conv-LSTM, Bangla BERT, and XML-Roberta techniques. As the dataset size was limited, some ML classifiers did not perform well for the dataset. DNN and BERT Variants accomplished a better F1 score of 88% and 89%, respectively. Das and others [8] obtained more than 7,425 Bengali comments to design an automatic offensive social media-related detection system. The authors manually collected the comments to introduce seven categories of hate speech. A Bangla Emot analysis tool was developed to assist in identifying the emotions expressed by emoticons and emojis. This action makes it easier to understand what constitutes hate speech in Bangla. The authors used supervised machine learning techniques, including attention-based decoder models. The GRU algorithm accomplished an accuracy of 74%. The performance was subsequently improved with 77% accuracy using the attention mechanism.

In [9], authors provided a framework for speech acquisition combined with the speaker's position and translating the audio signals into texts. A system based on long short-term memory (LSTM) and a variation of Recurrent neural networks (RNN) are used to categorize Bangla speeches as suspicious. Five thousand suspicious and non-suspicious samples have been used to construct the dataset. Various conventional data preprocessing techniques were applied, including tokenization, removing punctuations and duplicates, padding and word embedding. The LSTM technique accomplished the best performance with 94% accuracy and 0.95 F1-score. Hussain et al. [10] performed abusive text analysis using Bangla comment data and manually gathered the training and test data. The authors collected Bangla comments and discussions from various online news portals, YouTube channels and verified Facebook pages of famous personalities. Next, they performed a survey to label the obtained data into two categories. After applying traditional preprocessing approaches, a custom root-level NLP algorithm with a unigram feature extraction technique was applied. The implemented model achieved approximately 69% accuracy and an F1-coefficient of 0.67. Remon and coauthors [11] initiated Bangla hate speech recognition employing various machine learning techniques. The authors introduced a custom database of more than ten thousand Facebook comments from a wide range of public pages and groups. The SVM model with the fastText word embedding technique produced the highest detection accuracy.

Romim et al. [12] presented a comprehensive dataset of Bangla offensive social media comments. Fifty volunteers annotated individual observations thrice to reduce the bias. SVM model with BengFastText word embedding approaches achieved better performance with accuracy and F1-score of 87.5% and 0.88, respectively. Hossain et al. [13] compiled an extensive recorded Bangla voice-based database of approximately thirty hours. The audio speech was recorded by 50 participants comprising seven dialects. Scaled autoencoder and multi-label machine learning techniques were applied to identify regional hate speeches. Eshan and the team [14] developed an effective Bangla offensive text recognition system using machine learning approaches. The authors created their own dataset of two classes from eleven prominent public figures' Facebook profiles. The SVM model with TF-IDF Vectorizer provided the best results concerning classification accuracy and other metrics.

From the above paragraphs, considerable work has been performed on Bangla abusive speech detection based on natural language processing and deep learning methods. A single dataset of hate speech in Bangla, whether it be from an open source or collected by the authors, is used in the majority of these articles. This concern has motivated us to apply advanced transformer-based deep learning techniques on a combined dataset from different authentic open-source repositories.

III. Methodology

The datasets, preparation techniques, and detection algorithms used by the proposed Bangla hate speech detection system are described in this section. All the deep learning and NLP techniques have been trained and evaluated in the Google Colab environment. PyTorch framework has been used to implement the programming codes. Necessary Python dependencies have been downloaded to perform specific operations in this work. The working sequences of the proposed Bangla hate speech detection system are represented in Fig. 1. This paper uses a fine-tuned BERT model, Bi-LSTM, GRU and several machine-learning approaches to detect hate speech in Bangla.

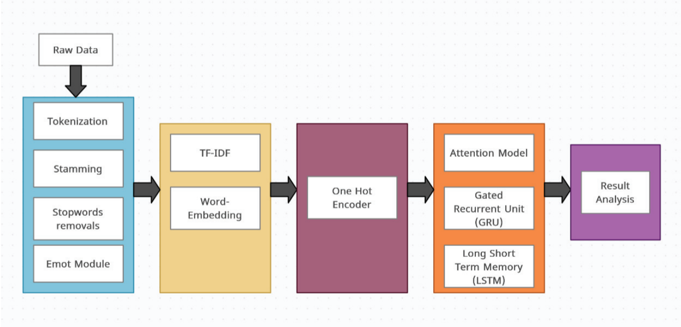
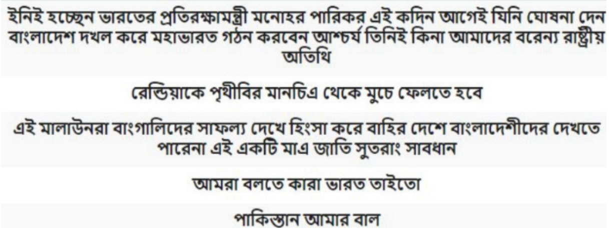


Fig. 1. Working steps of the proposed Bangla hate speech detection system.

A. Dataset and Corresponding Preprocessing Steps

The working process of the proposed Bangla hate speech detection system involves collecting hate speech from social media sites and performing exploratory data analysis, data cleaning, modeling, and evaluating the proposed system. In this work, we used more than 5,000 merged classified instances of two distinct categories (hate and non-hate). The dataset has been collected from two different authentic open-source Bengali hate speech datasets. These datasets were curated from individual comments and opinions of different popular Bangladeshi Facebook groups, most-followed celebrity Facebook pages and public profiles, YouTube channels, and newspaper articles. The data samples have been labeled into two categories: hate and non-hate. Fig. 2 demonstrates various sample Bangla hate texts.



ইনিই হচ্ছেন ভারতের প্রতিরক্ষামন্ত্রী মনোহর পারিকর এই কদিন আগেই যিনি ঘোষনা দেন বাংলাদেশ দখল করে মহাভারত গঠন করবেন আশ্চর্য তিনিই কিনা আমাদের বরেন্য রাষ্ট্রীয় অতিথি

রেন্ডিয়াকে পৃথীবির মানচিএ থেকে মুচে ফেলতে হবে

এই মালাউনরা বাংগালিদের সাফল্য দেখে হিংসা করে বাহির দেশে বাংলাদেশীদের দেখতে পারেনা এই একটি মাএ জাতি সুতরাং সাবধান

আমরা বলতে কারা ভারত তাইতো

পাকিস্তান আমার বাল

Fig. 2. Sample of labeled datasets.

In this research, the employed data are preprocessed using various preprocessing methods, and the related information has been extracted from emoticons and emojis to detect the corresponding sentiment. Tokenization involves breaking up a sequence of characters or words into smaller units called tokens, as illustrated in Fig. 3. In NLP, tokenization is typically the first step in processing raw text data [15]. It involves separating the text into individual units, which are then used as input for further processing. Tokenization can be performed at different levels, depending on the specific NLP task and the nature of the text data [16]. Some common types of tokenization include word tokenization, sentence tokenization, and character tokenization.

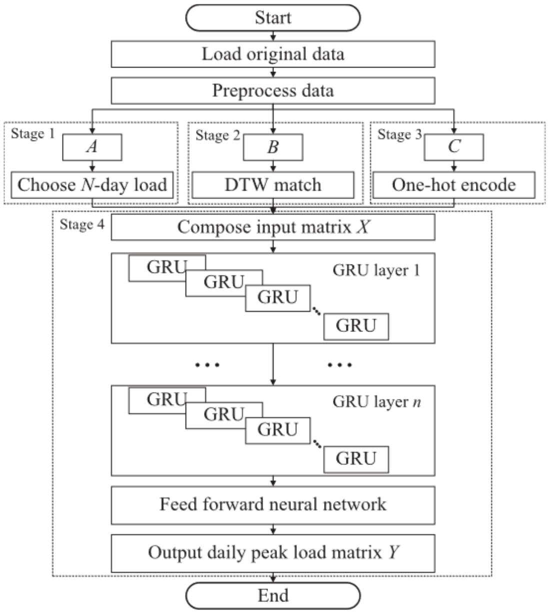
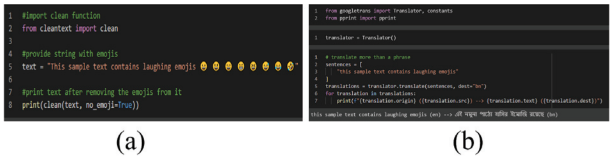


Fig. 3. Architecture of tokenization.



```
#import clean function
from cleantext import clean

#provide string with emojis
text = "This sample text contains laughing emojis 😀 😃 😄 😁 😆 😅 😂 🤣"

#print text after removing the emojis from it
print(clean(text, no_emoji=True))
```

(a)

```
from googletrans import Translator, constants
from pprint import pprint
```

```
translator = Translator()
```

```
# translate more than a phrase
sentences = [
    "this sample text contains laughing emojis"
]
translations = translator.translate(sentences, dest="bn")
for translation in translations:
    print(f"{translation.origin} ({translation.src}) --> {translation.text} ({translation.dest})")
```

```
this sample text contains laughing emojis (en) --> এই নমুনা পাঠ্যে হাসির ইমোজি রয়েছে (bn)
```

(b)

Fig. 4. Emoji and emoticons detection: (a) Original Text, (b) Detected Text.

Since we have collected the employed data samples from different sources in this work, it comprises many emojis and emoticons. We have implemented the translator method from the Googletrans Python package to detect emoticons and emojis and clean those tags. A sample text containing emojis and emoticons and its corresponding detected text have been illustrated in Fig. 4.

Features are extracted using TF-IDF vectorization, and the word embedding technique is also used. Multiple classification approaches such as BERT, Bidirectional LSTM, attention-based deep learning model and GRU are applied. Finally, the classification approaches' performances have been analyzed and compared. A summary of the employed deep learning and NLP techniques to detect Bangla hate speech in this work has been illustrated in the subsequent paragraphs.

B. Applied Models

1) BERT: Bidirectional Encoder Representations from Transformers (BERT) model considers each component of the incoming data differently based on its significance. The main applications of this model are computer vision (CV) and natural language processing (NLP) [17]. BERT is a deep learning algorithm trained on massive amounts of text data, using a bidirectional approach, which means it reads the text in both directions (from left to right and from right to left) [18]. The main innovation of BERT is its use of a transformer architecture, which allows it to capture the context of words in a sentence better than previous language models. Its pre-trained model has been made available for researchers and developers to use in their applications. A fine-tuned BERT model has been employed in this work.

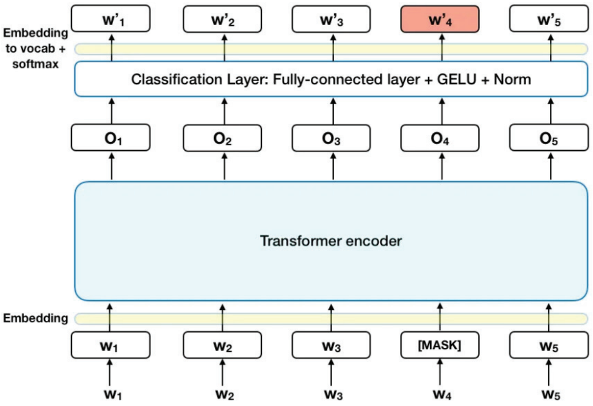

Embedding to vocab + softmax
w'₁
w'₂
w'₃
w'₄
w'₅
Classification Layer: Fully-connected layer + GELU + Norm
O₁
O₂
O₃
O₄
O₅
Transformer encoder
Embedding
w₁
w₂
w₃
[MASK]
w₅
w₁
w₂
w₃
w₄
w₅

Fig. 5. Architecture of the BERT Model.

Fig. 5 depicts the architecture of the employed fine-tuned BERT framework. The same frameworks are used in pre-training and fine-tuning in addition to the output layers.

2) Bi-LSTM: The bidirectional Long Short-Term Memory (Bi-LSTM) approach has been used for sentiment analysis, machine translation, and named entity recognition [19]. The outputs of both LSTM layers are then merged using different methods, e.g., sum, average, multiplication, and concatenation. This model can efficiently handle the vanishing gradient problem and preserve long-term dependencies in sequential data. By combining the advantages of both LSTMs and bidirectional processing, the Bi-LSTM model has been developed.

3) GRU: There is a similarity between GRU and the LSTM model. In LSTM, there is only an input gate and an output gate, but the GRU model has three gates input, output, and update gate, as shown in Fig. 6. The GRU model uses the update gates to control the flow of information, and because of this feature, the GRU model is more significant than the last model [20]. The GRU network consists of a set of recurrent layers, each containing a set of nodes or "cells."

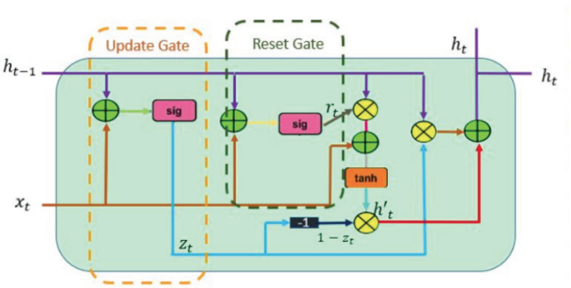
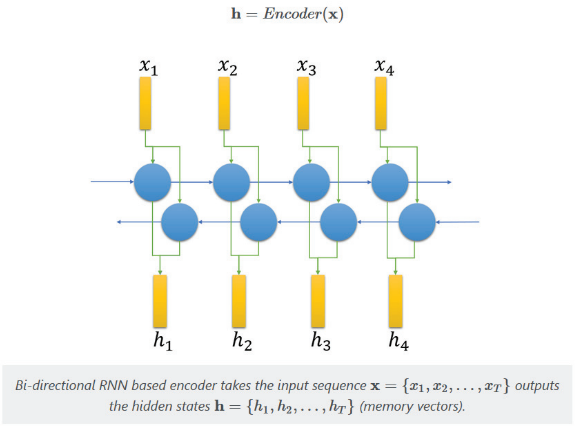


Fig. 6. Architecture of GRU model.

Fig. 6 represents the architecture of the GRU model, which trains faster and performs better than its predecessor LSTM technique. In addition to being more straightforward to modify, GRU is also faster and uses less memory than LSTM when adding new gates.

4) Attention-based model: Attention-based model is a sequence-to-sequence model aiming to produce an output sequence according to the given input sequence in different lengths [21]. The primary mechanism of this model is to improve the performance of the encoder to decoder in machine translation. The idea of the attention-based model is that it permits the decoder to utilize the most relevant part of the input sequence in a specific manner. In every output sequence, the decoder generates a word at a time, then takes the word in the previous step $(t-1)$ as input for generating the following word in the resulting output. This method is more reliable in short-term sequences but has complexity in long-term sequences. The architecture of the proposed attention model for Bangla hate speech detection has been demonstrated in Fig. 7.



Bi-directional RNN based encoder takes the input sequence $\mathbf{x} = \{x_1, x_2, \ldots, x_T\}$ outputs the hidden states $\mathbf{h} = \{h_1, h_2, \ldots, h_T\}$ (memory vectors).

Fig. 7. Attention model architecture.

IV. Results and Discussion

This section discusses the simulation results of the proposed deep learning-based Bangla hate speech recognition system.

Fig. 8 demonstrates the training and testing accuracies vs. epochs of the applied BERT model. According to this figure, the fine-tuned BERT approach achieved 94% and 77% training and test accuracies, respectively.

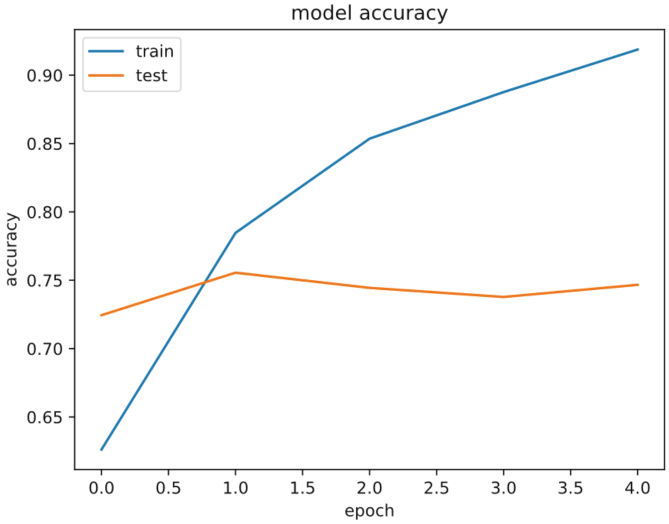


Fig. 8. Training and validation graphs of the BERT model.

TABLE I
VARIOUS EVALUATION METRICS OF THE APPLIED BERT TECHNIQUE

	Precision	Recall	F1-score	Support
0 (non-hate)	0.77	0.76	0.765	2534
1 (hate)	0.76	0.77	0.765	2466
accuracy	-	-	0.77	5000
macro avg	0.77	0.77	0.77	5000
weighted avg	0.77	0.77	0.77	5000

The Bi-LSTM model comprises 2,630,154 parameters. The dictionary-based Bi-LSTM approach accomplished an accuracy of 71.68%. However, there is an issue when we use a sentence from the Dictionary as input because the prediction is better. However, the input model cannot distinguish sentences given at random. We do not have to separate the training or test data for the pre-trained model. We can only obtain precise outcomes that are expected with training and test data.

The GRU model produced the best result of all the applied deep learning models. The GRU model attained the highest accuracy of 98.87%. Fig. 9 describes the GRU Model confusion matrix without normalization.

Fig. 10 and Fig. 11 illustrate the GRU model's loss graphs for the positive (hate) and negative (non-hate) classes, respectively.

Finally, Table II and Table III show the various performance metrics of the applied deep learning and natural

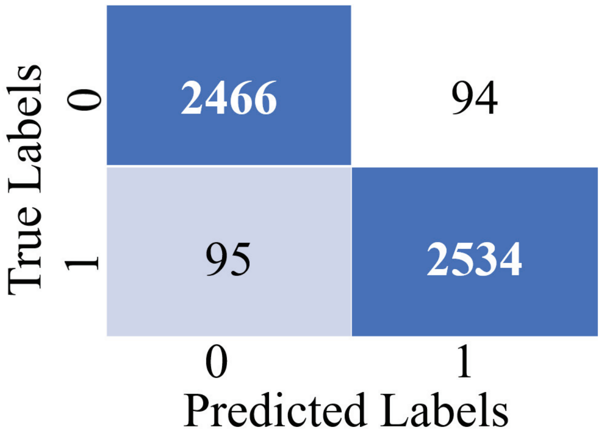


Fig. 9. GRU Model confusion matrix (without Normalization).

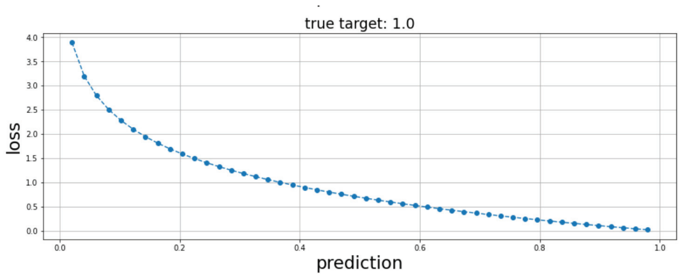


Fig. 10. GRU model loss graph for the positive (hate) class.

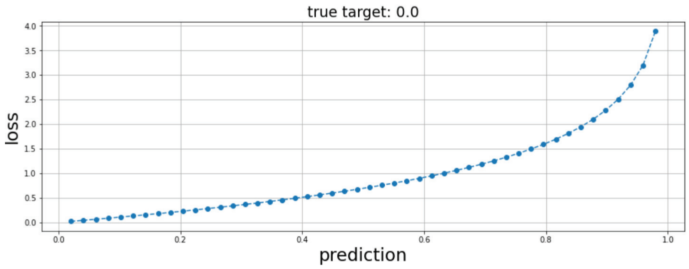


Fig. 11. GRU model loss graph for the negative (non-hate) class.

TABLE II
VARIOUS PERFORMANCE METRICS OF THE APPLIED MODELS

Model	Class	Precision	Recall	F1-Score	Accuracy
GRU	0	0.00	0.00	0.00	98.87%
	1	0.98	1.00	0.99	
Attention	0	0.00	0.00	0.00	98%
	1	0.98	1.00	0.99	
BERT	0	0.95	0.94	0.95	95%
	1	0.94	0.95	0.94	

TABLE III
VALIDATION ACCURACIES OF VARIOUS APPLIED MODELS

Applied Model	**Validation Accuracy**
BERT	94.04%
Bi-LSTM	71.68%
GRU	98.87%
Attention	98%

language processing-based Bangla hate speech identification approaches. According to Table III, the attention and GRU models attained the highest accuracies for the proposed Bangla hate speech classification system

V. Conclusion and future work

Hate speech refers to communication that disparages an individual or a group based on various characteristics like race, religion, ethnic and national origin, etc. It is frequently employed to promote hatred or violence toward members of specific communities. Hate speech may have terrible effects on people and communities and can foster an environment of intolerance and fear. This research aims to detect Bangla hate speech using deep learning and NLP techniques. Approximately five thousand samples have been collected by ourselves and from authentic sources. Various data preprocessing approaches have been applied, such as tokenization, stemming, extraction from emoticons and emojis, and removal of stopwords. Multiple deep learning and NLP classifiers have been used to detect Bangla hate speech. Google API has been employed to convert text from Bangla to English. The GRU and Attention techniques achieved the best performance among all the applied models.

A diverse dataset with more samples and hate speech categories can be collected in the future. We will try to enrich our dataset by collecting different sources, especially the video content, and preprocessing them for better results. We will also work on a real-time platform where we can show the results and visualize the implementation.

References

[1] M. A. Paz, J. Montero-Díaz, and A. Moreno-Delgado, “Hate speech: A systematized review,” *SAGE Open*, vol. 10, 2020.

[2] N. F. Azman and N. A. K. Zamri, “Conscious or unconscious: The intention of hate speech in cyberworld – A conceptual paper,” *Proceedings*, vol. 82, 2022.

[3] A. Tontodimma, E. Nissi, A. Sarra, and L. Fontanella, “Thirty years of research into hate speech: Topics of interest and their evolution,” *Scientometrics*, 2020.

[4] S. A. Kaiser, S. Mandal, A. K. Abid, E. Hossain, F. B. Ali, and I. T. Naheen, “Social media opinion mining based on Bangla public post of facebook,” in *International Conference on Computer and Information Technology*, pp. 1–6, 2021.

[5] M. I. Hossain Junaid, F. Hossain, and R. M. Rahman, “Bangla hate speech detection in videos using machine learning,” in *Annual Ubiquitous Computing, Electronics & Mobile Communication Conference*, pp. 0347–0351, 2021.

[6] M. R. Karim, B. R. Chakravarthi, J. P. McCrae, and M. Cochez, “Classification benchmarks for under-resourced bengali language based on multichannel convolutional-LSTM network,” in *International Conference on Data Science and Advanced Analytics*, pp. 390–399, 2020.

[7] M. Jahan, I. Ahamed, M. R. Bishwas, and S. Shatabda, “Abusive comments detection in Bangla-English code-mixed and transliterated text,” in *International Conference on Innovation in Engineering and Technology*, pp. 1–6, 2019.

[8] A. K. Das, A. A. Asif, A. Paul, and M. N. Hossain, “Bangla hate speech detection on social media using attention-based recurrent neural network,” *Journal of Intelligent Systems*, vol. 30, pp. 578–591, 2021.

[9] M. Rahman, A. Mohammad, H. M. H. Mohammad, and Kayes, “Towards a framework for acquisition and analysis of speeches to identify suspicious contents through machine learning,” *Complexity*, vol. 2020, pp. 1–14, 2020.

[10] M. G. Hussain and T. A. Mahmud, “A technique for perceiving abusive Bangla comments,” *GUB Journal of Science and Engineering*, vol. 4, pp. 11–18, 2017.

[11] N. I. Remon, N. H. Tuli, and R. D. Akash, “Bengali hate speech detection in public facebook pages,” in *International Conference on Innovations in Science, Engineering and Technology*, pp. 169–173, 2022.

[12] N. Romim, M. Ahmed, H. Talukder, and M. Saiful Islam, “Hate speech detection in the Bengali language: A dataset and its baseline evaluation,” in *International Joint Conference on Advances in Computational Intelligence*, pp. 457–468, 2021.

[13] P. S. Hossain, A. Chakrabarty, K. Kim, and M. J. Piran, “Multi-label extreme learning machine (MLELMs) for Bangla regional speech recognition,” *Applied Sciences*, vol. 12, 2022.

[14] S. C. Eshan and M. S. Hasan, “An application of machine learning to detect abusive Bengali text,” in *International Conference of Computer and Information Technology*, pp. 1–6, 2017.

[15] N. N. Prachi, M. Habibullah, E. H. Rafi, E. Alam, and R. Khan, “Detection of fake news using machine learning and natural language processing algorithms,” *Journal of Advances in Information Technology*, vol. 13, pp. 652–661, 2022.

[16] S. Siddique, S. Islam, E. E. Neon, T. Sabbir, I. T. Naheen, and R. Khan, “Deep learning-based bangla sign language detection with an edge device,” *Intelligent Systems with Applications*, vol. 18, pp. 1–12, 2023.

[17] T. Junaid *et al.*, “A comparative analysis of transformer-based models for figurative language classification,” *Computers and Electrical Engineering*, vol. 101, 2022.

[18] X. Han and L. Wang, “A novel document-level relation extraction method based on BERT and entity information,” *IEEE Access*, vol. 8, pp. 96912–96919, 2020.

[19] A. A. Sharfuddin, M. N. Tihami, and M. S. Islam, “A deep recurrent neural network with BiLSTM model for sentiment classification,” in *International Conference on Bangla Speech and Language Processing*, pp. 1–4, 2018.

[20] R. B. Islam, S. Akhter, F. Iqbal, M. S. U. Rahman, and R. Khan, “Deep learning based object detection and surrounding environment description for visually impaired people,” *Heliyon*, vol. 9, pp. 1–19, 2023.

[21] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, “Diabetes prediction using machine learning and explainable AI techniques,” *Healthcare Technology Letters*, vol. 10, pp. 1–10, 2023.