MIJST

MIST International Journal of Science and Technology

Analyzing the Performance of Deep Learning Models for Detecting Hate Speech on Social Media Platforms

Md Ariful Islam Arif[1], Md. Mahbubur Rahman[*2], Md. Golam Rabiul Alam[3], and M. Akhtaruzzaman[4]

[1,2,4]Department of Computer Science and Engineering, Military Institute of Science and Technology, Dhaka-1216, Bangladesh
[3]Department of Computer Science and Engineering, BRAC University, Merul Badda, Dhaka-1212, Bangladesh

emails: *[1]arif.arifulpbcn@gmail.com, [*2]mahbub@cse.mist.ac.bd, [3]rabiul.alam@bracu.ac.bd, and [4]akhter900@gmail.com*

ARTICLE INFO

Article History:

Received: 05th June 2024
Revised: 11th October 2024
Accepted: 19th November 2024
Published: 26th December 2024

Keywords:

Social media platform
Hate speech detection
Deep learning models
Word embedding
LSTM

ABSTRACT

Currently social media and online platforms have become a major source of cyberbullying and hate speech. It is currently affecting people and communities in harmful ways. Hate speech on social media is rising in Bangladesh and it is creating a need for effective tools to prevent and detect these incidents. This study introduces a deep learning model to mitigate this issue of identifying hate speech in text using three types of word embedding methods: Word2Vec, FastText, and BERT. The text data was labeled to mark hate speech and non-hate speech content. After that, these texts are preprocessed by removing punctuation and symbols to help improve model accuracy. Five deep learning models Bi-GRU-LSTM-CNN, Bi-LSTM, CNN, LSTM, and XGBoost were trained to classify the text as hate speech or non-hate speech. The study found that the LSTM model accomplished the highest accuracy at 95.66% with the Word2Vec embedding method, while CNN reached 87.70% with FastText embeddings. Word2Vec is effective for capturing word meanings in general text classification. FastText works well with rare words and languages that have complex word forms. These findings help advance effective hate speech detection techniques. It could promote more respectful and inclusive interactions on social media. This proposed deep-learning model can help stop cyberbullying and hate speech on social media.

1. INTRODUCTION

Artificial intelligence (AI) is an advanced and commonly used technology. Computers and machines cannot think and make decisions on their own like human do. AI technologies help the computer or machine make decisions based on their previously used or stored data. In this sector of computer science, these new technologies are engaged in preparing machines that can work normally to execute numerous tasks that need human intelligence. These kinds of tasks include recognizing speech or text, making decisions, classification, and recognizing different objects. Machine Learning (ML) is a subsection of AI. This allows machines to learn with a given data source without requiring precise orders and instructions. ML algorithms can find similar patterns and connections by learning from provided data sources. This learning allows the machines to make decisions and predictions. Deep Learning (DL) is a special type of ML process that implements algorithms designed with a combination neural network. These neural networks are structured similarly to the human brain. DL has recently exhibited massive performance in different aspects of applications and usages like, image classification, speech recognition, text classification, natural language processing, object detection, and more. These techniques are practically used in many sectors. AI can analyze medical images and predict disease in healthcare. ML technologies are used in credit scoring and fraud detection in the financial section. AI and ML have combined applications in miscellanies industries like transportation, manufacturing, and retail sectors. In these sectors, AI and ML can be used to optimize operations and expand productivity. The fourth industrial revolution (4IR) is noticeable by the union of social, digital, physical, and biotic systems.

Different DL and ML methods are applied in the context of 4IR to change productions such as manufacturing, healthcare, transportation, and others (Javaid *et al.*, 2022). ML and DL strategies are mostly used for object detection, face recognition, face identification, and text classification. These kinds of works are involved with assigning predefined categories to a given text document. ML algorithms such as Naive Bayes, Decision trees, Logistic regressions, and SVMs are mostly used for text classification and hate speech detection. These algorithms and classifiers need labeled data to train the model properly. A set of labels is allocated to each document, and the model learns from this provided labeled data to classify new and unlabeled documents. DL algorithm uses neural networks. This neural network can learn from huge amounts of datasets without explicit programming. DL models like CNNs, LSTMs, Bi-LSTMs and RNNs have shown amazing success rate in text classification and hate speech detection. CNN algorithm is useful for extracting features from the text, while RNNs can capture the sequence of words from given sentence or document. ML and DL methodologies for text classification have many real-world applications, such as, sentiment analysis, hate words filtering, topic classification, word classification, and content moderation (Ahmed *et al.*, 2023). These methodologies have capabilities to enhance the accuracy and performance of the aforementioned tasks. It can also utilize time and resources for businesses and organizations.

Now, social media is a large part of people daily activities. In recent times, the most popular online platforms are Facebook, Twitter, and Instagram. In 2024 Facebook is still continuing its leading spot in the social media landscape. Facebook has many features and tools. Users from any part of the world can connect with friends and family, share different contents and discover new interests as well. Facebook and others social media are also expanding their activities. The inclusion of e-commerce, gaming, augmented reality and virtual reality enhance the reachability. It had more than 2.8 billion active users till 2022. The popularity of Facebook makes it a powerful tool for business. This can enhance the reachability and the engagement with their target audience easily. The platform includes targeted advertising as well as a whole set of analytics. These are helping the companies to come across with their campaigns at a time; they measure how effectively your business mastermind is executed. Facebook has been at the center of criticism over user data and privacy. This has led it to enforce tighter rules on data protection. It also brought new tools to provide more insight about data.

Hate speech is a communication that demeans any individual on the basis of their culture, religion, gender or other distinctive. As hate speech on social media spreads incredibly fast and can destroy an individual, society or even humanity on large scale, so currently this might be a major concern. Hate speech has a lot of influence on social media. It tends to spread rumor and false news, pull people into negative online spaces, and also promote violence and discrimination in the society. Hate speech can in fact cause harm to those upon whom it is directed, manifesting itself into emotional consequences such as anxiety and depression amongst a host of other negative results. It is the accountability of social platforms to fight hate speech and make their users feel safe. It can be done by having a community guidelines-based policy, rules or even amongst obviously harmful content may also include hate speech. The social media platforms could as well make investments in technology with human moderators to detect and remove hate speech faster.

Hate speech on social media platforms is also a major concern in Bangladesh like in other countries. The rise of hate speech on social media and online platforms such as Facebook, Twitter, Instagram, and YouTube have contributed to a growth in online abuse against minorities like the Rohingya, Hindus and Christians. Hate speech on social media targeting message in Bangladesh, has turned fatal occasionally. Social media has proliferated over the past decade and wreaked havoc across society, dividing us further entrenched in our ideological camps spreading misinformation. This has resulted in a toxic online environment, particularly for those who are subject to hate speech and may suffer abusive language that could harm their mental health. Facebook and YouTube have also been used to spread fake news, which led to communal violence in Bangladesh (Deutsche Welle, 2019). Certain religious posts and comments made the conflict between these two groups visible at times on Facebook. Social media interactions on celebs in Bangladesh comment section of Chanchal Chowdhury (The Business Standard, 2021). There are some acute differences and issues, which the social media behaviors have reflected off lately. Bangladeshi civil society, journalists and normal citizens indicate that they are taking action to tamp down on hate speech in social media. This includes through reporting hate speech to the platforms, but also having constructive dialogue that contributes to greater mutual understanding and respect.

In the recent past, it is significant to realize the growing trend of hate speech in Bangladesh. In the context of social media hate speech has become one of the most concerning social issues in Bangladesh. Hate speech can also lead to social conflict. As the sharing and usage of hate speech have increased, the area of research focused on automated detection of such content remains notably scarce, especially with regard to Bengali and other local languages. The majority of existing studies tend to focus on more dominant international languages disregarding the fact that language and culture of Bangle can be a barrier to meaningful detection approaches. It is imperative that these gaps be addressed in order to build effective algorithms which are designed and developed for local usage so that they do not negotiate the online safety of any user. Hate speech continues to be a major problem in the context of social media in Bangladesh. There are initiatives in practice to resolve this issue and develop a more secure, diverse and tolerant cyberspace for everybody.

Negative and impolite comments made online can have effects in ways. Impacting not just the individual being attacked but also the broader community by causing lasting harm and leaving emotional wounds. As manual detection of hate speech is difficult and it needs time to detect a particular user when peoples high-volume activity in Twitter. DL technology has demonstrated to detect hate speech far more accurately and efficiently. Different ML algorithms use neural networks to analyze large amounts of data. Deep learning models need to be trained on a corpus of texts, perhaps labeled as hate/no-hate in the context of hate speech detection. The models can later be used to classify new texts in hate speech or not. The inclusion of different features, such as emotion, sentimental text, inclusion on emoji, and reactions instead of stylistic features can have a significant impact on DL model. It will enhance DL model performance for hate speech detection. Social media-based posts, comments, photo comments, blogs or news articles could be the open data source for hate speech detection. Hate speech detection using DL is an evolving field. Though it has some shortcomings mainly due to the biased nature of the training dataset, labeling, and fairness in freedom of expression vs hate speech. The DL method for the detection and prevention of hate speech on social media platforms has a potential impact with modern technology. This also helps to secure a safer and more inclusive online space for upcoming generation.

Section - II contains the previous research work made by the researchers on the subject. The system framework is discussed in Section - III. The elements of hate speech detection system's methodology are thoroughly described in Section - IV. In Section - V the results of the experiments are presented. All the findings from the complete discussion as well as potential for further development are presented in Section VI. The last part of this article contains references.

2. LITERATURE REVIEW

In this hate speech detection using machine learning and deep learning study, we have gone through a complete literature review. These studies helped us in analyzing multiple other research work on the same field which highlights and gave deep cultural insights towards this topic. People find it easier to express their opinions and information on virtual communities via online media platforms. Sentiment analysis and behavioral analysis are two data analysis methods that aim at the exploration of different language functions, such as attitudes and emotions. One of these negative styles of speech that people can use is called hate speech. With comment and speech people show their opinions in a negative, neutral, positive way or discrimination towards other races, gender or even other forms.

Recently, an efficient method for hate speech detection in online social media using transfer learning with pre-trained BERT model is proposed (Mozafari *et al.,* 2020). The authors add that social media has come to serve as a major source of information exchange and communication. In social media people can freely express their views, thoughts and participate in debates. Hate speech on social media platforms has real-world consequences in terms of individuals and the society as a whole. Their detection model used the BERT-CNN module. In their study they employed Twitter dataset with three output classes. Their BERT-CNN model, on the other hand reaches 65% accuracy at most. Another study proposed a classification technique Multinomial Logistic Regression (MLR) to identify hate speech on twitter (BR Ginting *et al.,* 2019). Twitter has more than 330 million active users. This is a simple and fast method that works well when applying on small datasets. They tested on the Indonesian tweets dataset resulted in an outstanding accuracy rate of 87.68%.

In the following study, a model focused on detecting hate speech in Spanish was proposed to check how machine learning methods were performing (Plaza-del-Arco *et al.,* 2021). Everything from classic models, to state-of-the-art deep learning methods using large pre-trained language models like BERT or XLM and transfer learning for a better score. The result of their implemented models is promising, and reveal a new perspective for the automatic detection of hate speech in Spanish tweets dataset. They used Spanish tweets as their dataset with two output classes and achieved an accuracy of 87.29% on BETO. A hate speech identification model is developed to identify hate speech from Arab tweets (Al-Hassan & Al-Dossari, 2022). The target of their study is to divide Arabic tweets into five classes that include hate speech categories. There were 11K labeled tweets in their used dataset. In their study, SVM is used as a baseline model and its performance in compared with advanced deep learning-based models: LTSM; CNN+LTSM; GRU; CNN+GRU. By integrating the CNN+ LTSM model, they note an increase in overall average recall of hate speech by 72%.

A specific detection model called BiCHAT was applied in hate speech detection study. This model uses the BERT layer in conjunction with a deep CNN algorithm and a hierarchical attention-based Bi-LSTM network (Khan *et al.,* 2022). The task of the model was to classify tweets into hateful and normal categories by learning the previously trained tweet dataset. The authors also observed the effect of several neural network structural elements on BiCHAT's performance. They examined the embedding methods, activation functions, batch size and optimization methods on BiCHAT model. It was found that the performance of the model was the most affected by the deep convolutional layer's performance. Research was carried out to compare learning techniques in detecting hate speech in the Afaan Oromo language (Ganfure, 2022). The researcher analyzed Facebook and Twitter data using four output categories. They evaluate deep learning models efficacy in identifying hate speech in the Afaan Oromo language. The researchers labeled a dataset containing hate speech. They observed the model combine with CNN and Bi-LSTM showed the highest performance with an average F1 score of 87%. Another study conducted on identifying hateful speech on Facebook pages written in Bengali (Ishmam & Sharmin, 2019). During their research project they sorted 5,126 Bengali comments into six categories including hate speech and religious comments. This dataset represents the effort to identify Bengali content on social media platforms. Their study also made use of ML techniques. The random forest algorithm produced an accuracy of 52% while the GRUs based model saw an accuracy of 70%.

A study was conducted to developed Twitter hate speech text classification model with deep learning (Gambäck & Sikdar, 2017). For instance, the researcher used deep learning to structure a Twitter hate speech identification system. They made classification on tweets as racism or sexism or both and non-hate. The training and evaluation of the system were performed with word2vec embedding technology. Their model achieved 78.3% F1-score using all feature sets. They found the highest F1-score in comparison through different CNN models. A deep learning architecture was designed for detecting cyberbullying using advanced preprocessing techniques on Roman Urdu data (Dewani *et al.*, 2021). Formatting slang-phrase dictionary and removing cyberbullying domain-specific stop words were done in their study. They utilized RNN-LSTM and RNN-BiLSTM models and their applied models achieved validation accuracy of 85.5% and 85%, respectively.

For this purpose, researchers conducted a study to construct a multilingual Twitter corpus for hate speech detection and demographic bias assessment (Huang *et al.*, 2020). They considered different features like age, country, gender and race/ethnicity. The accuracies of four classify models were calculated. They also measure unfairness and bias of the baseline classifiers on demographic attributes. In another study, researcher worked on twitter data to identify the hate speech and offensive language (Bisht *et al.*, 2020). They considered hate speech and offensive language in social media as a problem. They proposed a model to solve with an LSTM-based classification system using word embeddings and neural networks for the separation of these two classes. With LSTM-based model they achieved 86% accuracy in their study.

A study was conducted on developing a deep neural language model to detect offensive language with C-BiGRU model (Mitrović *et al.*, 2019). They combined convolutional and recurrent neural networks with word2vec word embedding method. Their model obtained a macro F1-score of 79.40%. Their proposed model was effective in detecting hate speech in both English and German tweets. Another research that works on improving classification using Arabic and how it done by multi-labeling systems (El Rifai *et al.*, 2022). The study utilized a variety of shallow learning classifiers and an ensemble model to improve accuracy. Deep learning techniques including custom accuracy metrics and ten neural networks were employed in their model. Their study achieved most effective multi-labeling classifier with CGRU.

Various studies have utilized comparable pre-existing datasets, with the highest number of studies falling into this category. As a result, we are considering creating our dataset and labeling it according to the unique cultural and ideological perspectives of the population in question. Our familiarity with different techniques and models has increased our interest in developing a deep learning-based hate speech detection system. This research contributes to the landscape of hate speech detection by focusing on Bengali language data from social media platforms like Facebook and Twitter. While prior studies mainly addressed languages such as English, Spanish, and Arabic, this research emphasizes the need for detection methods tailored to local contexts. By using advanced word embedding techniques like Word2Vec, FastText, and BERT, the study aims to enhance the representation of Bengali text. Utilizing five distinct algorithms will allow for a comprehensive comparison of performance, addressing the gap in literature regarding hate speech detection in Bengali.

3. SYSTEM FRAMEWORK

This study suggests a model for detecting hate speech in comments posted on social media or online, which is depicted in Figure 1. In this model, users can write comments on social media, which will be received at their end. These comments will then be transmitted over the internet to an expert system that has been developed using a pre-existing model. The process of how people's comments on social media are used to obtain search query results through a pre-trained model involves several steps. First, the comments made by people on social media platforms are taken as input. These comments are then passed through a connectivity channel, which allows them to be transmitted to a server where they can be processed. Once the comments are received at the server, they are sent to an inference engine.

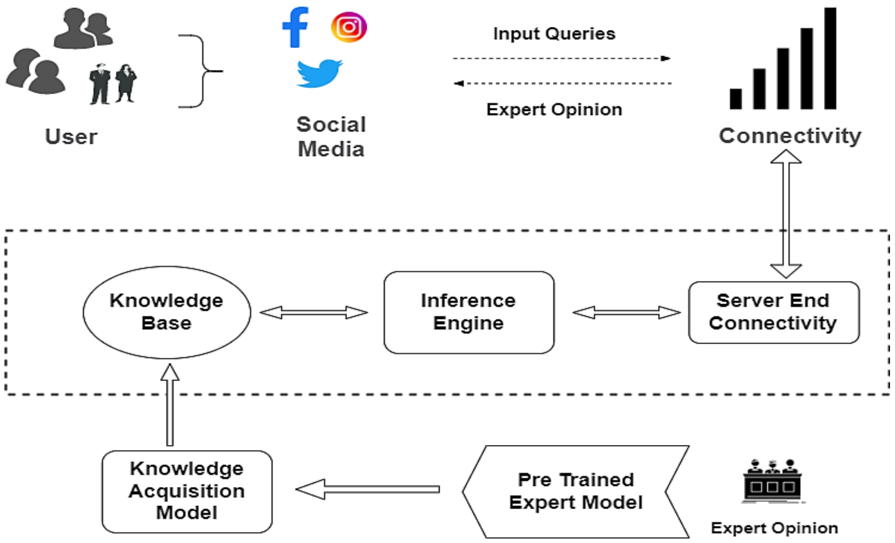


Figure 1: System architecture for social media-based hate speech detection model

The inference engine is responsible for analyzing the comments and extracting meaningful information from them. This information is then used to generate search queries based on the user's intent. To generate accurate search queries, the inference engine queries a knowledge base. The knowledge base contains a collection of information and data relevant to the user's search query. The engine queries this knowledge base to find information that is relevant to the user's query. This query search result along with the acquired knowledge is matched against the pre-trained model. The pre-trained models mean a machine learning or deep learning model trained with large dataset. The ML/DL models can now spot the patterns in an input based on their training. After the learning they can provide predictions based on their previous learning. The search engine can better its result accuracy using information that the knowledge retrieval model waives off. This sends the feedback with the comparison of improved results from the pre-trained model. This will help social media comments to detect hateful comments and speech.

4. METHODOLOGY

Detection of hate speech is a difficult task. It needs to be done using NLP techniques along with ML algorithms (Baki *et al.*, 2023). It starts with a large dataset of text data. The text datasets have different categories. The commonly used categories are 'hate speech' and 'non-hate speech'. This dataset is then used to train a ML/DL model. The model would typically be neural network based in order to recognize the patterns and features that contain hate speech. The data is then preprocessed using different techniques. The processed data then used in DL to extract features which the model can make use of. This study presents a detailed analysis of the different steps in chronological order for data collection, processing and model creation across two distinct sections.

A. Data Collection and Data Preprocessing

Facebook and Twitter data due to the methodology of this study was scraping from social media. Total 6,720 comments were collected from different posts in Facebook and hence a good spectrum of the population. Although Twitter is less popular in Bangladesh and that limited the data collection from twitter. A total of 960 tweets were also collected from Twitter. In total, 7,680 comments were collected and preprocessed for analysis. The dataset was split into the training and test sets as 80% and 20% of total data. A comprehensive labelling was accomplished. The comments were categorized into six different groups. The volunteers from various age ranges and socioeconomic backgrounds were engaged on voting to the comments. Based on the volunteers voting the comments were categorized. In case of disagreement between volunteers, a consultant provided the final interpretation upon which they settled. This method helped identify unacceptable comments in the context of Bangladesh society. During the labeling process, comments were divided into six categories according to age group of commenters. The comments were then handed to volunteers, who dutifully categorized them. Table I provides a synopsis of the labeling procedure applied on the collected data. This methodology also made sure that the comments were labeled correctly. This labelling process also valuable in understanding the people's opinions and emotions regarding Bangladesh. Label the temporary text/speech consistent with the final mark that received the mainstream of votes. This stage is quite important for assuring the validity and correctness of the data collection and the labeling process. The task of the labeling was also to annotate Arabic text with several labels (El Rifai *et al.*, 2022). The research aimed to enhance the accuracy and efficiency of classifiers. These classifiers will be utilized to handle texts applying with multi-label systems.

Table 1
The data labelling process was carried out by soliciting opinions from different groups of volunteers

Comments	**Vol$_1$**	**Vol$_2$**	**Vol$_3$**	**Vol$_4$**	**Vol$_5$**	**Vol$_6$**	**Label=Max** (Vol$_1$:Vol$_6$)
Bangladesh cricket team plays well	*positive*	*positive*	*neutral*	*positive*	*positive*	*positive*	*positive*
Bus accident korse, cholen driver ke amra marte jai (*There has been a bus accident, let's go and beat the driver*)	*negative*	*negative*	*racism*	*racism*	*negative*	*negative*	*negative*
Hey se** girl, what are you doing alone here?	*sexiest*	*negative*	*sexiest*	*sexiest*	*sexiest*	*sexiest*	*sexiest*
I have been working considering the people of the country in mind	*positive*	*positive*	*neutral*	*neutral*	*neutral*	*neutral*	*neutral*
Kalo cheleder ke khelay nibo na (*I won't let the black boys play*)	*racism*	*racism*	*racism*	*negative*	*negative*	*positive*	*racism*
Do not indulge in violence on the sports field	*neutral*	*neutral*	*positive*	*neutral*	*neutral*	*positive*	*neutral*

***Vol$_1$ = age (below 18)*
Vol$_2$ = age (19-25)
Vol$_3$ = age (26-34)
Vol$_4$ = age (35-44)
Vol$_5$ = age (45-60)
Vol$_6$ = age (61 and above)

Positive: People share thoughtful and new ideas on social media platform. These types of comments and opinions receive support and appreciations from different sector of the society. Constructive posts share a clear insight about a person or object to others. These kinds of posts do not contain any negative or offensive words. Any widely accepted viewpoints can be measured as a positive comment.

Negative: The comments on social media those hurt or insult anyone can be considered as negative or offensive comments. These kinds of comments are posted to show hate, disrespect and make another person feel low. These comments contain impolite words and language toward a person or group. The negative comment can create division and chaos in society as well.

Neutral: Neutral comments on Facebook stands for the comments do not take any side or show individual judgement. Neutral comment usually shares information regarding any certain topics without supporting or opposing any person. It uses productive language to explain the topic. It does not hurt or disrespect any person. It has less disagreement and less chance of conflict or misunderstanding. Usually, people accept neutral comments without any arguments.

Racism: Racist comments express disrespect to people based on their race, religion, color, or other characters. This category includes insult or denigrate speech or words on a particular group of people or religion. Racist comments disrespect anyone because of their identity. These comments hurt people and group of different society.

Sexism: Obscene and sexually expressive posts and comments on social media has indecent or sexual language to disrespect or annoy somebody. This category involves the use of words or phrases that express an ugly attitude towards a opposite gender. It promotes gender discrimination and reinforces gender categorizes. It includes rude words that can harm any person personal safety and dignity.

Preprocess the labeled data by removing stop words, stemming, and lemmatizing (Vo *et al.*, 2022). This step is necessary to prepare the data for further analysis and modeling. Clean the comments by removing special characters, symbols, and URLs. This step is necessary to remove irrelevant data and make the data consistent. Check for null values in the dataset and resolve any null value issues using the maximum likelihood technique (Kang, 2013). We also perform data normalization including converting the labels to numerical values, to prepare the data for analysis. We tokenize collected comments into words and sentences to prepare the collected data for further use and analysis. Tokenization is simply splitting up text into words, phrases and symbols. In this study, Word2Vec, fastText and BERT tokenization techniques were separately applied in this research. The data preprocessing method and steps are illustrated in Figure 2.

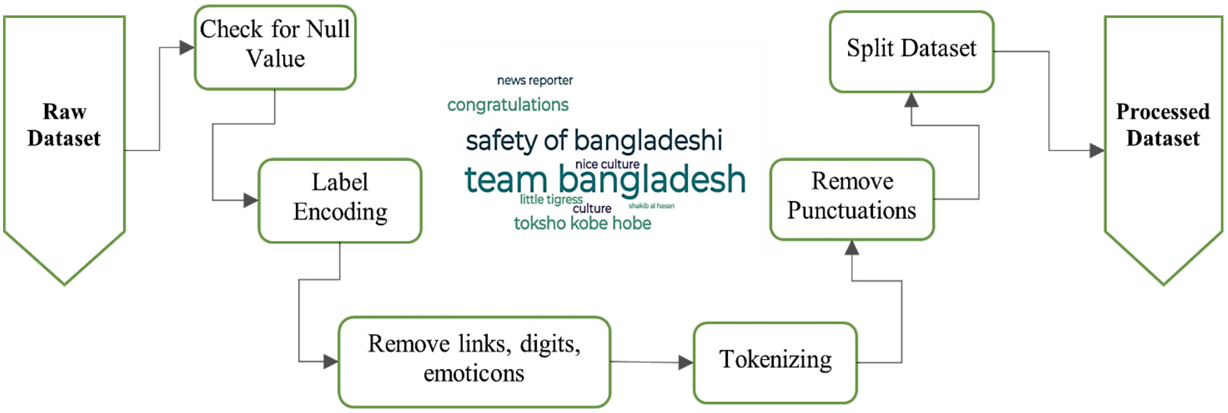


Figure 2: Data preprocessing technique applied in proposed hate speech detection model

Word2Vec: Word2Vec is a popular method for NLP. It creates word embeddings to represent words into vector. It highlights the semantic and syntactic relationships of words and phrases in a corpus. Word2Vec uses neural network to process the text. The input of the neural network is either word and the output are its context or vice versa. It predicts the context via given word and trained on large corpus to learn vector representations of words. The weights of the network are adjusted as part of the training process. This adjustment makes better predictions and allow for high-quality vector representations to be achieved (Bhardwaj *et al.*, 2018). Word2Vec model can be signified by the following equation where w_I stands for input word, w_O stands for output word, v_{wI} stands for input vector representation of w_I, v_{wO} stands for output vector representation of w_O, and v_w is vector representation of any other word inside vocabulary.

$$p(w_O|w_I) = \frac{\exp\,({v'_{w_O}}_{0}^{T} v_{w_I})}{\sum_{w=1}^{W} \exp\,({v'_{w}}_{0}^{T} v_{w_I})} \qquad (1)$$

FastText: FastText world embeddings technique is created and developed by Facebook AI Team. In FastText, a word is broken into chunks. N-grams are small chunks of the word. FastText uses these smaller blocks to get the meaning of incorrectly spelled words. The context of the words is used in the training on large amounts of data. FastText also has the advantage of being very fast and with better handling of words not found in train data compared to other models. Pointwise classification has gained attention for text classification, sentiment analysis and similar NLP tasks (Zhang *et al.,* 2023). Mathematically, FastText can be explained as shown in Equation (2) where v_w stands for the FastText embedding for word *w*, G_w stands for the set of all *n*-grams of *w*, z_g stands for the embedding vector for *n*-gram *g*, and $\sum$ represents the outline of overall n-grams in G_w.

$$v_w = \sum_{g \in G_w} z_g \tag{2}$$

BERT: BERT is a common word embedding technique developed by Google. It can read text bidirectionally to grasp the context of words. It uses transformer architecture. It processes all words in a sentence at the same time. It can easily understand the relationship between words. BERT executes language tasks for pre-trained large dataset. Two types of tasks are performed in BERT. One task is predicting missing words in a sentence. Another task is learning the order of sentences for next-sentence prediction. Table 2 presents the summary of main features of word embedding methods utilized in this study.

Table 2
Attributes of word embedding techniques

Word2Vec	Size = 100
	Window = 5
	Minimum count= 1
	Workers = 4
	Vocab size= None

FastText	Model word ngrams= 1 Model epoch= 5 Model minimum count= 5 Model min = 3 Model max = 6
BERT	Vocab size= 30522 Number of hidden layers= 12 Hidden act= gelu Max position embeddings= 512 Gradient check pointing= false

B. *Hate Speech Detection Deep Learning Model*

The hate speech detection model identifies hate and rude speech content in sentences from social media and online platform. Deep learning models are used to train the models. It assists identify words or phrases that suggest hate speech by analyzing the patterns in text. A social media comment dataset is applied to learn the model as well. CNN, LSTM and others promising DL models are used for the learning model. Models and techniques are being developed in the field of hate speech detection to identify and classify hate speech in online communication. DL is a subset of machine learning. It trains a neural network to make predictions based on data. DL models can be used to train on large datasets of text to classify hate speech by finding patterns in the text that are indicative of hate speech. Models with deep learning made good performance on hate speech identification. DL is more complicated to build and needs sufficient computing power. Bi-LSTM, Bi-GRU-LSTM-CNN, CNN, LSTM, and XGBoost algorithms are used to prepare the hate speech detection model for this study. This study assessed the performance of the models by using three different embedding techniques. The study evaluated how effective were the models in detecting objects correctly, and also separating their segmentation. The researches were conducted with three distinct embedding techniques to observe the impact of those embedding techniques on the performance of the hate speech detection model. The result of the study may have impact on creating detection model. The performance of the models also encounters for further research.

Bi-GRU-LSTM-CNN: Bi-GRU-LSTM-CNN is basically a hybrid deep learning model. It is generally used for text classification. It has the combination of Bi-GRU, LSTM and CNN layers. These three layers to grab both sequence and spatial features from the text. Bi-GRU layer processes the input text in both forward and backward directions. It guided the model to understand the framework from full sentence. LSTM layer captures long-term dependencies among words. It helps to understand the multifaceted sentence structures. CNN layers work on extracting local patterns and features from the text. These local patterns help the model to understand meaningful phrase. This hybrid model works well to handle complex data and get good accuracy.

Bi-LSTM: Bi-LSTM is one type of recurrent neural network (RNN). It learns the context from both past and future sequence of the text. It is able to process the input from both directions: forward and backward. Bi-LSTM is very effective for text classification. As it learns from both sides, it can compute well with complex sentences. It can learn the pattern of the text rapidly. It is a modified version of LSTM. It also comprises two LSTM layers. One-layer processes the input for forward direction and another layer executes for backward direction.

CNN: Convolutional Neural Network (CNN) is very popular DL model. It is used for text classification, image recognition and processing. CNN can extract features from raw text data. No manual feature engineering is required for the feature extraction. A fully connected neural network teaches the model and the convolutional layers extract local features from the input text. The architecture of CNN model for text classification can be expressed as presented in Equation (3).

$$Y = softmax\,(Wx + b) \tag{3}$$

Here, Y stands for the output probability distribution, *softmax* is the activation function. It converts the output scores into probabilities, W means the weight matrix. W connects the convolutional layer to the fully connected layer. x is the input text data and b is the bias term. CNN has several layers that perform different operations on the input data. The layers are as follows: embedding layer, convolutional layer, pooling layer, and fully connected layer. The convolutional layer identifies these features by passing over the text. The pooling layers facilitate information compression and recollect critical patterns.

LSTM: Long Short-Term Memory (LSTM) model is a kind of RNN. It processes multiple sequences. LSTMs are similar to RNNs. It is able to capture long-term dependencies in data and good for text classification problems. LSTMs will have a different cell structure. This cell structure is consisting of three types gates. These gates are the input, forget and output gates. The gates control information flow and decide to remember or forget over time. The LSTM networks with this design preserve the relevant information and keep track of important data from the first parts of a text. The output layer produces the final predictions based on the previously learned patterns (Zhang *et al.*, 2023). The equations of LSTM model are as presented in Equation (4), were input gate stands with I_t, forget gate F_t, and output gate O_t.

$$\left.\begin{aligned} I_t &= \sigma\,(X_t W_{xi} + H_{t-1} W_{hi} + b_i\,) \\ F_t &= \sigma\,(X_t W_{xt} + H_{t-1} W_{hf} + b_f) \\ O_t &= \sigma\,(X_t W_{xo} + H_{t-1} W_{ho} + b_o) \end{aligned}\right\} \quad (4)$$

XGBoost: XGBoost (Extreme Gradient Boosting) is a prominent ML classifier. It uses gradient-boosting methods to increase model performance. It works well for detection and classification models. XGBoost ensembles multiple learning methods. These methods take the prediction of several weak models. Then it combines them into a single strong model. XGBoost classifier works fast and produce good accuracy. It deals well with large datasets. It fits several decision trees on each iteration. It also fixes errors made in previous models. All predictions of trees are input to a specific aggregation process. This process increases accuracy of the model. XGBoost transforms text data into numerical form. It applies boosting to identify and learn patterns from the text (*Hands-On Gradient Boosting with XGBoost and scikit-learn. (n.d.)*). XGBoost model can be expressed by Equation (5).

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \quad (5)$$

Here, *f(x)* means predicted class label for the input text *x*, *K* stands for the number of decision trees in the model, and $f_k(x)$ is the prediction of the *k*-th decision tree.

These aforementioned algorithms were used for hate speech classification model for Facebook/Twitter. Figure 3 shows the overall process to prepare model and then next steps. The method of developing hate speech detection model contains some key steps. Accumulate data from multiple social media platforms like Facebook/Twitter. Various posts and comments from Facebook/Twitter are stored. The gathered data is then annotated based on expert opinions to find hate speech. The annotated data is then processed and prepared for analysis. Word vectorization techniques such as word2vec, fastText, and BERT are applied to the processed data. These techniques enable the transformation of words into numerical representations, which can be utilized in machine learning algorithms. A range of deep learning models such as Bi-GRU-LSTM-CNN, Bi-LSTM, LSTM, CNN, and XGBoost are applied to the data to determine which model performs best in detecting hate speech. The models are trained, and their accuracy and epoch performance are compared to determine the most appropriate model for the study. By following this process, a robust hate speech detection model can be developed that can effectively identify instances of hate speech within social media data.

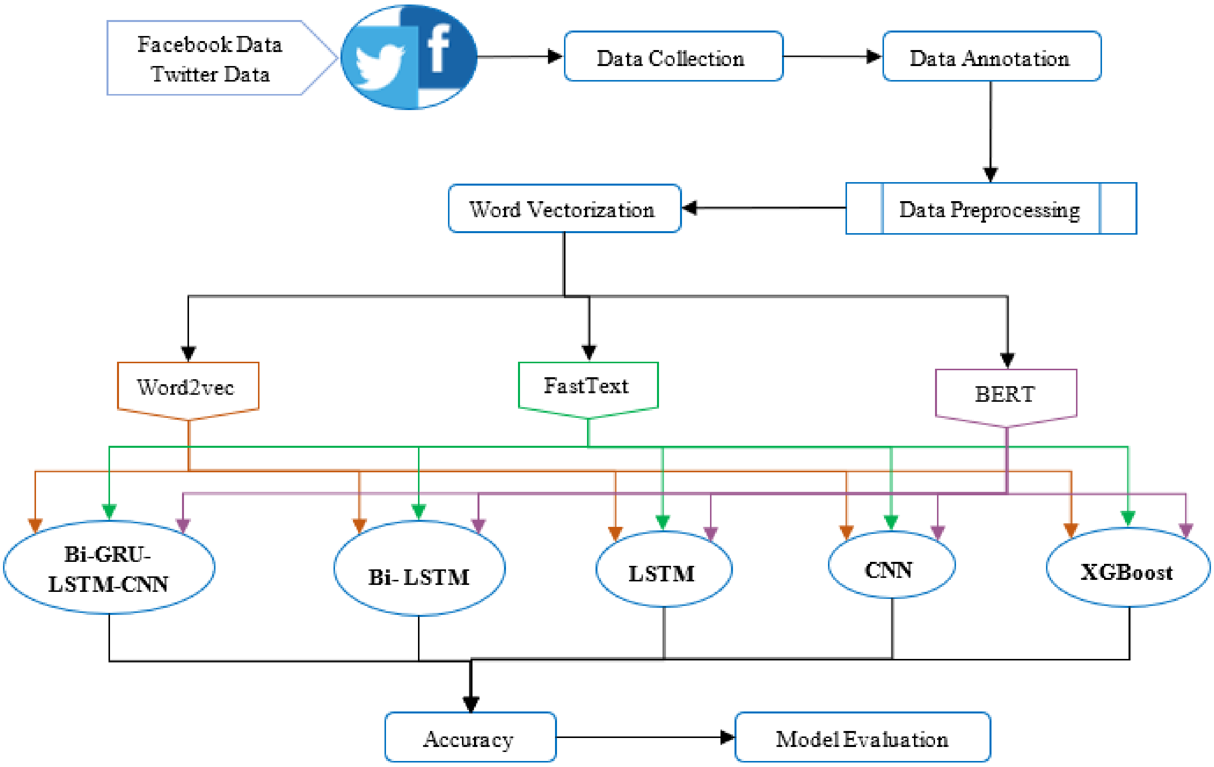


Figure 3: Development approach of proposed social media-based hate speech detection model

5. RESULTS AND DISCUSSION

The discussion and result analysis section of a hate speech detection model using five deep learning models is a critical component of the study. The section provides an in-depth analysis of the accuracy of each model. The discussion outlines the strengths and limitations of each model and highlights the areas where further improvement is necessary. In a hate speech detection study, a critical component is to compare the performance of different deep learning models and word embedding techniques. This is done to determine which combination of techniques can effectively identify hate speech in social media data. The study applies three different word embedding techniques, namely word2vec, fastText, and BERT, against five different deep learning algorithms. The performance of each model is compared with every embedding technique to determine the most effective combination. The performance of each algorithm is illustrated on Figure 4. The figure displays the accuracy scores of various models such as word2vec, fastText, and BERT, along with different architectures such as Bi-GRU-LSTM-CNN, Bi-LSTM, CNN, LSTM, and XGBoost. The models were evaluated based on their performance in hate speech detection tasks. The results show that LSTM had the highest accuracy score of 95.66%, closely followed by XGBoost with 95.27%. Bi-GRU-LSTM-CNN and Bi-LSTM also performed well, with accuracy scores of over 90%. While CNN had a high score of 94.09% for word2vec, its performance was lower for fastText and BERT. The performance of different word embedding techniques on the five algorithms varies due to their distinct characteristics. Word2vec and fastText are shallow models that learn word representations based on co-occurrence statistics. BERT, on the other hand, is a deep contextual model that captures semantic and syntactic information. The choice of algorithm also influences performance. Bi-GRU-LSTM-CNN and Bi-LSTM excel at capturing long-range dependencies, while CNN is better suited for local patterns. LSTM is effective for sequential data, and XGBoost is a powerful ensemble method. The combination of word embedding techniques and algorithms results in varying levels of accuracy in hate speech detection.

Figure 5 to Figure 9 demonstrate how different deep learning algorithms perform across various epochs, presenting their accuracy scores and highlighting their strengths and weaknesses. Figure 10 illustrates the cross-validation accuracy with 10-fold on each algorithm. Table 3 shows the performance comparison of word embedding techniques and algorithms for hate speech detection. In evaluating hate speech detection models, accuracy is a critical metric, measuring the percentage of correctly classified instances from the total number of instances. The equations used for computing accuracy, precision, recall, F1-score are presented in Equation (6) to Equation (9).

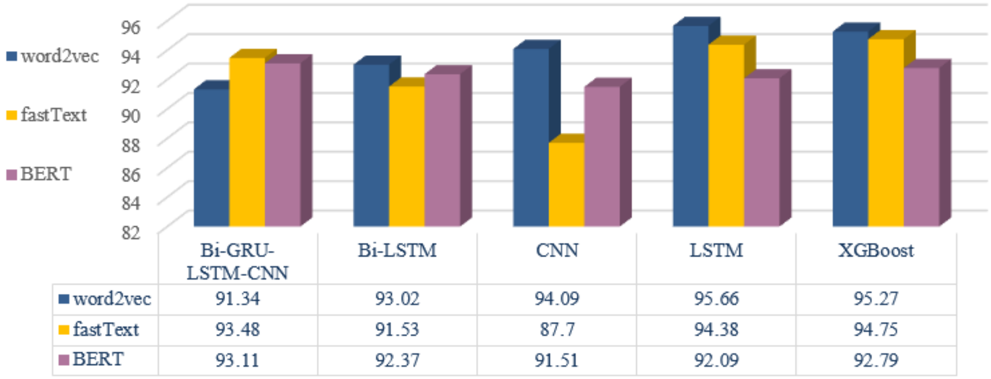
$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{6}$$

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

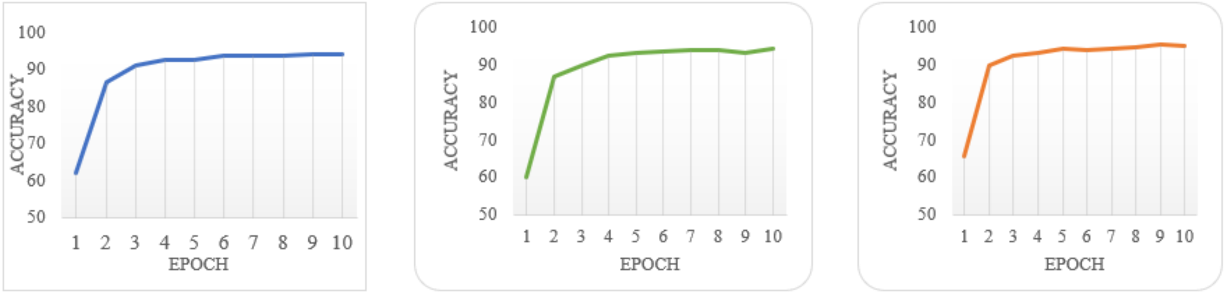
$$F1 - score = \frac{2 \times precision \times recall}{precision+recall} \tag{9}$$

Here *TP* = true positive, *TN* = true negative, *FP* = false positive and *FN* = false negative.



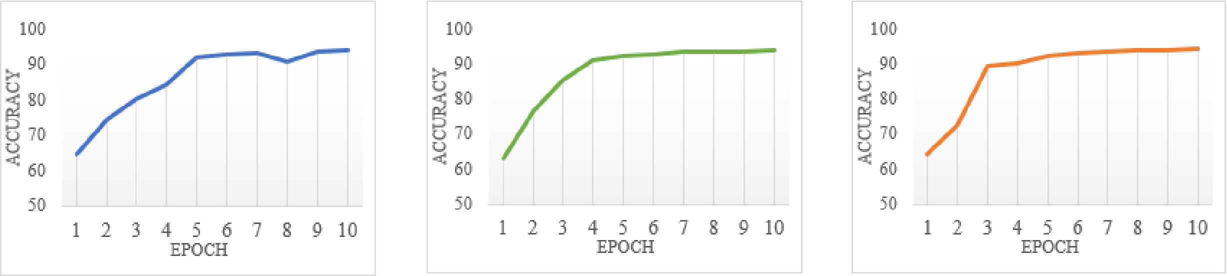
	Bi-GRU-LSTM-CNN	Bi-LSTM	CNN	LSTM	XGBoost
■ word2vec	91.34	93.02	94.09	95.66	95.27
■ fastText	93.48	91.53	87.7	94.38	94.75
■ BERT	93.11	92.37	91.51	92.09	92.79

Figure 4: Analyzing the efficiency score of applied DL algorithms



(a) word2vec (b) fastText (c) BERT

Figure 5: Efficacy of Bi-GRU-LSTM-CNN using various embedding techniques

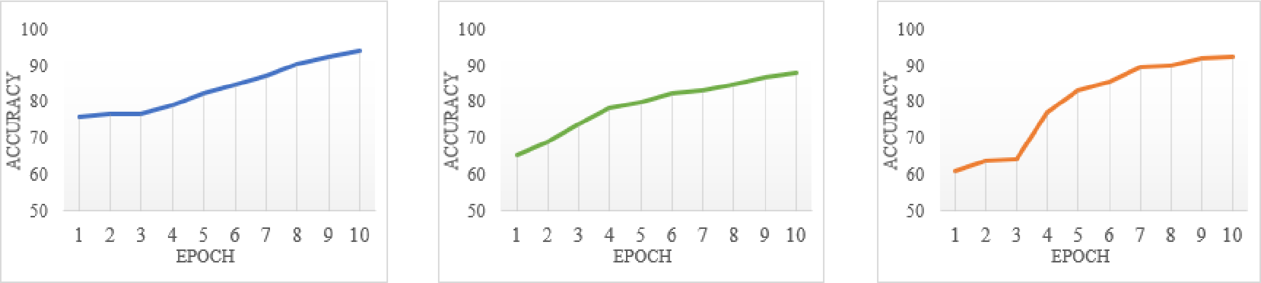


(a) word2vec

(b) fastText

(c) BERT

Figure 6: Efficacy of Bi-LSTM using various embedding techniques

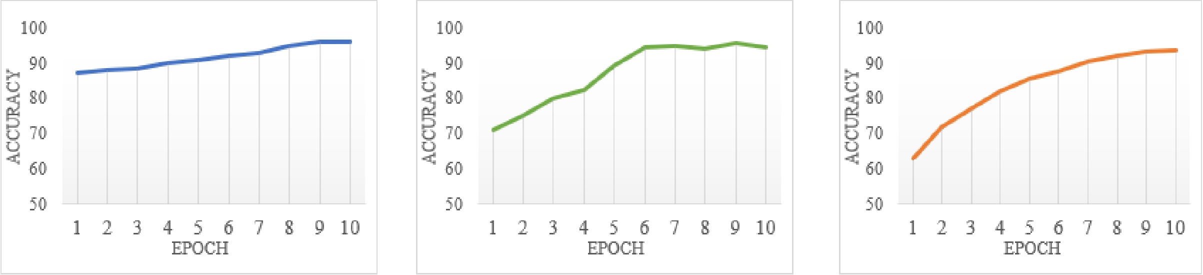


(a) word2vec

(b) fastText

(c) BERT

Figure 7: Efficacy of CNN using various embedding techniques

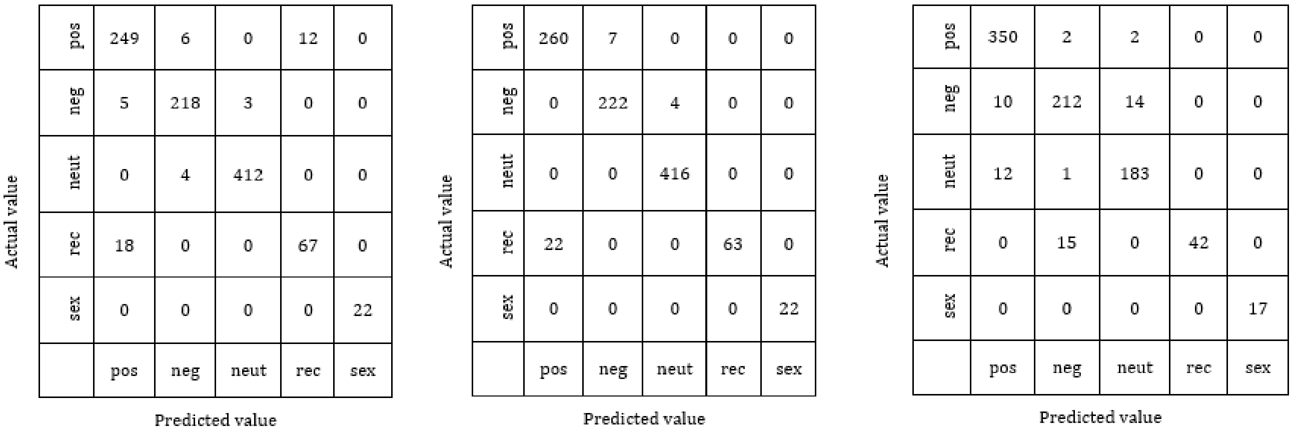


(a) word2vec

(b) fastText

(c) BERT

Figure 8: Efficacy of LSTM using various embedding techniques



Actual value

pos	249	6	0	12	0
neg	5	218	3	0	0
neut	0	4	412	0	0
rec	18	0	0	67	0
sex	0	0	0	0	22
	pos	neg	neut	rec	sex

Predicted value

Actual value

pos	260	7	0	0	0
neg	0	222	4	0	0
neut	0	0	416	0	0
rec	22	0	0	63	0
sex	0	0	0	0	22
	pos	neg	neut	rec	sex

Predicted value

Actual value

pos	350	2	2	0	0
neg	10	212	14	0	0
neut	12	1	183	0	0
rec	0	15	0	42	0
sex	0	0	0	0	17
	pos	neg	neut	rec	sex

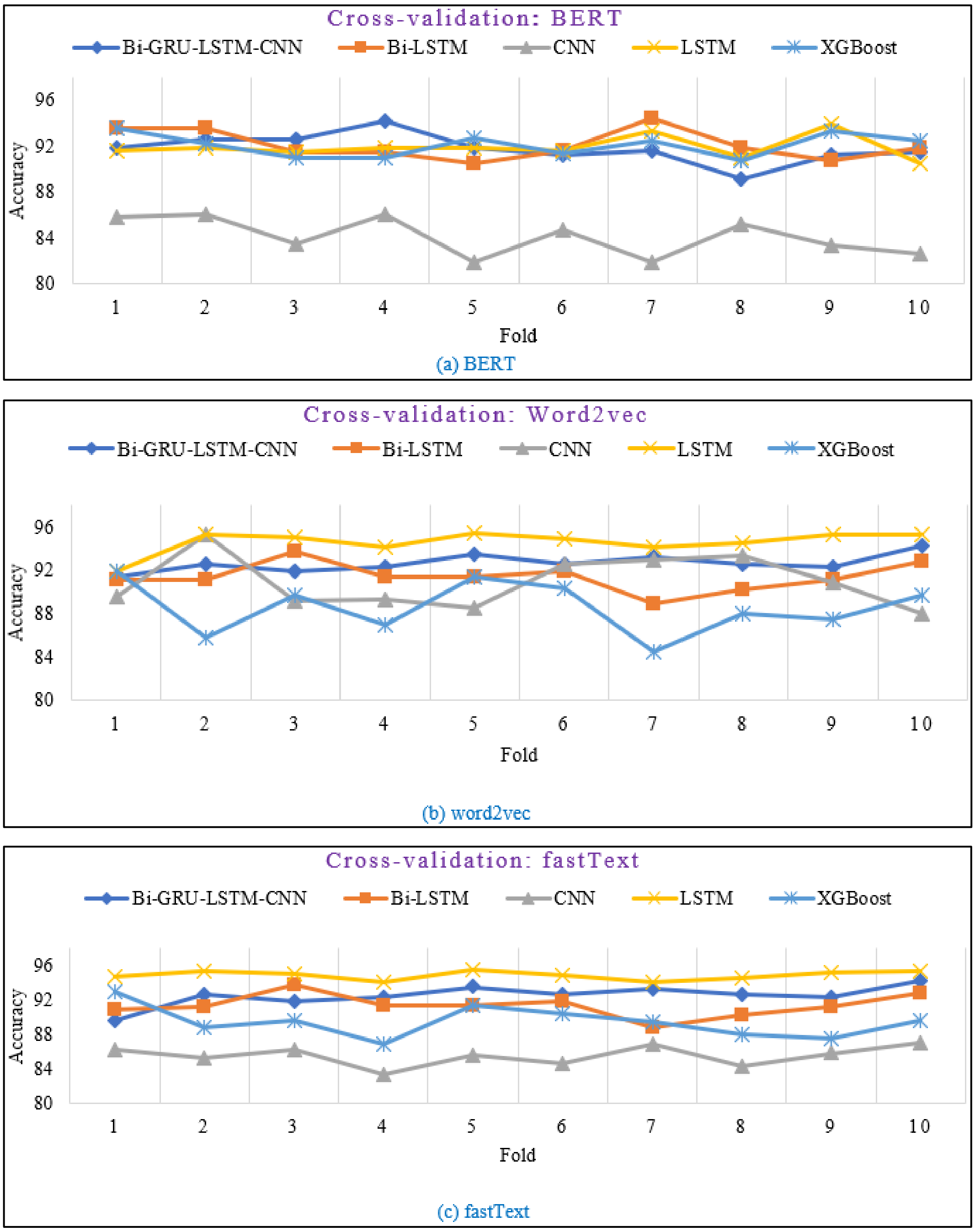
Predicted value

(a) word2vec

(b) fastText

(c) BERT

Figure 9: Efficacy of XGBoost using various embedding techniques


Cross-validation: BERT

(a) BERT

Cross-validation: Word2vec

(b) word2vec

Cross-validation: fastText

(c) fastText

Figure 10: Comparison of cross-validation accuracy for different algorithms and word embedding techniques in hate speech detection

Table 3
Performance comparison of word embedding techniques and algorithms for hate speech detection

	Word2vec			**fastText**			**BERT**		
Algorithms	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Bi-GRU-LSTM-CNN	91.35	91.43	91.77	92.75	92.37	92.37	93.53	93.4	93.27
Bi-LSTM	92.73	90.81	92.61	93.54	93.12	92.98	92.41	91.35	91.1
CNN	93.27	91.73	90.1	78.51	87.7	82.82	88.5	85.93	88.64
LSTM	93.54	92.49	94.51	94.8	94.39	93.92	91.45	90.12	90.24
XGBoost	91.57	89.83	91.96	91.57	91.73	91.3	94.04	93.95	93.84

The performance of DL algorithm relies on different features, including the selection of architecture, hyperparameters, and the size and quality of the training data. In hate speech detection and text classification, these aspects are critical in determining the accuracy and dependability of the outcomes. In this study, Table 4 highlights the hypermeters of the DL algorithms utilized.

In order to assess the effectiveness of our proposed hate speech detection system, we have conducted a comparison with recent and relevant research studies. It is important to note that the assumptions made by these researchers when collecting and reporting sample data will heavily influence the evaluation of our own performance. We have attempted to compare our work with others based on parameters such as data size, embedding techniques, platform, algorithm, and accuracy. The findings of our comparison are presented in Table 5, that provides an overview of both our work and that of others in the field.

Table 5 presents an overview of various NLP models used for hate speech detection on different platforms such as Twitter, Facebook, and news portals. It shows that with this study, LSTM had the highest accuracy score of 95.66% on Facebook and Twitter. BERT-CNN and BETO had scores of 92% and 87.29%, respectively, on Twitter and Spanish Tweets. However, CNN+LTSM, GRU, and C-BiGRU had lower accuracy scores of 72%, 70.10%, and 79.40%, respectively, on Arabic and German Tweets and Facebook. CGRU had the highest accuracy score of 94.85% on Arabic news portal.

This study and other works in the field of hate speech detection share a common goal of developing accurate models for identifying harmful content. However, they differ in terms of data sources, embedding techniques, methods, and performance metrics. This study utilizes Facebook and Twitter data, while others may focus on different platforms. The choice of embedding techniques and algorithms also varies, leading to diverse performance results. In Figure 11, the proposed model for hate speech detection is illustrated. It includes the prior data collection process from social media (Facebook and Twitter). Then data preprocessing has done and Word2vec is used as word embedding technique. A deep learning model prepared with LSTM will be applied for detection the hate speech on given dataset. Lastly the LSTM DL classifiers with word2vec is applied for hate speech classification and prediction model.

Table 4
Specific values assigned as the hyperparameters for applied in DL algorithms

Algorithms	Hyperparameter Name	Hyperparameter Value
Bi-GRU-LSTM-CNN	Max words	3000
	Kernel size	2
	Emb dim	20
	Activation	softmax
	Optimizer	adam
Bi-LSTM	Emb dim	20
	Loss	sparse_categorical_crossentropy
	Spatial Dropout 1D	0.5
	Activation	relu
	Optimizer	adam
CNN	Kernel size	4
	Activation	relu
	Optimizer	adam
	Loss	mean_squared_error
	Pool size	2
LSTM	Activation	softmax
	Loss	mean_squared_error
	Optimizer	adam
	Dropout	0.2
	Spatial Dropout 1D	0.4
XGBoost	Max depth	6
	Subsample	1
	Min child weight	1
	Lambda	1
	Tree method	auto

Table 5
Comparative analysis among our work and other relevant previous studies

Related Work	Modality	Embedding Technique	Method	Num of Class	Platform	Score
This study	Hate Speech Detection	Word2Vec	LSTM	5	Facebook, Twitter	95.66%
Mozafari *et al.,* 2020	Hate Speech Detection	BERT	BERT-CNN	3	Twitter	92%
Br Ginting *et al.,* 2019	Hate Speech Detection	NM	Multinomial Logistic Regression	2	Indonesian Tweets	87.68%
Plaza-del-Arco *et al.,* 2021	Hate Speech Detection	BERT, XLM	BETO	2	Spanish Tweets	87.29%
Al-Hassan & Al-Dossari, 2022	Hate Speech Detection	NM	CNN +LTSM	5	Arabic Tweets	72%
Khan *et al.,* 2022	Hate Speech Detection	BERT	BiCHAT	2	Twitter	88%
Ganfure, 2022	Hate Speech Detection	Word2Vec	CNN +LTSM	4	Facebook, Twitter	87%
Ishmam & Sharmin, 2019	Hate Speech Detection	NM	GRU	6	Facebook	70.10%
Gambäck & Sikdar, 2017	Hate Speech Classification	Word2Vec	CNN	4	Twitter	78.30%
Dewani *et al.,* 2021	Cyberbullying Detection	NM	RNN-LSTM	2	Twitter	85%
Huang *et al.,* 2020	Hate Speech Recognition	TF-IDF	RNN	2	Twitter	89.80%
Bisht *et al.,* 2020	Hateful Speech Detection	NM	LSTM	2	Twitter	86%
Mitrović *et al.,* 2019	Offensive Language Detection	Word2Vec	C-BiGRU	2	German Tweets	79.40%
El Rifai *et al.,* 2022	Text Classification	TF-IDF	CGRU	4	Arabic news portal	94.85%

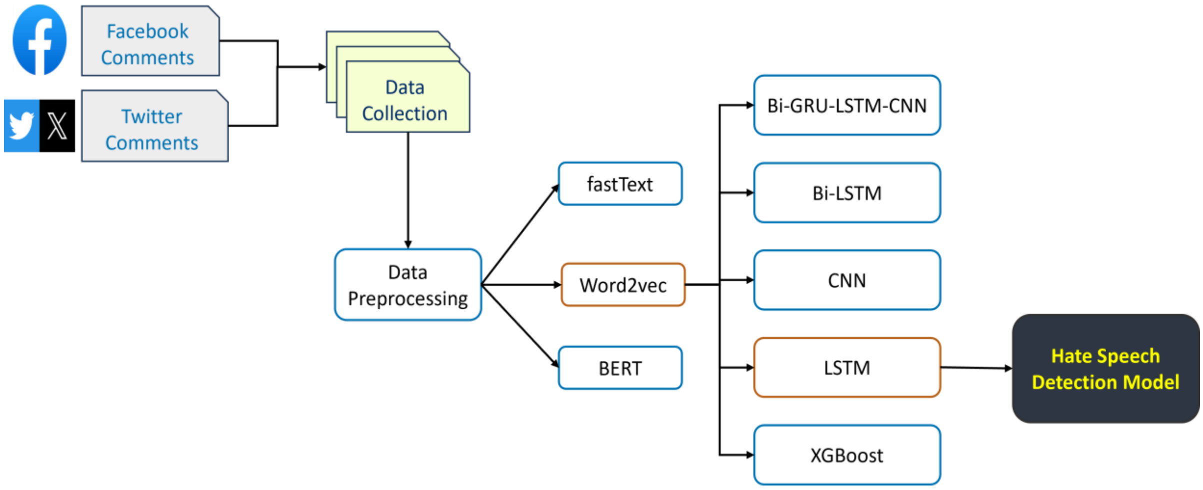


Figure 11: Hate speech detection model: DL approach using social media data

6. CONCLUSIONS AND FUTURE WORK

Last few years, hate speech and offensive comments have become an important issue for online and social media platforms. Sometimes these harmful and hate comments create division and chaos in the society. Hate speech is on the up and fighting against it requires effective measures. One of these solutions is hate speech detection models. This model has specific ML or DL algorithm. The model can identify and classify hate speech based on previous knowledge-based learning. The model is trained to identify and learn language patterns such as context, tone and word choice. It can differentiate between hate speech and other forms of expression based on the text patterns. A large dataset with hate speech non-hate speech allows the model to learn about the texts, words and patterns. After the training the model can distinguished between hate speech and regular speech. This learning model can integrate in a systematic approach for further use in hate speech detection in social media.

It provides an opportunity to critically evaluate the effectiveness of multiple DL models in detecting hate speech and presents valuable insights for improving hate speech detection in social media and online platforms. According to the results, the LSTM algorithm had the highest accuracy score of 95.66%, followed closely by XGBoost at 95.27%. The Bi-GRU-LSTM-CNN and Bi-

LSTM models also performed well. However, the performance of the CNN algorithm varied depending on the language model used. While it achieved a high score of 94.09% for word2vec, it had lower scores for fastText and BERT. These findings are useful for researchers and practitioners looking to enhance their hate speech detection or classification tasks. In future, a good number of labeled datasets can be used and analysis the native language and words for creating a better model for hate speech detection on social media platforms.

ACKNOWLEDGEMENTS

Authors would like to express their sincere gratitude to all the volunteers for participating the data labelling process. They are also thankful to Department of Computer Science and Engineering, Military Institute of Science and Technology (MIST), Bangladesh, for their cordial support, supervision and inspiration during the study. Data used in this study were collected from Facebook and Twitter.

REFERENCES

Ahmed, L., et al. (2023). Context based emotion recognition from bengali text using transformers. *In 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 1478-1484). IEEE.

Al-Hassan, A., & Al-Dossari, H. (2022). Detection of hate speech in Arabic tweets using deep learning. *Multimedia Systems*, 28(6), 1963–1974.

Baki, R. F., et al. (2023). Intelligent Head-bot, towards the Development of an AI Based Cognitive Platform. *MIST International Journal of Science and Technology*, 11(2), 01-14.

Bhardwaj, A., Di, W., & Wei, J. (2018). *Deep Learning Essentials: Your hands-on guide to the fundamentals of deep learning and neural network modeling*. Packt Publishing Ltd.

Bisht, A., Singh, A., Bhadauria, H. S., Virmani, J., & Kriti. (2020). Detection of Hate Speech and Offensive Language in Twitter Data Using LSTM Model. In S. Jain & S. Paul (Eds.), *Recent Trends in Image and Signal Processing in Computer Vision* (pp. 243–264). Springer.

Br Ginting, P. S., Irawan, B., & Setianingsih, C. (2019). Hate Speech Detection on Twitter Using Multinomial Logistic Regression Classification Method. *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, 105–111.

Deutsche Welle. (2019, October 24). *Bangladesh: Fake news on Facebook fuels communal violence*. DW. https://www.dw.com/en/bangladesh-fake-news-on-facebook-fuels-communal-violence/a-51083787.

Dewani, A., Memon, M. A., & Bhatti, S. (2021). Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for Roman Urdu data. *Journal of Big Data*, 8(1), 160.

El Rifai, H., Al Qadi, L., & Elnagar, A. (2022). Arabic text classification: the need for multi-labeling systems. *Neural Computing and Applications*, 34(2), 1135–1159.

Gambäck, B., & Sikdar, U. K. (2017). Using Convolutional Neural Networks to Classify Hate-Speech. *Proceedings of the First Workshop on Abusive Language Online*, 85–90.

Ganfure, G. O. (2022). Comparative analysis of deep learning based Afaan Oromo hate speech detection. *Journal of Big Data*, 9(1), 76.

Hands-On Gradient Boosting with XGBoost and scikit-learn. (n.d.). Packt. Retrieved April 16, 2023, from https://www.packtpub.com/product/hands-on-gradient-boosting-with-xgboost-and-scikit-learn/9781839218354.

Huang, X., Xing, L., Dernoncourt, F., & Paul, M. J. (2020). Multilingual Twitter Corpus and Baselines for Evaluating Demographic Bias in Hate Speech Recognition. *In Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 1440–1448). European Language Resources Association.

Ishmam, A. M., & Sharmin, S. (2019). Hateful Speech Detection in Public Facebook Pages for the Bengali Language. *2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 555–560.

Javaid, M., Haleem, A., Singh, R. P., & Suman, R. (2022). Artificial Intelligence Applications for Industry 4.0: A Literature-Based Study. *Journal of Industrial Integration and Management*, 07(01), 83–111.

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, 64(5), 402–406.

Khan, S., Fazil, M., Sejwal, V. K., Alshara, M. A., Alotaibi, R. M., Kamal, A., & Baig, A. R. (2022). BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection. *Journal of King Saud University - Computer and Information Sciences*, 34(7), 4335–4344.

Mitrović, J., Birkeneder, B., & Granitzer, M. (2019). nlpUP at SemEval-2019 Task 6: A Deep Neural Language Model for Offensive Language Detection. *Proceedings of the 13th International Workshop on Semantic Evaluation*, 722–726.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media. In H. Cherifi, S. Gaito, J. F. Mendes, E. Moro, & L. M. Rocha (Eds.), *Complex Networks and Their Applications VIII* (pp. 928–940). Springer International Publishing.

Plaza-del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2021). Comparing pre-trained language models for Spanish hate speech detection. *Expert Systems with Applications*, 166, 114120.

The Business Standard. (2021, May 21). Why Chanchal Chowdhury's comment box needs our attention. https://www.tbsnews.net/feature/panorama/why-chanchal-chowdhurys-comment-box-needs-our-attention-244456.

Vo, H. H.-P., Nguyen, H. H., & Do, T.-H. (2022). Analysis of the Effects of Stop-word Removal in Hate Speech Detection Problem for Vietnamese Social Network Data. In N.-T. Nguyen, N.-N. Dao, Q.-D. Pham, & H. A. Le (Eds.), *Intelligence of Things: Technologies and Applications* (pp. 299–309). Springer International Publishing.

Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2023). *Dive into Deep Learning* (No. arXiv:2106.11342). arXiv.