IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 July 2023, accepted 22 July 2023, date of publication 26 July 2023, date of current version 3 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3299021

RESEARCH ARTICLE

G-BERT: An Efficient Method for Identifying Hate Speech in Bengali Texts on Social Media

ASHFIA JANNAT KEYA[1], MD. MOHSIN KABIR[2], NUSRAT JAHAN SHAMMEY[1], M. F. MRIDHA[3], (Senior Member, IEEE), MD. RASHEDUL ISLAM[4], (Senior Member, IEEE), AND YUTAKA WATANOBE[5], (Member, IEEE)

[1]Department of Computer Science and Engineering, Bangladesh University of Business and Technology, Dhaka 1216, Bangladesh
[2]Superior Polytechnic School, University of Girona, 17071 Girona, Spain
[3]Department of Computer Science and Engineering, American International University-Bangladesh, Dhaka 1229, Bangladesh
[4]Department of Computer Science and Engineering, University of Asia Pacific, Dhaka 1205, Bangladesh
[5]School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu 965-8580, Japan

Corresponding author: Md. Rashedul Islam (rashed.cse@gmail.com)

ABSTRACT The rapid increase in Internet users has increased online concerns such as hate speech, abusive texts, and harassment. In Bangladesh, hate text in Bengali is frequently used on various social media platforms to condemn and abuse individuals. However, Research on recognizing hate speech in Bengali texts is lacking. The pervasive negative impact of hate speech on individuals' well-being and the urgent need for effective measures to address hate speech in Bengali texts have created a significant research gap in the Bengali hate speech detection field. This study suggests a technique for identifying hate speech in Bengali social media posts that may harm individuals' sentiments. Our approach utilizes the Bidirectional Encoder Representations from Transformers (BERT) architecture to extract Bengali text properties, whereas hate speech is categorized using a Gated Recurrent Units (GRU) model with a Softmax activation function. We propose a new model, G-BERT, that combines both models. We compared our model's performance with several other algorithms and achieved an accuracy, precision, recall, and F1-score of 95.56%, 95.07%, 93.63%, and 92.15%, respectively. Our proposed model outperformed all other classification algorithms tested. Our findings show that the strategy we have suggested is successful in locating hate speech in Bengali texts posted on social media platforms, which can aid in mitigating online hate speech and promoting a more respectful online environment.

INDEX TERMS Bidirectional encoder representations from transformers, deep learning, hate speech, gated recurrent unit, social platform.

I. INTRODUCTION

Social media has become an essential component of our everyday lives in the modern world, enabling us to connect with others, share our thoughts and experiences, and stay updated on current events. However, along with its advantages, social media has drawbacks, such as the frequency of online harassment and bullying. Cyberbullying is a serious problem where people use the Internet to bully others [1]. This can lead to many negative consequences for the bullied person, including feeling bad emotionally, acting out negatively, and experiencing health problems [2]. The anonymity these platforms provide can embolden individuals to engage in hate speech, which they might hesitate to express face-to-face [3]. Speech that provokes, instigates, or insults individuals or groups is known as hate speech, and it frequently considers elements such as race, skin tone, gender, sexual orientation, nationality, religion, or health [4], [5]. Such behavior undermines the online environment and can severely impact victims. As a result, the problem of hate speech and cyberbullying on social media platforms needs to be addressed immediately to promote a safer and more inclusive online space. Given the limited resources available in this language, this study aims to thoroughly examine this

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna Dulizia.

issue and suggest a computational method for Bengali hate speech detection.

Increased offensive language on social media has created a toxic online environment that harms users. Using hate speech based on various factors such as ethnicity, gender, and disability has become a significant issue for social media platforms, resulting in severe emotional and mental impacts on individuals [6]. Bauman et al. [7] recently examined how often children and teenagers experience online hate speech. They found many different ways of measuring cyberhate, and few studies have explicitly focused on cyberhate. Hate-speech can have devastating consequences, such as severe depression and even suicide [8]. Consequently, effective ways to spot and deal with abusive language on social media are urgently required. The difficulties in addressing hate speech on social media include the complexity of language, disagreements on what qualifies as hate speech, and restrictions on accessing data for algorithm training and testing [9]. This paper focuses on the challenge of detecting and fighting hate speech, and proposes the development of new technologies and a dataset to address this issue. This problem is of utmost importance as it affects many people worldwide, and its potential implications include creating a safer online environment and protecting vulnerable individuals from harm.

The rapid growth of internet users has increased online concerns, including the proliferation of hate speech, abusive texts, and harassment. In Bangladesh, hate speech in the Bengali language is frequently observed on various social media platforms, targeting individuals and perpetuating harmful sentiments [10]. However, research on recognizing and addressing hate speech in Bengali texts remains scarce [11], [12]. This study aims to fill this research gap by proposing a technique for identifying hate speech in Bengali social media posts, intending to mitigate the negative impact of hate speech on individuals and promote a more respectful online environment. The significance of improving hate speech detection cannot be overstated. Hate speech poses severe consequences for individuals, communities, and society. It fuels hostility, perpetuates discrimination, and undermines social cohesion. Our research focuses on developing a practical approach for detecting hate speech in Bengali texts recognizing the urgent need to address this issue. Detecting hate speech in Bengali texts presents unique challenges. The Bengali language is complex, with its own linguistic nuances and cultural context. Existing research and resources for hate speech detection are limited in the Bengali language, despite Bengali being the 7th most spoken language globally, with nearly 300 million speakers [13], [14]. However, several studies have been conducted to detect hate speech in different languages [15], [16]. Bengali Language Processing's lack of development and the scarcity of high-quality Bengali text datasets are to blame for this lack of interest [17]. As a result, there is an urgent need to develop specialized approaches tailored to the specific characteristics of the Bengali language. This study aims to bridge the research gap by proposing a novel model called G-BERT, which combines the power of BERT for extracting Bengali text properties and GRU with a Softmax activation function for hate speech classification. By leveraging the strengths of these models, we aim to achieve a high level of accuracy and effectiveness in identifying hate speech in Bengali and Banglish texts.

Through numerous studies, there has been a growing emphasis on applying computational algorithms to detect hate speech on social media over the past few years [18]. Several supervised learning-based approaches using Machine Learning (ML) techniques have been developed for hate speech detection using English texts [14], [19]. However, detecting hate speech in other languages presents unique challenges that require language-specific approaches [20]. Ketsbia et al. [21] conducted a study to identify hate tweets using four ML-based classifiers: Multinomial Naive Bayes (MNB), Bernoulli Naive Bayes (BNB), and Logistic Regression (LR), on two datasets. Additionally, they explored various Doc2Vec methods. The researchers found that two Convolutional Neural Network (CNN) models built using Continuous Bag-of-Words and skipgram outperformed the ML-based classifiers on both datasets with accuracies of 89.70% and 92.88%, respectively. However, Deep Learning (DL) based classifiers have recently gained popularity in identifying hate speech, with numerous research papers published on the subject [3], [22].

After examining previous studies, CNNs and recurrent neural networks (RNN) have demonstrated promising outcomes in detecting hate speech [23], [24]. However, self-supervised learning-based approaches are relatively rare in this field, and most of the existing approaches have been developed using English text or comments for detecting hate speech. In this study, we concentrate on identifying hate speech in the Bengali language on various social media platforms using a novel approach combining BERT and GRU models. We provide a detailed description of our proposed G-BERT model and present experimental findings for Bengali hate speech detection.

The following is the summary of major contributions:

- To automatically identify hate speech from Bengali writings on multiple social media sites, we propose the G-BERT model, a deep learning-based approach.
- The G-BERT model efficiently captures contextual and semantic information in the Bengali text by fusing the GRU and BERT models.
- We evaluate the performance of the G-BERT model using various performance metrics and compared it with existing studies in the area of hate speech detection.
- To reduce the risk of hate speech in Bengali social media sites, this paper thoroughly explains the G-BERT model and shows how it can be used to identify hate speech from Bengali text.

The remaining parts of the paper are divided into the following sections. Section II contains related works. Section III clarifies the methodologies. Section IV discusses the findings and results and also a comparison with existing models.

A discussion section is added in section V. Finally, Section VI brings our work to a conclusion.

II. RELATED WORK

A supervised learning-based approach is widely used to identify hate speech circulating on social media [25], [26]. A work by Jahan et al. [27] suggested a strategy to use ML techniques to discover offensive remarks on public websites (Support Vector Machine, Random Forest, and Adaboost). They gathered public comments posted on Facebook pages to train their model and considered Bangla text, transliterated Bangla text, and Bangla-English code-mixed text. William et al. [28] proposed various approaches to utilize machine learning to find hate speech on social media. They compared several feature engineering strategies to discover the most effective machine learning algorithm. According to the study, the support vector machine (SVM) method combined with bigram features had the best overall accuracy at 79%. The results offer crucial insights into the application of machine learning to the fact of hate speech on social sites and can act as a baseline for further investigation into automatic text classification algorithms.

Currently, Deep Learning-based work is gaining more popularity in hate speech detection. Several studies have been published using Neural networks that outperformed state-of-the-art ML classifiers [29], [30]. Roy et al. [29] explained a Deep Convolutional Neural Network (DCNN) utilizing tweet text with a pre-trained GloVe embedding to extract semantic features. Many baseline Machine Learning classifiers and Long Short Term Memory (LSTM) is used for comparison with the suggested model. The proposed DCNN model outperformed and acquires an F1-score of 0.92, with a misclassification rate of approximately 2%. Another study by Khan et al. [31] introduced an innovative deep learning model, called BiCHAT, for identifying hate speech on Twitter. To develop tweet representations, this model employs Bidirectional LSTM, DCNN, and hierarchical attention-based techniques. Doing so surpasses the accuracy of previously established methods and benchmark techniques concerning precision, recall, and f-score. Furthermore, the model displays promising accuracy in both training and validation. the authors also explored the effects of diverse neural network components, embedding techniques, activation functions, batch size, and optimization algorithms on the overall performance of the BiCHAT model.

In addition, Transfer Learning (TL) is utilized in the DL-based approach for hate speech recognition. Melton et al. [30] proposed a novel framework with three datasets. To identify hate speech, a hybrid model is constructed that integrates CNN, GRU, and fully connected (FC) layers. This model was created by utilizing both transfer learning and weakly supervised learning techniques. The ensemble approach trained with comparatively weaker supervision acheived a hate recall of 67% on posts from Gab. Another study by Ali et al. [32] proposed a hate speech lexicon for the Urdu language and utilized it to annotate a collection of 10,526 Urdu tweets. They experimented with several machine learning techniques to evaluate their approach, including transfer learning using pre-trained FastText Urdu word embeddings and multi-lingual BERT embeddings. Four different BERT variants were tested, resulting in F1-scores of 0.68, 0.68, and 0.69 for BERT, xlm-roberta, and distil-Bert, respectively, outperforming baseline models. Recently, multimodal learning has become popular in hate speech detection [33]. A study by Rana and Jha [34] suggested a deep learning framework that utilizes multiple modes to identify hate speech in multimedia, emphasizing the speaker's emotional state and its impact on the spoken language. They have introduced a fresh Hate Speech Detection Video Dataset (HSDVD) designed for multimodal learning. The findings demonstrate that integrating emotional features results in substantial enhancement compared to models based solely on text in identifying hateful multimedia material. D'Sa et al. [35] proposed Dl-based binary and multi-class classification for toxic speech detection. In the feature-based approach, fastText and BERT embeddings are used as input features for CNN and Bi-LSTM classifiers. And a fine-tuned pre-trained model BERT is utilized for the second approach. On a Twitter corpus, they found that fine-tuning the BERT model have outperformed feature-based methods. Lee et al. [36] designed a decision system that identifies offensive texts using unsupervised learning. They used the word embeddings of Word2vec's skip-gram and the cosine similarity to identify newly created abusive terms. They verified their system's performance using word messages and comments on various platforms. Transformer-based approaches that utilize contextual embedding are limited to this domain.

Thus, we propose a transformer-based approach to detect Bengali hate speech using the Bidirectional Encoder Representation of the Transformer and Gated Recurrent Unit. By combining both models, we introduce the G-BERT model to identify hate speech in Bengali language text across different social media platforms.

III. METHODOLOGY

The following sections demonstrate the proposed methodology:

A. DATA COLLECTION

Instead of manual processing, a data crawling process was used for automatic data collection procedure. To gather data, various sources such as different Bengali online news portals (e.g., banglanews24.com, prothomalo.com, kalerkontho.com) and social media sites (e.g., Facebook, Twitter) were used. The data crawling process involves leveraging the power of the BeautifulSoup Python library to extract posts and comments related to hashtags and emojis from various social media platforms. In addition, specific criteria and keywords were used to select data from Bengali online

TABLE 1. Some samples from our dataset.

Sentence	English Translation	Label
কুত্তার বাচ্চা তোর খবর আছে	You son of a dog, you're finished.	Hate Speech
আইডি স্টক করতেছেন মনে হচ্ছে	It feels like you have been stalking my id.	No Hate Speech
তোকে ঘাড়ে ধাক্কা দিয়ে বিতাড়িত করা হবে	You will be pushed out by the shoulder.	Hate Speech
তুই একটা নাস্তিক শালা	You are an atheist bastard.	Hate Speech
অনেক ছবি তুলে রাখবো	I will take many pictures and keep them.	No Hate Speech
রাজাকারের বাচ্চা	Rajakar's child.	Hate Speech
জীবনের নিশ্চয়তা নাই	There is no certainty in life.	No Hate Speech
একবারে মাইরা কবর দিয়া দিমু	I will kill you and bury you in a grave all at once.	Hate Speech
তাদেরকে তাদের মতো বাঁচতে দেই	Let them live as they wish.	No Hate Speech

news portals and social media sites. The criteria included focusing on content published within a specific time period and targeting platforms with a significant user base and active engagement. The keywords used for data selection were carefully chosen to cover topics related to hate speech, abusive language, and online harassment in Bengali. Some examples of keywords used include Hate speech, Abusive language, Offensive remarks, Cyberbullying, Online harassment, Discrimination, Threats, Insults, Racist comments and Misogyny. From several Bengali websites, 20000 posts, comments and memes were accumulated. By analyzing those, almost 50% of the total content was found offensive. In the dataset, the total number of offensive or hate words is 110083 and the remaining 102782 words were non-offensive. The data labeling is manually done with eight graduate students from the Advanced Machine Learning (AML) lab in Bangladesh Business and Technology University (BUBT) [37]. The data labels undergo a thorough review process by the expert team at the Advanced Machine Learning (AML) lab. Our team consists of esteemed professionals, including professors, Ph.D. students with over five years of experience, and researchers with extensive knowledge in the field of Bengali language processing. This dataset has been utilized in another research work addressing the detection and analysis of offensive text in Bengali text [37]. The dataset attributes in our research work include the number of documents, Bengali sentences, Banglish sentences, total words, total unique words, average number of words, largest text length, smallest text length, and total size in bytes.

All permalinks, user details, dates, and times were eliminated for higher accuracy. The dataset was named Bengali offensive text from the social platform (BHSSP) dataset. Table 1 gives a few samples from our dataset. The data were collected and stored according to the published year, and the sources were also kept track of. The dataset was divided into three sets: a 60% training set, a 20% testing set, and a 20% validation set. The dataset division was carried out to ensure an unbiased evaluation and assess the algorithm's generalization capabilities. The dataset was divided using a stratified sampling approach, where the proportions of each class were maintained in all sets. The training set was used to train the algorithm, the testing set was used to evaluate its performance, and the validation set was utilized for parameter tuning and model selection.

B. DATA SET PRE-PROCESSING

Text processing is prevalent in performing common tasks such as emotion analysis, language translation, sentiment analysis, spam filtering, and many more tasks used in machine learning applications. Punctuation and case conversions are involved in general text processing. Each language has its own distinct syntactic and grammatical structure, as is the case with Bengali, which requires individual command and practice. Thus, a unique method must be applied to specify the Bengali language. The following section presents the text processing techniques used for Bengali text.

1) EMOJI AND EMOTICON CONVERSION

Text processing can be a complex task due to the presence of various types of data, such as emojis, emoticons, and words. Many people usually provide comments on various social platforms using different kinds of emojis and emoticons. These are actually used to express feelings as well as attitudes. However, these feelings may also be offensive too. Therefore, emojis and emoticons play a vital role in detecting offensive comments. Some of the offensive emojis are rooster which is often used to refer to male genital, and screw is often used to harass someone in a sexual way; shit, often used as a curse word, etc. Sometimes a combination of several non-offensive emojis can create an offensive meaning. For example, a combination of some emojis means 'go to hell.' Here particular emojis are not offensive, but when combined, they create an offensive meaning. Unicode has been released, defined by the Unicode organization, and emojis and emoticons have been converted to text format. Python library named Beautifulsoup4 was used to shred all emojis and emotions defined on Unicode Organization's website. This actually determines all the textual presentation for each emoji and emoticon. Subsequently, all emoji texts were converted from English to Bengali. To accomplish this task, the Python Translation Library package was used. Also, offensive and non-offensive emojis were identified manually. To ensure the accuracy and reliability of the manual identification of offensive and non-offensive emojis, we followed a rigorous validation procedure. We formed a team of experts in the field of linguistics and language processing in the AML lab. The team consisted of individuals with extensive knowledge and experience in analyzing linguistic content, including emojis and their associated meanings. The experts reviewed and discussed each emoji's potential offensiveness and agreed on its categorization as offensive or non-offensive. To further validate the categorization, we conducted cross-validation with a separate group of experts. This group consisted of individuals who were not involved in the initial categorization process. They independently reviewed and categorized a subset of emojis based on their own expertise. The results were compared with the initial categorization, and any discrepancies or disagreements were resolved through discussion and consensus.

2) HASHTAG SEGMENTATION

Hashtags (#) refer to using a pound or number symbols to identify a keyword or topic. These are often used on social media, especially Twitter, Instagram, and Facebook, to comment on a post. These hashtags are used as a keyword to discover related posts. This is why it would be effortless to identify offensive or hate text if all the hashtags were considered. Therefore, all hashtags (#) were removed and replaced by space during preprocessing. Then the text was saved for further use. Sometimes, the hashtag is in English, and sometimes, it can also be in Bengali. If not in Bengali, this method cannot identify texts. Therefore, all English hashtag texts will be translated into Bengali format. For example, a widely used hashtag quote is #depression which is in English. The first task was to remove the "#" symbol and replace it with a space. The text was then translated into Bengali. #depression => depression => বিষণ্ণতা.

3) MISCELLANEOUS

Various techniques have been employed to preprocess text data, such as removing punctuation, converting numbers to words, removing accent marks, removing white space, and eliminating stop words. However, identifying stop words in Bengali is more challenging than in English because there is no exact assortment of them in the Bengali format. "হচ্ছে = is, করবে = will do, দেয় = gives, হয়েছিল = had been, থাকে = remains, চান = want, ছিলেন = were, করছে = is doing, পারেন = can, খুব = very, উপরে = above, মধ্যেই = among, আছে = exists, উপর = on, উচিত = appropriate, করিয়ে = by doing, করছে = is doing, কাছে = near, দুটি = two, দেখা = see, ছাড়া = without" are often used as stop words during sentiment analysis. However, these words are also used to write offensive writing. Therefore, these words can not be considered stop when identifying hate speech. The process of identifying stop words in our research involved two main steps: manual identification and the use of regular expressions (regex) to remove punctuation and contextually irrelevant words from the text. During the manual identification step, we carefully examined and curated a list of non-objectionable stop words and irrelevant words specific to the Bengali language. These words were identified based on their common usage and context. We ensured that the identified words were not associated with offensive or objectionable content.

C. FEATURE EXTRACTION

N-gram features are extracted from the Bengali sentences and Banglish sentences, where n ranges from one to three (unigrams, bigrams, and trigrams). These were also weighted using term frequency-inverse document frequency (TF-IDF) values. These values help reduce the impact of lower informative tokens as they emerge repeatedly in the data corpus. The TF-IDF calculation formula of term t present in document d is:

$$tfidf(d,t) = tf(t) * idf(d,t) \tag{1}$$

tf(t) refers to the term frequency of term "t" in document "d." It measures the frequency of occurrence of the term within the document. idf(d, t) represents the inverse document frequency of the term "t" in the entire document collection. It measures the rarity or importance of the term across the Bengali and Banglish sentences collection.

While carrying out the task, both L1 and L2 (Euclidean) normalization of TF-IDF has been considered. L1 normalization can be noted as:

$$U_{norm} = \frac{v}{|v_1| + |v_2| + \ldots + |v_n|} \tag{2}$$

In this equation, n is calculated as the total number of documents. "v" represents a vector containing n elements, denoted as v = [v1, v2, ..., vn]. |v1|, |v2|, ..., |vn| refers to the absolute values of the vector elements. It calculates the magnitude or absolute value of each element in the vector. Likewise, L2 normalization can be defined as:

$$U_{norm} = \frac{v}{\sqrt{v_1^2 + v_2^2 + \ldots + v_n^2}} \tag{3}$$

All of these features are then fed into the proposed model.

D. BERT ARCHITECTURE

Bidirectional Encoder Representations of Transformers is a masked language model for natural language processing (NLP) [38]. BERT has been applied in various natural language processing tasks such as sentiment analysis, named entity recognition, question answering, text classification, and machine translation [39]. It helps the computer to understand the meaning of the obscure language of the text. This is done using sidebar text to set the context. It is based on transformers. BERT represents a deep learning-based bidirectional self-supervised model. The weight between the entities involved in the bidirectional self-supervised model, specifically the encoder and decoder, is dynamically calculated based on their connection. Transformers are utilized as both encoders and decoders with the aim of performing the translation. The BERT model utilized in this study is illustrated in Figure 1.

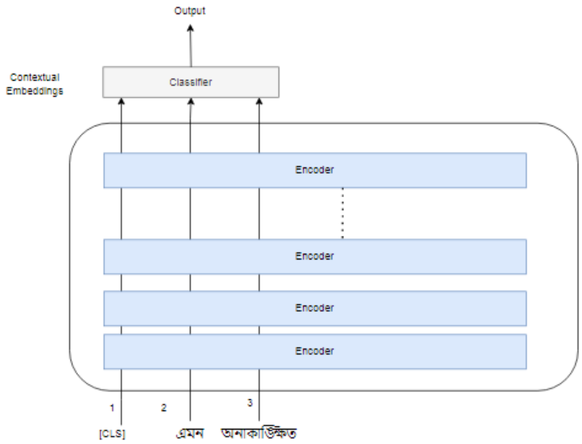


FIGURE 1. Architecture of BERT model for text classification.

The BERT positional embedding technique was employed in this study to extract properties of Bengali and Banglish text. This method enables the representation of word order in the sentence matrix, thereby capturing the contextual relationships between words. Absolute position embedding (AE) and relative position embedding (RE) are two different forms of position embedding. The space element of representation is mapped with AE, whereas the positional distance among words is mapped in RE. The weighted value (Wv), weighted key (Wk), and weighted query (Wq) are considered in the attention head during calculating the BERT transformer [37]. For instance: let's assume that two positions are $x \epsilon N$ and $y \epsilon N$. Then, WV_x is the x positional word vector, while Ex is the embeddings. Besides, the embedding of the relative position is defined as Ex-y. The formula for calculating the query (q), value (v), and key (k) vector for the x positional word is:

$$AE : \begin{bmatrix} q_x \\ k_x \\ v_x \end{bmatrix} = (WV_x + E_x) \times \begin{bmatrix} W_q \\ W_k \\ W_v \end{bmatrix} \tag{4}$$

$$RE : \begin{bmatrix} q_x \\ k_x \\ v_x \end{bmatrix} = (WV_x + E_x) \times \begin{bmatrix} W_q \\ W_k \\ W_v \end{bmatrix} + \begin{bmatrix} 0 \\ E_{x-y} \\ E_{x-y} \end{bmatrix} \tag{5}$$

The final output is the sum of the values from all attention heads (a). The symbol qk^t represents the dot product between the query vector (q) and the transposed key vector (k^t). The symbol d_k represents the key vector. Here, the attention weight depends on qk^t. So,

$$Attention(q, v, k) = softmax(qk^t/\sqrt{d_k})v \tag{6}$$

E. GRU MODEL

Gated Recurrent Units are a commonly used gating mechanism in RNNs that enable connections through a sequence of nodes. The primary objective of GRUs is to facilitate machine-learning tasks that involve memory and clustering. They also help to adjust the input weights of neural networks by solving the basic problems of gradient and explosion in standard RNN while controlling long-term dependence. Moreover, they are designed to update or reset the memory contents properly [40]. The GRU model is similar to Long LSTM in that it has a forget gate. However, GRU lacks an output gate, resulting in fewer parameters than LSTM. Instead, GRU features an update gate and a reset gate, which modify the outputs by controlling the flow of information through the model by using these gates.

As Figure 2 explains, multiple neurons consist of a single input level, and the location of space determines the number of neurons. Equally, the output levels are similar to neurons in the output space.

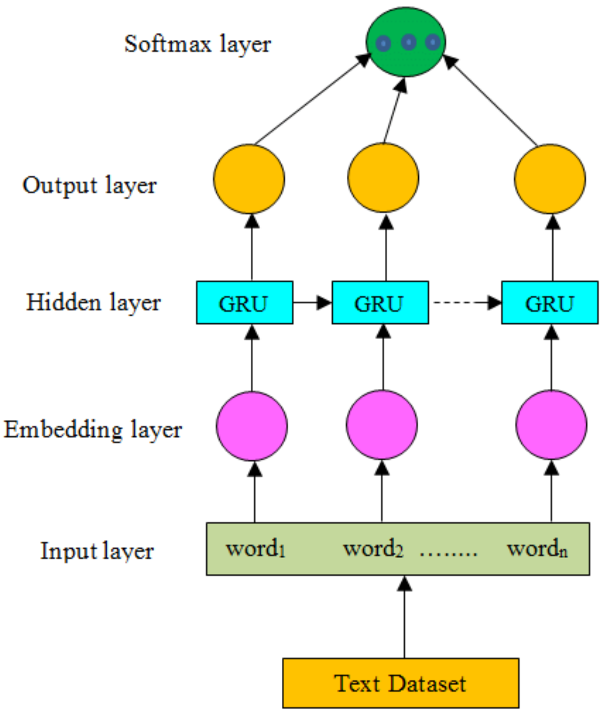


FIGURE 2. Architecture of GRU model for text classification.

At t time, the linear interpolation among the previous activations h^i_{t-1} and candidate activations $\widehat{h^i_t}$ is considered as the activation h^i_t of GRU:

$$h^i_t = (1 - z^i_t)h^i_{t-1} + z^i_t\widehat{h^i_t} \tag{7}$$

The amount that the unit updates its activation is determined by the update gate:

$$z^i_t = sigm(W_z x_t + U_z h_{t-1})^i \tag{8}$$

The calculation of the candidate activation $\widehat{h^i_t}$ is identical to the update gate:

$$\widehat{h^i_t} = tanh(W_z x_t + U_z h_{t-1})^i \tag{9}$$

Here, x_t denotes input vectors, and h_t stands for hidden state vector. Reset gates are presented by r_t. If the reset gate is closed, the unit forgets the previous information. The evaluation of the reset gate is performed using the subsequent equation.

$$r^i_t = sigm(W_r x_t + U_r h_{t-1})^i \tag{10}$$

The amount of previous data thrown is determined by the update gate z. It also determines which new state should be added. In the case of short-term dependencies, the units of reset gate r_t^i are active satisfactorily. But in long-term dependencies, units have an active update gate.

F. PROPOSED G-BERT MODEL

In this study, a combination of BERT and GRU has been used to predict those texts that are offensive where there is an input layer, a BERT embedding layer, GRU layer, and also an output layer. The model focused on the tasks related to word levels, whereas BERT divides words into multiple sub-words, which are considered tokens. Then, attention weights are added to these tokens to emphasize the sub-words. Average weights help to calculate split word attention. The attention amount is maintained as one during these conversions [40].

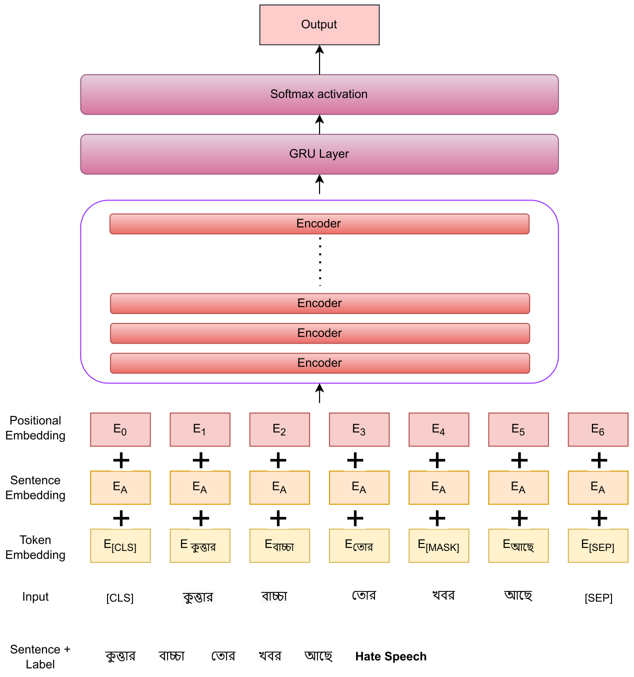


FIGURE 3. Architecture of G-BERT model for text classification.

Figure 3 shows the word embedding of our proposed G-BERT model which is an integral component that helps describe the model structure. To overcome the unidirectional constraints of the masked language model, BERT leveraged a mask language model. Additionally, BERT uses a next-sentence prediction (NSP) task to pre-train text-pair representations. The base model encoder comprises 12 layers and utilizes the position, segment, and token embeddings to represent the input sequences. Token embeddings involve tokenizing and embedding input data using the WordPiece embedding technique, which separates words using special tokens such as SEP and CLS. WordPiece embedding has a vocabulary size of 30K, resulting in an embedding size of V × H, where V is the vocabulary size and H is the hidden size. Furthermore, Segment embedding is used to represent the connection between two sentences.

GRU model contains an update gate and a reset gate. By utilizing these gates, the model modifies the outputs by controlling the flow of information through the model. In addition, this model can help to adjust the input weights of neural networks by solving the vanishing gradient problems. These operations are performed by the following:

Update gate:

$$z_t = \sigma(W_z[Uh_{t-1}), x_t] \tag{11}$$

Reset gate:

$$r_t = \sigma(W_r[Uh_{t-1}), x_t] \tag{12}$$

Candidate hidden state:

$$\widehat{h_t} = tanh(W \cdot [r_t * h_{t-1}, x_t] \tag{13}$$

$$\widehat{h_t} = LRelu(W \cdot [r_t * h_{t-1}, x_t] \tag{14}$$

Hidden state:

$$h_t = (1 - z_t) * h_{t-1}^i + z_t * \widehat{h_t} \tag{15}$$

Here, x_t denotes input vectors, and W, U presents weight matrices. σ stands for logistic sigmoid activation function, h_t denotes the hidden state.

Our proposed G-BERT approach can be summarized as follows:

Algorithm 1 G-BERT Model for Hate Speech Detection in Bengali Texts

```
Input: Input sequence $X = (x_1, x_2, \ldots, x_n)$ of
       Bengali text tokens
Output: Predicted class label $\hat{y}$
Initialize arbitrary values for learning parameters in
  the GRU model (e.g., weight, bias);
Extract contextualized representations
  $H = (h_1, h_2, \ldots, h_n)$ of input tokens using BERT;
for $t \leftarrow 1$ to $n$ do
    Compute GRU cell $c_t = \sigma(W_c x_t + U_c h_{t-1} + b_c)$;
    Compute output $h_t = (1 - c_t) \odot h_{t-1} + c_t \odot$
      $\tanh(W_h x_t + U_h(c_t \odot h_{t-1}) + b_h)$;
end
Compute predicted class distribution
  $\hat{y} = \text{softmax}(W_y h_n + b_y)$;
Enroll optimization method to minimize cross-entropy
  loss between predicted and true class labels;
while model performance is not satisfactory do
    Update GRU model parameters using
      backpropagation and optimization algorithm;
end
return predicted class label $\hat{y}$;
```

G-BERT Model for Hate Speech Detection in Bengali Texts. This algorithm combines the strengths of two deep learning models, BERT and GRU, to detect hate speech in Bengali texts. The input sequence of Bengali text tokens $X = (x_1, x_2, \ldots, x_n)$ is first processed by a pre-trained BERT model to extract contextualized representations $H = (h_1, h_2, \ldots, h_n)$, where each hidden state h_i is a vector in a high-dimensional embedding space. The contextualized representations are then fed into a GRU model, which uses a gated recurrent neural network architecture to perform the classification task. Specifically, at each time step t, the GRU model computes a cell state c_t and an output state h_t as follows:

$$c_t = \sigma(W_c x_t + U_c h_{t-1} + b_c) \tag{16}$$

$$h_t = (1 - c_t) \odot h_{t-1} + c_t \odot \tanh(W_h x_t + U_h(c_t \odot h_{t-1}) + b_h) \tag{17}$$

where σ is the sigmoid activation function, $\odot$ denotes element-wise multiplication, and $W_c, W_h, U_c, U_h, b_c, b_h$ are weight and bias parameters of the GRU model. The output of the GRU model at the final time step is then fed through a softmax activation function to obtain a probability distribution over the two possible classes (hate speech or not). The model is optimized using an optimization algorithm to minimize a cross-entropy loss function, which measures the difference between the predicted and true class labels. The algorithm iteratively updates the model parameters using backpropagation and the optimization algorithm until the model performance is satisfactory. The final predicted class label is returned by the algorithm. This approach provides a powerful and effective solution for detecting hate speech in Bengali texts, with potential applications in social media monitoring and online content moderation.

In general, the hybrid approach works better than manual methods. We combined these two models to get the highest prediction in terms of hybridity and brilliant prediction. This model is mentioned as the G-BERT model. This model's dataset order is reformed according to the results, which are used as the model's input in the following iteration.

IV. EVALUATION

First, the evaluation metrics were clarified. Next, we describe the observed setup. Finally, we present the assessment along with a comprehensive analysis.

A. EVALUATION METRIC

Recall, F1-score, accuracy, and precision evaluation metrics were used based on the confusion matrix. The prediction results to estimate machine learning are encapsulated in a confusion matrix. This includes the issue of deep learning classification and contains four measures: true positive (TP), false positive (FP), true negative (TN) and False Negative (FN). Later, the architectural performance of the system is evaluated by comparing it with the following measurements.

1) PRECISION

The ratio of validly classified positive samples to the total quantity of positive samples (false positive and true positive) is denoted as Precision. It focuses on the efficiency of the model in identifying positive samples. The following formula summarizes the concept of precision.

$$Precision = \frac{TP}{TP + FP} \tag{18}$$

2) RECALL

The ratio of validly classified positive samples to all quantity of samples in the actual class(true positive and false negative) is denoted as Recall which can be presented by the following equation.

$$Recall = \frac{TP}{TP + FN} \tag{19}$$

3) F1-SCORE

The F1-score is a function of precision and recall. The F1-score takes into consideration both false positive and false negative classified samples. Though the concept of accuracy seems easier to grasp on, the F1-score is way more efficient. Finding accuracy will be more effective if the quantity of false positive and false negative samples is equal. But in case of uneven class distribution, the F1-score works better. The formula of the F1-score can be stated as follows:

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{20}$$

The range of the $F1 - score$ is 0 to 1. It can also be stated that the model's efficiency can be determined by analyzing how closer it is to 1.

B. EXPERIMENTAL SETUP

The process of data collection, pre-processing, testing, and evaluation of a deep learning model often requires the use of Python programming language. Python was utilized to complete the neural network architecture in this study, and Keras was employed to apply the deep learning model. The Adam optimization function was implemented to train the dataset model with a total learning rate percentage of 0.1%. Basic mathematical operations were carried out using NumPy. Furthermore, TensorFlow was used to execute the neural network's GPU performance. The dataset was divided into three parts, namely train, test, and validity, with a partition percentage of 60%, 20%, and 20%, respectively. The validation dataset was employed to determine the standard of the deep learning model, and as the concluding dataset, the test data was assessed.

C. RESULT ANALYSIS

We have done several tests to achieve a good result. The experiments are represented below:

1) EMBEDDING AND CLASSIFICATION

In this study, we utilized BERT base uncased, a pre-trained deep learning model from Huggingface's transformers library, and BERT tokenizer to train hate speeches. The BERT base uncased model consists of 12 layers and 110 million parameters. We used frozen and unfrozen weights of the BERT base uncased model for training our proposed model. To prepare the dataset, we used the BERT tokenizer. After generating the word embeddings with BERT we utilized the GRU layer for classification following a softmax layer.

We conducted several experiments using a range of parameter values to identify the optimal performance configuration for the G-BERT model. These experiments involved adjusting the batch size, the number of epochs, the learning rate, and the maximum sequence length of the input data. We trained the model using different combinations of these parameter values and evaluated its performance on the validation dataset.

TABLE 2. Parameter values for the G-BERT model.

Parameter	Value
Batch Size	32
Epochs	50
Learning Rate	2e-5
Max Sequence Length	128

We adjusted various parameters in our model by selecting specific values. For instance, we varied the maximum sequence length from 128 to 50 and altered the learning rate from 1×10^{-5} to 2×10^{-3}, and the batch size from 5 to 64. The parameter values listed in the table were adopted as they gave us the best performance. The batch size refers to the number of samples that the algorithm processes at each iteration. It is an essential parameter that can affect the performance of the classifier. As shown in Figure 4, the batch-wise and epoch-wise approach significantly influences the classifier's performance, which can improve GPU performance.

Based on our evaluation results, we selected the parameter values that provided the highest classification accuracy for the final training of the G-BERT model. Our proposed G-BERT model was assigned specific parameter values, as shown in Table 2. It is worth noting that the parameter values are interconnected, and changes in one parameter value can significantly impact the model's overall performance. Our findings demonstrate the importance of selecting appropriate parameter values in achieving high classification accuracy for natural language processing tasks. Our proposed model's performance highlights the effectiveness of BERT-based models for identifying hate speeches. These insights may have implications for future studies in this field and may contribute to advancing natural language processing techniques for hate speech detection.

2) EVALUATION AND COMPARISON

To analyze the efficiency of our proposed G-BERT model for the recognition process, we compared our G-BERT model with other models using a similar test environment and data set.

Figure 5 presents the accuracy, precision, recall, and F1-score for the different existing machine learning models and the proposed G-BERT model. Among the existing models, Bangla BERT achieved an accuracy of 88.77%, LSTM-BERT achieved an accuracy of 89.52%, AdaBoost-BERT achieved an accuracy of 92.26%, and L-BOOST achieved an accuracy of 95.10%. The Random forest (RF) model obtained the lowest accuracy with an accuracy of 80.06%.

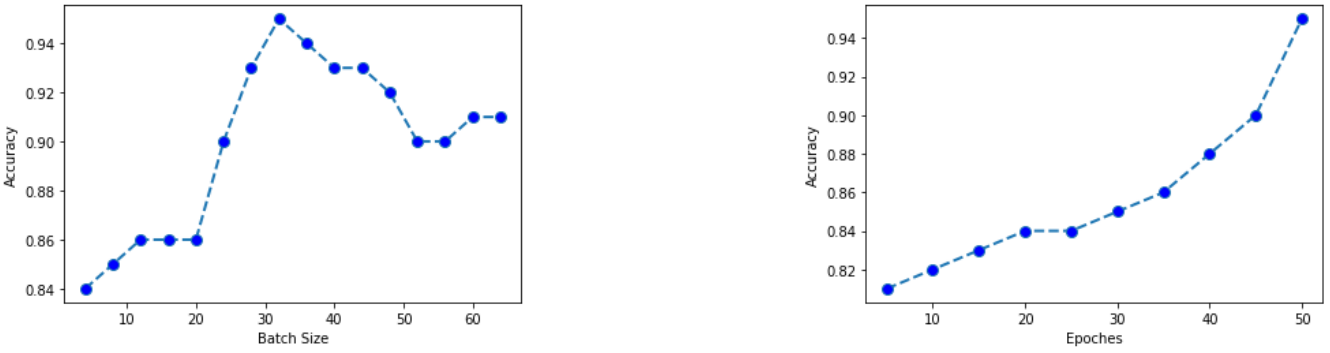
However, the proposed G-BERT model outperformed all the existing models, achieving the highest accuracy of 95.56%. In terms of the F1-score, G-BERT also outperformed all the other models. Specifically, the F1-score for G-BERT was 92.15%, which was the highest among all the models. Moreover, G-BERT's precision and recall values were also very high, with 95.07% and 93.63%, respectively.

Table 3 displays contemporary Bengali hate speech classification research using a different dataset called Bengali Hate Speech Dataset [41], which is publicly available. Karim et al. [41] utilized the Bengali Hate Speech Dataset and employed various models, including BengFastText with GBT and BengFastText with RF, achieving precision, recall, and F1-Score values of 0.842, 0.845, and 0.845, and 0.861, 0.857, and 0.862, respectively. They also applied BengFastText with MC-LSTM and performed MAE on top-3 models (GBT, RF, MC-LSTM), achieving precision, recall, and F1-Score values of 0.881, 0.883, 0.882, and 0.896, 0.894, and 0.891, respectively. In a similar vein, Karim et al. [42] employed Conv-LSTM, Bangla BERT, XLM-RoBERTa, and mBERT-uncased on the Extended Bengali Hate Speech Dataset, achieving precision, recall, and F1-Score values of 0.79, 0.78, and 0.78, 0.80, 0.79, and 0.79, 0.82, 0.82, and 0.82, and 0.81, 0.81, and 0.81, respectively. Additionally, our proposed G-BERT model achieved precision, recall, and F1-Score values of 0.91, 0.91, and 0.90, respectively, using Bengali Hate Speech Dataset.

In summary, based on the results shown in Figure 5 and Table 3, the G-BERT model proposed in this study appears to be the most effective for the given task. It achieved the highest accuracy, F1-score, and precision/recall values compared to the other models. Our proposed model showed significant improvements in classification accuracy compared to other state-of-the-art models. The findings of this study could potentially advance the field of deep learning models and natural language processing in detecting hate speech across different domains. By utilizing the proposed G-BERT model, the accuracy and efficiency of hate speech detection in Bengali text on social media and other platforms can be improved.

V. DISCUSSION AND FUTURE RESEARCH

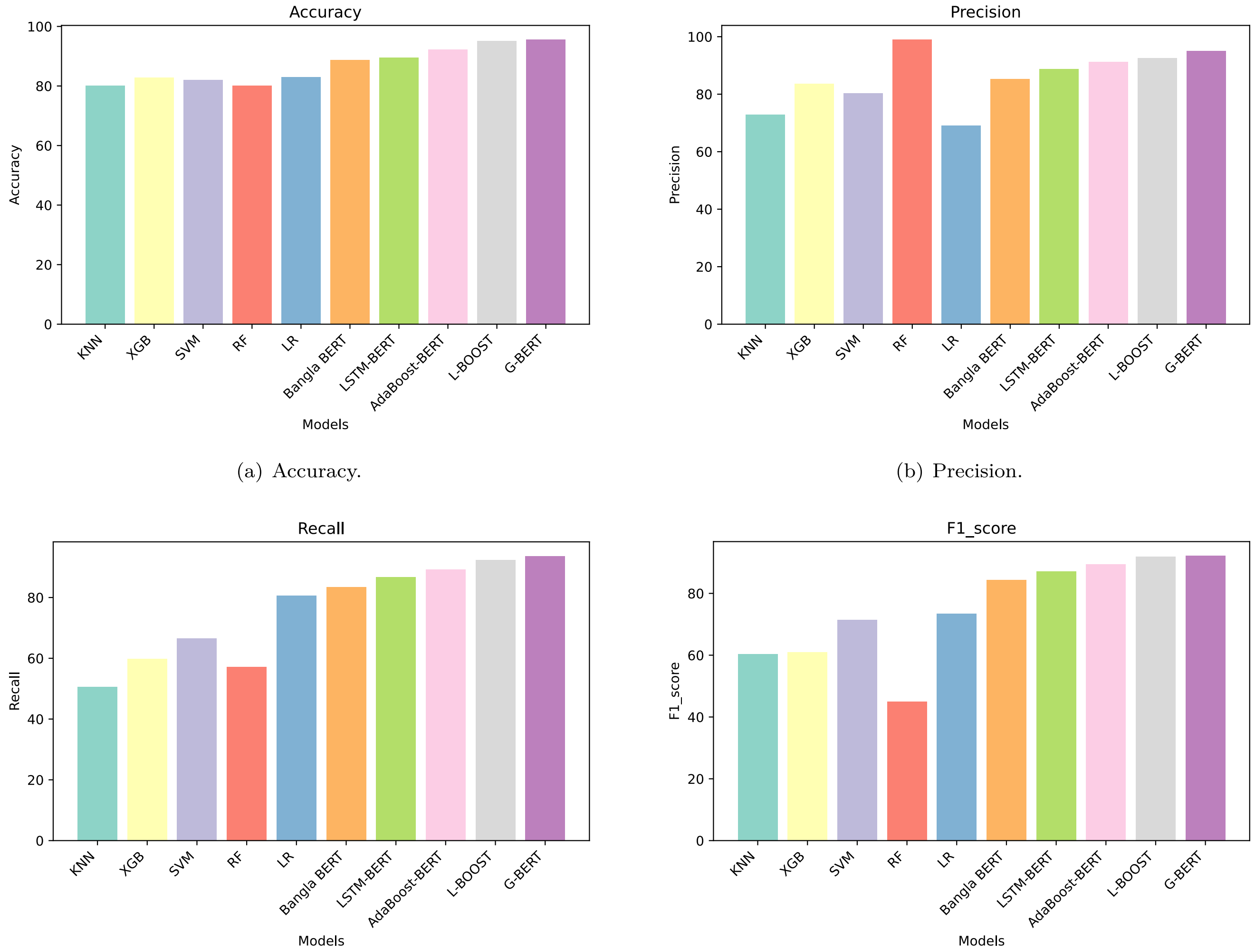
In this study, we proposed a novel method for identifying hate speech in Bengali texts on social media platforms called G-BERT. The proposed method combines the Bidirectional Encoder Representations from Transformers architecture with a Gated Recurrent Units model and Softmax



(a) Accuracy in terms of batch size.

(b) Accuracy in terms of epochs.

FIGURE 4. Accuracy of G-BERT Model regarding epoch and batch size.



(a) Accuracy.

(b) Precision.

(c) Recall.

(d) F1-score.

FIGURE 5. Comparison of the evaluation metrics of the existing models with G-BERT. The X-axis represents the different models considered in the evaluation, while the Y-axis represents the following evaluation metrics: a) accuracy, b) precision, c) recall, and d) F1-score.

activation function. This approach achieved an accuracy of 95.56%, precision of 95.07%, recall of 93.63%, and F1-score of 92.15%, outperforming all other classification algorithms tested. Our results indicate that G-BERT effectively detects

TABLE 3. Performance comparison on bengali hate speech dataset [41].

Dataset	Ref.	Model	Precision	Recall	F1-Score
Bengali Hate Speech Dataset	Karim et al. [41]	BengFastText with GBT	0.842	0.845	0.845
		BengFastText with RF	0.861	0.857	0.862
		BengFastText with MC-LSTM	0.881	0.883	0.882
		BengFastText with MAE	0.896	0.894	0.891
	Karim et al. [42]	Conv-LSTM	0.79	0.78	0.78
		Bangla BERT	0.80	0.79	0.79
		XLM-RoBERTa	0.82	0.82	0.82
		mBERT-uncased	0.81	0.81	0.81
	Proposed Approach	G-BERT	0.91	0.91	0.90

GBT= Gradient Boosted Trees, MC-LSTM = Multichannel Convolutional LSTM, MAE = Model Averaging Ensemble.

hate speech in Bengali texts on social media platforms. The performance of our proposed model is particularly significant given the challenges of identifying hate speech in Bengali, which has a unique script and grammar. Most earlier studies have concentrated on identifying hate speech in the English language, highlighting the significance of our research in natural language processing for low-resource languages.

The G-BERT method is promising because it utilizes the BERT architecture to extract Bengali text properties and a GRU model with a Softmax activation function to classify hate speech. BERT has demonstrated its effectiveness in various natural language processing tasks, including but not limited to sentiment analysis and text classification, as it can comprehend a sentence's context. Moreover, the GRU model is well-suited for sequential data processing, making it an excellent choice for text classification tasks. Compared to other methods, such as CNN, BERT is better at understanding the context of a sentence, which is critical in identifying hate speech. CNNs are typically used for image classification tasks, whereas BERT is designed for natural language processing tasks, making it a better choice for text classification tasks. Moreover, GRUs have demonstrated similar or better performance in text classification tasks compared to other recurrent neural networks such as LSTM and RNN. Additionally, GRUs are faster, more memory-efficient, and require fewer parameters than these other models. Moreover, the GRU model has a gating mechanism that allows it to capture relevant information and ignore irrelevant information, making it well-suited for sequential data processing. By combining BERT and GRU, the G-BERT model can utilize the strengths of both the models. BERT can extract contextual information from Bengali text, whereas the GRU model can process this information sequentially. This combination results in a highly effective model that outperforms the other classification algorithms.

Future research can expand on this study by incorporating other languages and evaluating the performance of the G-BERT model on larger datasets. In addition, research can be conducted to assess the effectiveness of the G-BERT model in real-time situations, such as identifying hate speech in live chats and video streams. Another possible direction is to investigate the effectiveness of G-BERT in identifying other types of offensive language, such as sexist or racist languages. Finally, the G-BERT model can be further optimized for better performance, such as by fine-tuning the model on domain-specific data or incorporating other types of neural networks.

VI. CONCLUSION

Hate speech is a significant issue in social media platforms, and it is crucial to detect such speech automatically to create a safer environment. In this study, we proposed a G-BERT model that can detect hate speech in Bengali texts. We collected a large dataset of 16,800 posts and comments from various Bengali websites, social media platforms, and blogs. The proposed model combines the GRU and BERT architectures, which showed outstanding performance compared to other baseline models, with an accuracy of 95.56% and precision, recall, and F1-score of 95.07%, 93.63%, and 92.15%, respectively. Our study contributes to the field of Bengali language processing, where limited research has been conducted. In the future, we aim to expand our research to Banglish texts to detect malicious speech more accurately. In summary, our proposed model has the potential to effectively identify hate speech in Bengali texts, thus helping to create a safer online community.

ACKNOWLEDGMENT

The authors would like to thank the Advanced Machine Intelligence Research Lab (AMIRL) for supervision and resources.

REFERENCES

[1] K. Subaramaniam, R. Kolandaisamy, A. B. Jalil, and I. Kolandaisamy, "Cyberbullying challenges on society: A review," *J. Positive School Psychol.*, vol. 6, no. 2, pp. 2174–2184, 2022.

[2] G. W. Giumetti and R. M. Kowalski, "Cyberbullying via social media and well-being," *Current Opinion Psychol.*, vol. 45, Jun. 2022, Art. no. 101314.

[3] Z. Al-Makhadmeh and A. Tolba, "Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach," *Computing*, vol. 102, no. 2, pp. 501–522, Feb. 2020.

[4] A. S. Saksesi, M. Nasrun, and C. Setianingsih, "Analysis text of hate speech detection using recurrent neural network," in *Proc. Int. Conf. Control, Electron., Renew. Energy Commun. (ICCEREC)*, Dec. 2018, pp. 242–248.

[5] M. Mondal, L. A. Silva, and F. Benevenuto, "A measurement study of hate speech in social media," in *Proc. 28th ACM Conf. Hypertext Social Media*, Jul. 2017, pp. 85–94.

[6] A. F. Colladon and P. A. Gloor, "Measuring the impact of spammers on e-mail and Twitter networks," *Int. J. Inf. Manag.*, vol. 48, pp. 254–262, Oct. 2019.

[7] S. Bauman, V. M. Perry, and S. Wachs, "The rising threat of cyberhate for young people around the globe," in *Child and Adolescent Online Risk Exposure*. Jan. 2021, ch. 8, pp. 149–175.

[8] J. A. Garcia-Diaz, S. M. Jimenez-Zafra, M. A. Garcia-Cumbreras, and R. Valencia-Garcia, "Evaluating feature combination strategies for hate-speech detection in Spanish using linguistic features and transformers," *Complex Intell. Syst.*, vol. 9, no. 3, pp. 2893–2914, 2022.

[9] A. B. Pawar, P. Gawali, M. Gite, M. A. Jawale, and P. William, "Challenges for hate speech recognition system: Approach based on solution," in *Proc. Int. Conf. Sustain. Comput. Data Commun. Syst. (ICSCDS)*, Apr. 2022, pp. 699–704.

[10] A. K. Das, A. Al Asif, A. Paul, and M. N. Hossain, "Bangla hate speech detection on social media using attention-based recurrent neural network," *J. Intell. Syst.*, vol. 30, no. 1, pp. 578–591, Apr. 2021.

[11] M. Atikuzzaman and S. Akter, "Hate speech in social media: Personal experiences and perceptions of university students in Bangladesh," *Global Knowl., Memory Commun.*, Apr. 2022.

[12] Md. R. Karim, S. K. Dey, T. Islam, S. Sarker, M. H. Menon, K. Hossain, Md. A. Hossain, and S. Decker, "DeepHateExplainer: Explainable hate speech detection in under-resourced Bengali language," in *Proc. IEEE 8th Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2021, pp. 1–10.

[13] S. Banerjee, S. Ghosh, and A. N. Chowdhury, "Expression of depression in Bengali language and culture," Roy. College Psychiatrists, DocsBay, TPSIG Newslett., Tech. Rep., 2016.

[14] M. A. Awal, M. S. Rahman, and J. Rabbi, "Detecting abusive comments in discussion threads using naive Bayes," in *Proc. Int. Conf. Innov. Sci., Eng. Technol. (ICISET)*, 2018, pp. 163–167.

[15] W. S. S. Fernando, R. Weerasinghe, and E. R. A. D. Bandara, "Sinhala hate speech detection in social media using machine learning and deep learning," in *Proc. 22nd Int. Conf. Adv. ICT Emerg. Regions (ICTer)*, Nov. 2022, pp. 166–171.

[16] M. Almaliki, A. M. Almars, I. Gad, and E.-S. Atlam, "ABMM: Arabic BERT-mini model for hate-speech detection on social media," *Electronics*, vol. 12, no. 4, p. 1048, Feb. 2023.

[17] M. A. H. Wadud, M. M. Kabir, M. F. Mridha, M. A. Ali, M. A. Hamid, and M. M. Monowar, "How can we manage offensive text in social media—A text classification approach using LSTM-BOOST," *Int. J. Inf. Manag. Data Insights*, vol. 2, no. 2, Nov. 2022, Art. no. 100095.

[18] N. Shakeel and R. K. Dwivedi, "Performance analysis of supervised machine learning algorithms for detection of cyberbullying in Twitter," in *Intelligent Sustainable Systems*. Berlin, Germany: Springer, 2022, pp. 381–401.

[19] K. Nugroho, E. Noersasongko, A. Z. Fanani, and R. S. Basuki, "Improving random forest method to detect hatespeech and offensive word," in *Proc. Int. Conf. Inf. Commun. Technol. (ICOIACT)*, Jul. 2019, pp. 514–518.

[20] I. Priyadarshini, S. Sahu, and R. Kumar, "A transfer learning approach for detecting offensive and hate speech on social media platforms," *Multimedia Tools Appl.*, vol. 82, pp. 27473–27499, Feb. 2023.

[21] L. Ketsbaia, B. Issac, and X. Chen, "Detection of hate tweets using machine learning and deep learning," in *Proc. IEEE 19th Int. Conf. Trust, Secur. Privacy Comput. Commun. (TrustCom)*, Dec. 2020, pp. 751–758.

[22] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on Twitter using a convolution-GRU based deep neural network," in *Proc. Eur. Semantic Web Conf.* Cham, Switzerland: Springer, 2018, pp. 745–760.

[23] M. U. S. Khan, A. Abbas, A. Rehman, and R. Nawaz, "HateClassify: A service framework for hate speech identification on social media," *IEEE Internet Comput.*, vol. 25, no. 1, pp. 40–49, Jan. 2021.

[24] F. Alkomah and X. Ma, "A literature review of textual hate speech detection methods and datasets," *Information*, vol. 13, no. 6, p. 273, May 2022.

[25] H. Sahi, Y. Kiliç, and R. B. Saglam, "Automated detection of hate speech towards woman on Twitter," in *Proc. 3rd Int. Conf. Comput. Sci. Eng. (UBMK)*, Sep. 2018, pp. 533–536.

[26] A. Bhat, S. Adhikari, K. Jha, and H. B. Sadat, "Deep learning based hybrid word representation for detection of hate speech," in *Proc. 2nd Int. Conf. Advance Comput. Innov. Technol. Eng. (ICACITE)*, Apr. 2022, pp. 2128–2133.

[27] M. Jahan, I. Ahamed, Md. R. Bishwas, and S. Shatabda, "Abusive comments detection in Bangla-English code-mixed and transliterated text," in *Proc. 2nd Int. Conf. Innov. Eng. Technol. (ICIET)*, Dec. 2019, pp. 1–6.

[28] P. William, R. Gade, R. E. Chaudhari, A. B. Pawar, and M. A. Jawale, "Machine learning based automatic hate speech recognition system," in *Proc. Int. Conf. Sustain. Comput. Data Commun. Syst. (ICSCDS)*, Apr. 2022, pp. 315–318.

[29] P. K. Roy, A. K. Tripathy, T. K. Das, and X.-Z. Gao, "A framework for hate speech detection using deep convolutional neural network," *IEEE Access*, vol. 8, pp. 204951–204962, 2020.

[30] J. Melton, A. Bagavathi, and S. Krishnan, "DeL-haTE: A deep learning tunable ensemble for hate speech detection," in *Proc. 19th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2020, pp. 1015–1022.

[31] S. Khan, M. Fazil, V. K. Sejwal, M. A. Alshara, R. M. Alotaibi, A. Kamal, and A. R. Baig, "BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 7, pp. 4335–4344, Jul. 2022.

[32] R. Ali, U. Farooq, U. Arshad, W. Shahzad, and M. O. Beg, "Hate speech detection on Twitter using transfer learning," *Comput. Speech Lang.*, vol. 74, Jul. 2022, Art. no. 101365.

[33] A. Chhabra and D. K. Vishwakarma, "A literature survey on multimodal and multilingual automatic hate speech identification," *Multimedia Syst.*, vol. 29, pp. 1203–1230, Jan. 2023.

[34] A. Rana and S. Jha, "Emotion based hate speech detection using multimodal learning," 2022, *arXiv:2202.06218*.

[35] A. G. D'Sa, I. Illina, and D. Fohr, "BERT and fasttext embeddings for automatic detection of toxic speech," in *Proc. Int. Multi-Conf., 'Org. Knowl. Adv. Technologie' (OCTA)*, Feb. 2020, pp. 1–5.

[36] H.-S. Lee, H.-R. Lee, J.-U. Park, and Y.-S. Han, "An abusive text detection system based on enhanced abusive and non-abusive word lists," *Decis. Support Syst.*, vol. 113, pp. 22–31, Sep. 2018.

[37] M. F. Mridha, M. A. H. Wadud, M. A. Hamid, M. M. Monowar, M. Abdullah-Al-Wadud, and A. Alamri, "L-boost: Identifying offensive texts from social media post in Bengali," *IEEE Access*, vol. 9, pp. 164681–164699, 2021.

[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[39] S. Bano and S. Khalid, "BERT-based extractive text summarization of scholarly articles: A novel architecture," in *Proc. Int. Conf. Artif. Intell. Things (ICAIoT)*, Dec. 2022, pp. 1–5.

[40] M. Zulqarnain, R. Ghazali, Y. M. M. Hassim, and M. Rehan, "Text classification based on gated recurrent unit combines with support vector machine," *Int. J. Electr. Comput. Eng. (IJECE)*, vol. 10, no. 4, p. 3734, Aug. 2020.

[41] Md. 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 *Proc. IEEE 7th Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2020, pp. 390–399.

[42] M. R. Karim, S. K. Dey, T. Islam, M. Shajalal, and B. R. Chakravarthi, "Multimodal hate speech detection from Bengali memes and texts," 2022, *arXiv:2204.10196*.

ASHFIA JANNAT KEYA received the B.Sc. degree in computer science and engineering from the Bangladesh University of Business and Technology (BUBT), in 2021. She is currently pursuing the master's degree in a prestigious institution with the Military Institute of Science and Technology (MIST). She is a dedicated educator and has been a Lecturer with the Department of CSE, BUBT, in addition to her research work as a Researcher with the Advanced Machine Learning Laboratory. With her strong academic background, passion for research, and commitment to education, she is devoted to advancing the field of computer science and engineering and preparing the next generation of professionals for success in the field. She has gained experience working with various programming languages, such as C++, Python, and libraries, such as Keras, TensorFlow, Sklearn, NumPy, Pandas, and Matplotlib. In addition, she has contributed to the scientific community by publishing conference paper in ICSCT conference as well as in reputed journals, such as IEEE Access, *Cancers*, and *Applied Sciences*. Her research interests include deep learning, natural language processing, and computer vision.

MD. MOHSIN KABIR received the Bachelor of Science degree in CSE from the Bangladesh University of Business and Technology (BUBT), Bangladesh, in 2021. He is currently pursuing the joint master's degree in intelligent field robotics systems (IFRoS) with the University of Girona, Spain, and Eötvös Loránd University, Hungary, funded by the Erasmus Mundus Scholarship (2022–2024). He was a Research Assistant with BUBT and a Researcher with the Advanced Machine Learning Laboratory. Also, he has had the privilege of collaborating with several prominent research laboratories around the globe, including the Computer Vision and Pattern Recognition Laboratory, University of Asia Pacific, Bangladesh, and the Database System Laboratory, The University of Aizu, Japan. In addition to his studies, he holds a position as a Lecturer with the Department of Computer Science and Engineering, BUBT (study-leave). With an extensive research background, he has authored over ten articles in high-impact journals, such as IEEE Access, *Sensors*, *Computer Systems Science and Engineering*, *Biology*, *Mathematics*, and more. In addition, he has contributed to the scientific community by publishing over ten conference papers and actively participating in well-established conferences, including IEEE HONET, ICCIT, ICIEV, icIVPR, DASA, BIM, and ICSCT. Moreover, some of his research work has been published as a chapter in a few esteemed books related to machine learning and AI. His research interests include artificial intelligence, machine learning, deep learning, computer vision, the IoT, and robotics.

NUSRAT JAHAN SHAMMEY received the B.Sc. degree in computer science and engineering from the Bangladesh University of Business and Technology (BUBT), in 2023. She has majored in artificial intelligence (AI). Her research interests include artificial intelligence, machine learning, deep learning, computer vision, and the IoT-based projects.

M. F. MRIDHA (Senior Member, IEEE) received the Ph.D. degree in AI/ML from Jahangirnagar University, in 2017. He is currently an Associate Professor with the Department of Computer Science, American International University-Bangladesh (AIUB). Before that, he was an Associate Professor and the Chairperson with the Department of CSE, Bangladesh University of Business and Technology. He was the CSE Department Faculty Member with the University of Asia Pacific and as the Graduate Head, from 2012 to 2019. For more than ten years, he has been with the master's and bachelor's students as a supervisor of their thesis work. His research experience, within both academia and industry, results in over 120 journals and conference publications. His research work contributed to the reputed journal of *Scientific Reports* (Nature), *Knowledge-Based Systems*, *Artificial Intelligence Review*, IEEE Access, *Sensors*, *Cancers*, and *Applied Sciences*. His research interests include artificial intelligence (AI), machine learning, deep learning, natural language processing (NLP), and big data analysis. He has served as a program committee member for several international conferences/workshops. He served as an Associate Editor for several journals, including *PLOS ONE*. He has served as a Reviewer for reputed journals and IEEE conferences, such as HONET, ICIEV, ICCIT, IJCCI, ICAEE, ICCAIE, ICSIPA, SCORED, ISIEA, APACE, ICOS, ISCAIE, BEIAC, ISWTA, IC3e, ISWTA, CoAST, icIVPR, ICSCT, 3ICT, and DATA 2021.

MD. RASHEDUL ISLAM (Senior Member, IEEE)received the B.Sc. degree in computer science and engineering from the University of Rajshahi, Rajshahi, Bangladesh, in 2006, the M.Sc. degree in informatics from Högskolan i Borås (University of Boras), Boras, Sweden, in 2011, and the Ph.D. degree in electrical, electronic, and computer engineering from the University of Ulsan, Ulsan, South Korea, in 2016. He was a Senior Architect with the Research and Development Department, Exvision Corporation, Tokyo, Japan; a Visiting Researcher (Postdoctoral Researcher) with the School of Computer Science and Engineering, The University of Aizu, Japan; a Graduate Research Assistant with the Embedded System Laboratory, University of Ulsan; an Assistant Professor with the Department of Computer Science and Engineering, University of Asia Pacific (UAP), Dhaka, Bangladesh; and a Lecturer with the Department of Computer Science and Engineering, Leading University, Sylhet, Bangladesh. He is currently a Chief Researcher of computer vision and AI with Chowagiken Corporation, Japan, and also an Associate Professor (on leave) with the Department of Computer Science and Engineering, UAP. Also, he has a good experience in professional IT system analysis and development. His research interests include machine learning, signal and image processing, HCI, health informatics, bearing fault diagnosis, and others. He is also a PC member of several international conferences. He is a member of the IEEE Computer Society and the IEEE Computational Intelligence Society. He has also served as the Secretary for the Organizing Committee of the 19th International Conference on Computer and Information Technology 2017 (ICCIT 2017), an Organizing Chair for the Organizing Committee of the ACM-ICPC Dhaka Regional Site 2017, the Head of the Self-Assessment Committee (SAC) for the Department of CSE under IQAC, University of Asia Pacific, a Coordinator for the MCSE Program, Department of CSE, UAP, a Convener for the Software and Hardware Club, Department of CSE, UAP, a Coordinator for the Admission Committee, Department of CSE, UAP, and a Treasurer for the Bangladesh Advanced Computing Society. He is a Reviewer of several journals, such as the IEEE Transactions on Industrial Electronics, IEEE Access, *Applied Science*, *Multimedia Tools and Applications*, *Cluster Computing*, *Shock and Vibration*, *Journal of Information Processing Systems*, and others.

YUTAKA WATANOBE (Member, IEEE) received the master's and Ph.D. degrees from The University of Aizu, Japan, in 2004 and 2007, respectively. He was a Research Fellow with the Japan Society for the Promotion of Science (JSPS), The University of Aizu, in 2007. He is currently a Senior Associate Professor with The University of Aizu. His research interests include visual programming, smart learning, data mining, and robotics.