Social Network Analysis and Mining (2022) 12:24
https://doi.org/10.1007/s13278-021-00852-x

ORIGINAL ARTICLE

Abusive Bangla comments detection on Facebook using transformer-based deep learning models

Tanjim Taharat Aurpa[1] · **Rifat Sadik**[1] · **Md Shoaib Ahmed**[1]

Received: 27 June 2021 / Revised: 6 October 2021 / Accepted: 1 November 2021 / Published online: 29 December 2021

Abstract

In the era of social networking platforms, user-generated content is flooding every second on online social media platforms like Facebook. So observing and identifying many contents, including threats and sexual harassment, are more accessible than traditional media. Online content with extreme toxicity can lead to online harassment, profanity, personal attacks, and bullying acts. As Bangla is the seventh most spoken language worldwide, the utilization of Bangla language in Facebook has raised current times. The use of abusive comments on Facebook with Bangla also has increased alarmingly, but the research regarding this is very low. In this research work, we concentrate on identifying abusive comments of Bangla language in social media (Facebook) that can filter out at the primitive stage of social media's affixing. To classify abusive comments swiftly and precisely, we apply transformer-based deep neural network models. We employ pre-training language architectures, BERT (Bidirectional Encoder Representations from Transformers) and ELECTRA (Efficiency Learning an Encoder that Classifies Token Replacements Accurately). We have conducted this work with a novel dataset comprises 44,001 comments from multitudinous Facebook posts. In this classification process, we have exhibited an average accuracy, precision, recall, and $f1$-score to evaluate our proposed models. The outcomes have brought a percipience of our applied BERT and ELECTRA architecture that performs notably with 85.00% and 84.92% test accuracy, respectively.

Keywords Deep learning (DL) · Bangla abusive comments · Cyberbulling · Bangla cyberbulling · BERT · ELECTRA · Transformers

1 Introduction

Online social networks (OSNs) are a podium where human interactions occur by posting texts, images, videos, etc. The mode of communication in social media materializes via messages, comments, and chats Ahmed et al. (2020b). The impact of social media nowadays is increasing in both people's personal lives and their professional circumstances. Comments are the most common and straightforward way which capacitate a reciprocal way of providing an individual point of view. Commenters can easily express their sentiments, opinions, and also responses Ahmed et al. (2021b). Besides bringing ease of communication among people, this has some drawbacks too. Among these, abusive comments and cyberbullying have become an alarming threat to social media users.

Bangla is the world's sixth-largest language that more than 260 million users speak in daily life. This language also has historical importance. UNESCO also declared the international mother language day based on the Bangla Language Movement happened in 1952. Recently Facebook has become the most influential social media platform for Bengali users. According to a statistics[1] 3.3% users of Facebook uses the Bangla language, which constitutes 71 million population. A project[2] to predict the hate crimes in

Rifat Sadik and Md Shoaib Ahmed have contributed equally to this work.

✉ Tanjim Taharat Aurpa
taurpa22@gmail.com

Rifat Sadik
rifat.sadik.rs@gmail.com

Md Shoaib Ahmed
shoaibmehrab011@gmail.com

[1] Department of Computer Science and Engineering, Jahangirnagar University, Dhaka 1342, Bangladesh

[1] https://wearesocial.com/digital-2021.

[2] https://migrationdataportal.org/data-innovation/Facebook-data-predict-hate-crimes-against-refugees-Germany.

Germany derived a relationship between hate crimes and social media like Facebook. In Bangladesh, cyberbullying has become a burning problem. Observing some past events of Bangladesh, communal crimes such as rioting erupted in a community due to Facebook. Besides, defamation, bullying, and harassment are also some major crimes that take place on social media platforms. In recent years, Bangladeshi celebrities and influencers are becoming victims of abusive comments after posting about different topics[3]. So, it is mandatory to monitor Facebook posts, comments, shares to prevent any kind of cybercrimes.

The importance of research on these topics based on different languages is so demandable that many recent works have been conducted. Authors in Nobata et al. (2016) accomplished a machine learning-based abusive comments detection method. Another work on this topic is proposed in Janardhana et al. (2021). In this work, authors classified comments as abusive or non-abusive using convolutional neural network (CNN).

Bidirectional Encoder Representations from Transformers (BERT) has gained popularity since it was introduced in natural language processing. This transformer-based model outperforms many NLP segments like text classification, entity recognition, question answering, etc. Authors in Adhikari et al. (2019), Yu et al. (2019), Ostendorff et al. (2019), and Chia et al. (2019) used BERT for text/document classification and proposed different modified BERT architechtures for better performence. This transformer-based pre-trained language model brings higher accuracy in other NLP sectors like sentiment analysis (e.g., Yu and Jiang (2019), Li et al. (2019), and Su et al. (2020)), question answering (e.g., Yuan (2019)), entity extraction and recognition (Xue et al. (2019), Souza et al. (2019), and Ashrafi et al. (2020)).

Once BERT disclosed to the NLP world, it ruled. Nevertheless, in 2020 another pre-train language model (PLM) ELECTRA was proposed, and it overcame the limitations that come with mask language models (MLM). ELECTRA train a model with the generator that train like MLM and discriminator responsible for identifying the token replace by the generator. ELECTRA has been proved efficient in many NLP domains like sentiment/emotion analysis (e.g., Xu et al (2020), Al-Twairesh (2021)) , Fake news analysis (e.g., Das et al. (2020)), text mining (e.g., Ozyurt (2020)), etc. It also shows good performance in cyber bullying related works in Pericherla and Ilavarasan (20218), the domain we are focusing on in this paper.

The automated systems on the Bangla language can be helpful for a large number of users around the world. So classifying these Bangla abusive comments and taking proper steps are also as crucial as classifying English ones. However, less work has been accomplished on this issue in other online media (e.g., YouTube) rather than Facebook. Existing ones do not use the latest technologies to bring out results with higher accuracy. In this paper, we want to contribute to this issue using the latest NLP technologies and propose a solution that classifies Bangla abusive comments more accurately. The whole contribution of this work is summarized below-

- Propose a superlative framework that classifies different types of abusive comments using the latest pre-trained language model (PLM) BERT and ELECTRA.
- Justify the model's efficiency based on real-world Bangla abusive comments.
- Determining significant evolution methods for analyzing the performance of our proposed model.

2 Related work

Sentiment analysis on social media is an emerging research topic nowadays. Different word embedding techniques with ML models achieved results while analyzing sentiments. Samad et al. (2020) used word embedding to classify sentiments from tweeter posts. Salur and Aydin (2020) used different word embedding techniques like with different deep learning techniques. Combined features from Word-2Vec, FastText, and char-level embedding were then used in LSTM, GRU, BiLSTM, and CNN models to classify Turkish tweets. Moreover, a hybrid model is built by combining CNN and BiLSTM, which outperforms other models. Classification of three types of sentiments from tweets was studied by Alzamzami et al. (2020) where light gradient boosting machine (LGBM) framework was used. The proposed model compared with six other conventional models like linear regression, support vector machine, random forest, gradient boost, etc., and achieved the best classification results than others. Sentiment classification by using the interaction between tasks was proposed by Zhang et al. (2020) for Chinese blog posts. BiLSTM with attention and CRF were combined to extract features from the text and ERNIE model to classify texts.

Researches are now focusing on developing methods to monitor social media platforms like Facebook, Twitter, Snapchat, and so many. General machine learning approaches are used to classify and detect abusive or toxic comments on social media. For the classification of hateful comments, Salminen et al. (2020) used a machine learning approach. Different ML (machine learning) algorithms like logistic regression, naïve Bayes, support vector machines, XGBoost, and neural networks were used for the classification task. A dataset had been constructed by collecting

[3] https://www.dhakatribune.com/bangladesh/2020/08/24/-cyberbullying-how-to-report-crimes-online.

comments from social media like YouTube, Twitter, Wikipedia, and Reddit. Different methods were used for feature representation like BOW, TF-IDF, Word-2Vec, BERT, and their combination. The XGBoost classifier with a combined feature showed an excellent result by reaching to $F1$ score of 92%. An approach for classifying comments Kurnia et al. (2020) proposed a model that used SVM as the classifier with Word2Vec embedding. Pre-processing included tokenizing, cleaning, and removal of stop words were compromised. $F1$ score of 79% was reported in this experiment while classifying comments.

Nowadays, deep learning is widely used in natural language processing tasks. Park and Fung (2017) implemented CNN models for detecting abusive tweets from Twitter. The proposed models were based on character level, word level, and the combination of character- and word-level CNN. The highest $F1$ measure was achieved using the hybrid CNN model compared to other models when classifying tweets based on racism and sexism. Cyberbullying for the English language, Iwendi et al. (2020) used deep learning algorithms, namely RNN, GRU, LSTM, and BiLSTM. The preprocessing stage consists of cleaning, tokenizing, stemming, and lemmatization. BiLSTM had achieved test accuracy of 82.18% outperformed other models.

There is not much research done for Bengali sentiment analysis on social media due to insufficient data. Some works are done by collecting a small amount of data manually with different ML approaches. To detect abusive comments in social media, Awal et al. (2018) proposed a classifier using the Naïve Bayes algorithm. In this approach, English comments were collected from YouTube, then translating into Bangla. While pre-processing, comments were first tokenized; then, selective words were computed by preparing a bag of word (BOW) vector. The classifier reported accuracy of 80.57%. Emon et al. (2019) proposed an approach in detecting abusive texts based on a deep learning algorithm as well as compare the results with several machine learning algorithms. Different social media and news sites, namely YouTube, Prothom Alo, and Facebook, were used as a data source in this study. Pre-processing step included removal of unwanted digits, punctuation or whitespaces, stemming. For feature extraction, count vectorizer and TF-IDF vectorizer, and word embedding were used. The experimental result showed that the deep learning-based approach RNN (LSTM) achieved the highest accuracy of 82.2% and outperformed other ML algorithms such as naïve Bayes, logistic regression, random forest, and ANN. To classify sentiment and emotions in the Bangla language, Tripto and Ali (2018) built a deep learning model. This model was designed to identify Bangla sentences that were belonged to multi-labeled emotion and sentiment classes. The dataset consists of comments in Bangla, English, and romanized Bangla language from YouTube videos that were used to train the model. The Word2Vec algorithm was used for vector representation, and two models, LSTM and CNN, were used to analyze both sentiments and emotions. LSTM reported accuracy of 53% for five class sentiments and 59% accuracy for emotion classes, while CNN achieved 52% and 54% accuracy, respectively.

Transformer-based models like BERT and ELECTRA are gaining popularity day by day for NLP-related works. BERT model was applied by Yadav et al. (2020) to detect cyberbullying. Two different datasets were used to train and test the model. Reported accuracies were 98% and 96% for Formspring and Wikipedia datasets, respectively. An approach for detecting hostile posts from social media was presented by Shukla et al. (2021) using relational graph convolutional network (RGCN) with BERT embedding. Tweets in the Hindi language were collected and translated into English for training and validation. Furthermore, posts were classified into different categories like offensive, fake, hate. The proposed model achieved an $F1$ score of 97%. Logistic regression with TF-IDF and DBOW was used to build a Chatbot by Bauer et al. (2019) for detecting sexual harassment and their types. Furthermore, a fine-tuned BERT model was used for the Named entity recognition task. The proposed methods identify harassing comments with over 80% accuracy while location and dated with over 90% accuracy. For biomedical text analysis, Ozyurt (2020) used the pre-trained ELECTRA model and showed that the proposed model performs better than the Bert model. For the detection of sentiments and sarcasm from the Arabic language ELECTRA model was used by Farha and Magdy (2021). To identify the fake news spreader on Twitter, Das et al. (2020) used ensembled ELECTRA models on Spanish and English languages.

3 Preliminary and proposed framework

3.1 Transformer-based learning

Transformer-based learning brings revolutionary changes in the field of natural language processing. This architecture operates sequential inputs using an attention mechanism. Like RNNs, it also has an encoder–decoder structure. Here the encoder maps input sequence $(x_1,...,x_n)$ to a continuous representation z $(z_1,...,z_n)$ with auto-regressive steps. Lastly, the decoder generates an output sequence $(y_1,...,y_m)$. The encoder and decoder stacks have point-wise and fully connected layers and a self-attention mechanism.

Encoder and Decoder Stacks Both Stacks contains $N = 6$ layers with 2 sublayers. The first sublayer is the multi-head self-attention mechanism and the second one is about a feed-forward network, which is position-wise fully connected. For the sublayer function, *Sublayer (x)* the output of sublayers is *LayerNorm (x + Sublayer (x))* and the dimension of output

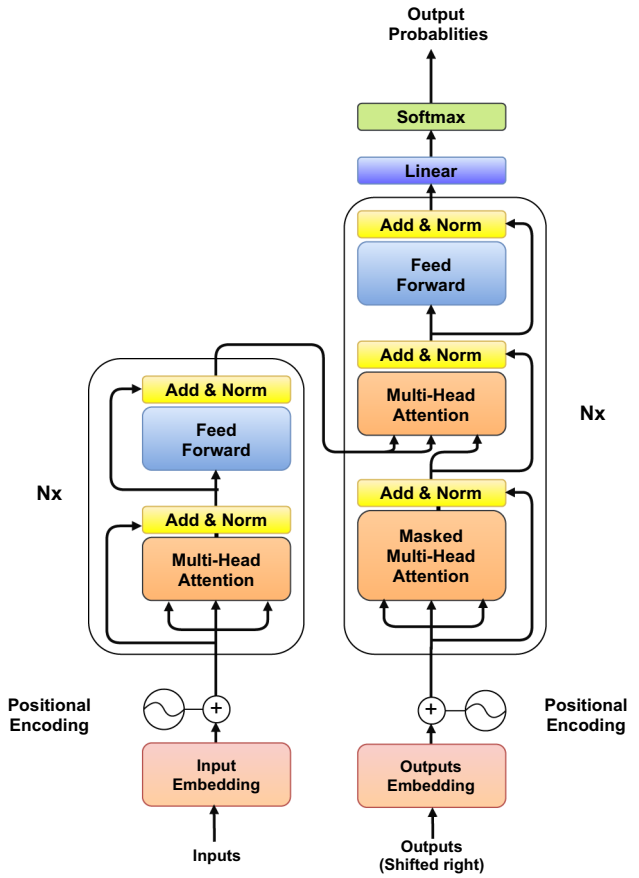


Fig. 1 Transformer model architecture (this figure's left and right halves sketch how the encoder and decoder of the transformer, respectively, work using point-wise fully connected layers with stacked self-attention)

$d_{\text{model}} = 512$. Unlike the encoder, the decoder has a third layer to apply multi-head attention to the output.

Attention Transformers utilize the multi-head self-attention mechanism. Three different uses of this attention mechanism have been implemented here. They are:

- The layered decoder passes the queries to the next, and the output of the encoder generates the memory keys and values in encoder–decoder attention layers.
- For encoder self-attention layers, all queries, keys, and values are generated from the same place, which is the output of the previous layer's encoder.
- The auto-regressive property is maintained in the decoder by preventing the leftward information flow. This is implemented inside the scaled dot product attention by masking out values (setting to -) for the softmax's input corresponding to all illegal connections.

Figure 1 shows the visual representation of transformer-based model architecture.

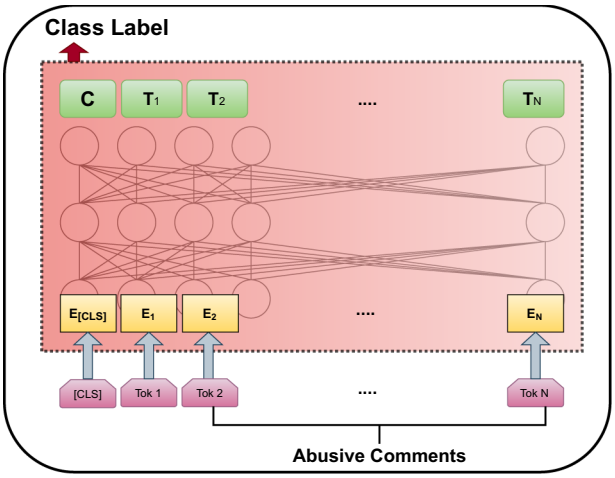


Fig. 2 BERT Architecture (to initialize the model, BERT uses the pre-trained model parameters and, during the fine-tuning, it fine-tunes the parameters. [CLS] added to the front of the input stream)

3.1.1 Bidirectional encoder representations from transformers (BERT)

BERT is a powerful transformer-based architecture that provides state-of-the-art results in various NLP tasks. It is a multilayered bidirectional transformer encoder Devlin et al. (2018). Input for BERT can be unambiguously represented as a token sequence consist of one sentence or a couple of sentences. In this sequence, the first token is the classification token [CLS]. For a couple of sentences packed together as input, and after that, BERT separates the sentence into two steps. Firstly a special token [SEP] is used. Then learning embeddings are added to each token. It indicates whether the separated sentence was the first or the second one in the packed couple.

BERT Framework has two steps, and they are pre-training and fine-tuning. These two steps are explained below.

Pre-training BERT Unlike left-to-right or right-to-left models, BERT is pre-trained as a mask language model (MLM). It has been pre-trained with unlabeled data using unsupervised learning. During this process, some input tokens are randomly masked and then predicted. It is also pre-trained to capture the relationship in coupled sentences. BERT is pre-trained with BooksCorpus (800M words) Zhu et al. (2015) and English Wikipedia's text passages (not list, headers, or tables) (2500M words).

Fine-tuning BERT For both single and coupled sentences, BERT is allowed for many downstream tasks. For that, it swaps proper input and outputs. During fine-tuning initially, BERT uses pre-trained parameters, and then, all these parameters are fine-tuned using labeled data downstream tasks. We sketch the BERT architecture in Fig. 2.

3.1.2 Multilingual BERT (mBERT)

Multilingual BERT (mBERT) Libovickỳ et al. (2019) facilitates 104 languages to be pre-trained in BERT. This architecture is able to splinter between language-neutral components and language-specific components. The probing tasks evaluated on mBERT are:

Language Identification The linear classifier is trained on the top of sentence representation and trying to identify the sentence's language.

Language Similarity On average, languages with similarities have similarities in POS tagging Pires et al. (2019). These similarities are quantified with V-measure Rosenberg and Hirschberg (2007) on language clusters by language families.

Parallel Sentence Retrieval For each sentence in parallel pair, the cosine distance between its representation and all sentence's representation on the same parallel side is computed. The sentence with the smallest distance is selected.

Word Alignment Word alignment is determined as a minimum weighted edge cover of a bipartite graph.

Machine Translation (MT) Quality Estimation The cosine distance of the source sentence's representation and MT output's reflection used to evaluate.

3.2 ELECTRA

To detect and classify abusive comments, we have utilized another transformer-based architecture, namely ELECTRA Clark et al. (2020). ELECTRA is a comparatively smaller transformer with satisfying high performance. ELECTRA uses two different neural networks, Generator G and Discriminator D. Both of them have an encoder, and it maps the input tokens' sequence $x = [x_1, \ldots, x_n]$ into contextualized vector representations' sequence $h(x) = [h_1, \ldots, h_n]$. Using the softmax layer for the generation of a specific token x_t at given position t the output is:

$$p_G(x_t \mid \mathbf{x}) = \mathbf{exp}(e(x_t)^T h_G(x)_t) / \sum_{x'} exp(e(x')^T h_G(x)_t) \tag{1}$$

Here e represents the token embeddings. The discriminator predicts the realness of token x_t at position t. That means whether this token comes from data or the generator distribution. For this prediction, it uses a sigmoid output layer given below:

$$D(x, t) = \mathbf{sigmoid}(w^T h_D(x)_t) \tag{2}$$

The generator is trained for performing MLM. For input $x = [x_1, x_2, \ldots, x_n]$, MLM mask out $m = [m_1, \ldots, m_k]^3$ by selecting a set of some random position. [MASK] token replace tokens from those positions as

$$x^{\text{masked}} = \text{REPLACE}(x, m, [\text{MASK}]) \tag{3}$$

Now the generator is able to predict the original identities for masked-out tokens. Now the discriminator learns to differentiate tokens that have been alternated in the generator. For example, if masked-out tokens replace x^{corrupt} by generator MLM, the discriminator is trained to predict the tokens in x^{corrupt} are matched with input x. The construction of these model input follows:

$$\begin{aligned} m_i &\sim \text{unif}\{1, n\} \text{ for } i = 1 \text{ to } k \\ x^{\text{masked}} &= \text{REPLACE}(x, m, [\text{MASK}]) \end{aligned} \tag{4}$$

$$\begin{aligned} \hat{x}_i &\sim p_G(x_i \mid x^{\text{masked}}) \text{ for } i \in m \\ x^{\text{corrupt}} &= \text{REPLACE}(x, m, \hat{x}) \end{aligned} \tag{5}$$

The loss functions are-

$$\mathcal{L}_{\text{MLM}}(x, \theta_G) = \mathbb{E}\left(\sum_{i \in m} -\log p_G(x_i \mid x^{\text{masked}}) \right) \tag{6}$$

$$\begin{aligned} \mathcal{L}_{\text{Disc}}(x, \theta_D) = \mathbb{E}\Bigg(\sum_{t=1}^{n} -\mathbb{1}(x_t^{\text{corrupt}} = x_t) \\ \log D(x^{\text{corrupt}}, t) - \mathbb{1}(x_t^{\text{corrupt}} \neq x_t) \\ \log(1 - D(x^{\text{corrupt}}, t)) \Bigg) \end{aligned} \tag{7}$$

Following the same process, the generator is also trained. However, training the generator is a more complicated task and has some key differences. If the generator finds any correct token, it labels it as 'real.' The training process focuses on maximum likelihood.

The maximum combined loss over a large corpus is computed by-

$$\min_{\theta_G, \theta_D} \sum_{x \epsilon \mathcal{X}} \mathcal{L}_{\text{MLM}}(x, \theta_G) + \lambda \mathcal{L}_{\text{Disc}}(x, \theta_D) \tag{8}$$

Figure 3 shows the visual presentation of the pre-train procedure of ELECTRA.

BERT and ELECTRA, both transformer-based architecture, have several layers, hidden sizes, and parameters. We broach these layers, hidden sizes, and parameters in Table 1.

3.3 Proposed framework

Figure 4 represents our proposed Framework. We have preprocessed the comment texts from the dataset before train our model. We train our model using preprocessed comments and apply transformer-based learning and enable the model to classify abusive comments. Finally, we fine-tuned the model using different values of hyperparameters. This step helps the model to predict classes more accurately.

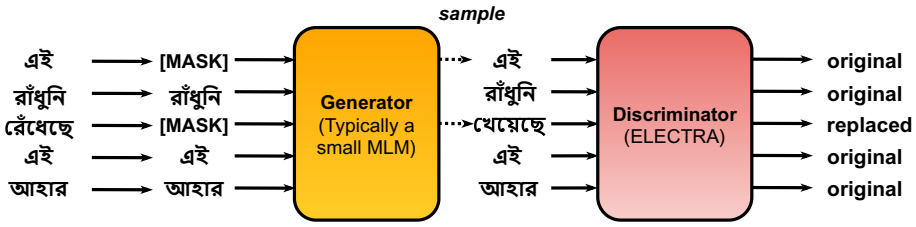


Fig. 3 Pre-training procedure of ELECTRA (it depicts how replaced tokens are detected. The generator is trained with the maximum likelihood that brings out an output distribution over tokens, and then the discriminator is fine-tuned for downstream tasks)

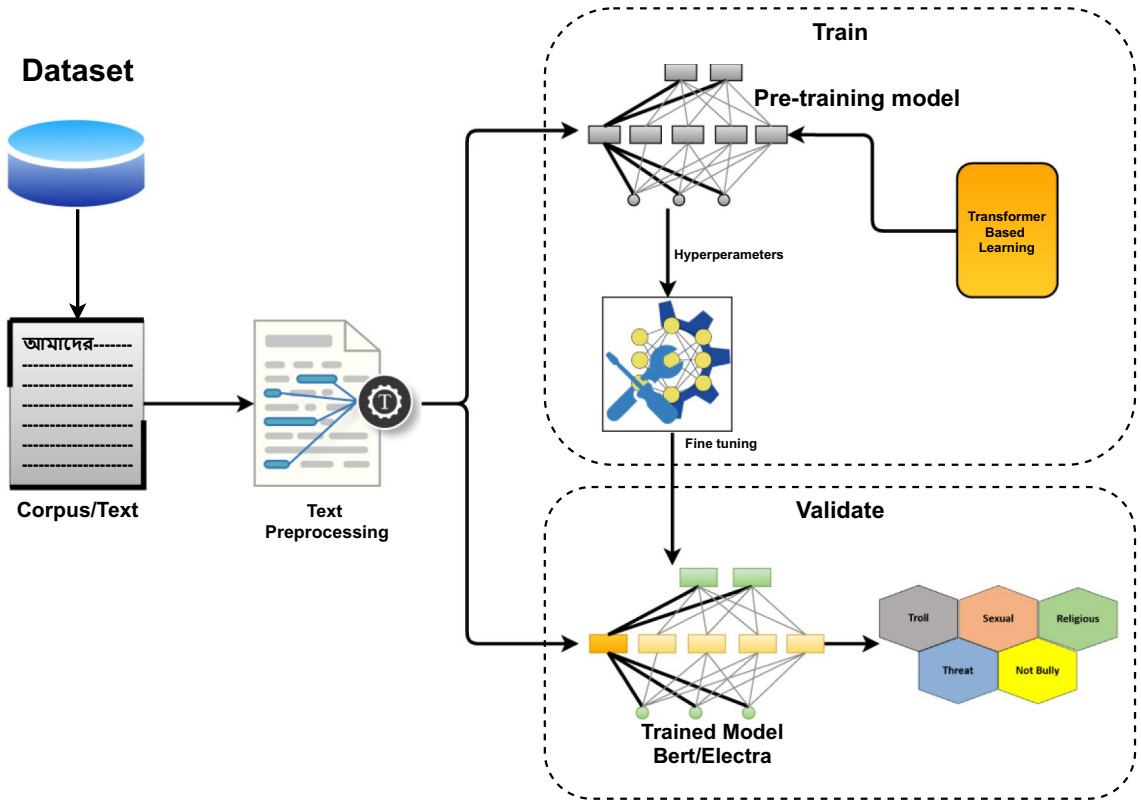


Fig. 4 Our proposed framework for classifying Bangla abusive comments (here depicts the working procedure of our classifier, starting from the text retrieving to classifying abusive comments using transformer-based architectures)

Table 1 Layers, hidden size and parameters of different models

	Layers	Hidden size	# Total parameters
BERT-Base (Multilingual cased)	12	768	110M
BERT-Large (cased)	24	1024	340M
ELECTRA-Small	12	256	14M
ELECTRA-Base	12	768	110M
ELECTRA-Large	24	1024	335M

4 Experiments and results

4.1 Environment specifications

To train deep learning models, very high computing power is needed for the parallel processing of tasks. In this regard, we have used Google Colab, which is a cloud-based Jupyter notebook platform with required hardware options such as GPU and TPU on cloud Carneiro et al. (2018). Google Colab

Table 2 Examples of different comments from the dataset with their respective classes

Text/Comment	Class
নারীর সৌন্দর্য যে নগ্নতায় নয় আবরনে, তা এখান থেকেই দৃশ্যমান।	Sexual
বাংলাই বলতে পারে না।	Not Bul
এটা কেমন গান রে ভাই, মিউজিক ভিডিওর কোনো কুল কিনারা নাই।	Troll
এটা একটা এক নাম্বারের নাস্তিকদের।	Religious
অভিযোগ দিছো ভালো কথা কিন্তু তুই ভুয়া তথ্য দিলি কেন?	Threat

Table 3 Overall dataset splitting

Class	Quantity	Training	Testing	Avg. length	Max length	Avg. token length	Max token length
Sexual	8928	7142	1786	12	160	15	196
Not Bully	15340	12272	3068	26	169	33	205
Troll	10462	8370	2092	19	173	31	191
Religious	7577	6061	1516	17	156	27	171
Threat	1694	1355	339	14	168	21	189

provides python runtime with pre-configured libraries and packages for deep learning-based tasks. It operates under Ubuntu OS with Tesla k-80 GPU of NVIDIA with 12 GB of GPU memory.

4.2 Experimental dataset

There is a scarcity of dataset which contains Bangla abusive comments in a categorized manner. Recently, Ahmed et al. (2021a) published a dataset that contains labeled comments from Facebook to aid NLP researchers. This dataset is focused on detecting comments about bully or harassment. A total of five classes of harassment that are mostly used in social media are presented here. These five classes include sexual, troll, religious, threat, and not bully. The total number of comments this dataset includes is 44001. we have also shown some examples of different comments based on their respective classes in Table 2.

Overall dataset splitting and some statistical information about our dataset with five classes show in Table 3. Here, the average non-tokenized sequence length and maximum non-tokenized sequence length are mentioned by the Average Length and Max Length column, respectively. Furthermore, the average and maximum tokenized length are pointed out to the column of Average Token Length and Max Token Length.

4.3 Data preprocessing

The dataset contains raw comments with special characters such as #, @, and - along with emoji, white spaces, HTML tag, URLs, and punctuations. These are unnecessary and removed from the texts. Since we focus on the only language, any comments or texts containing more than 20% of other languages have been removed. For feature extraction, we have applied the Tri-gram model with word tokenization. Tokenization is done in such a way that it worked efficiently with the models. Firstly, basic tokenization is applied, followed by wordpiece tokenization. Along with tokenization, we also applied Lemmatization, Stemming, and Sentence Segmentation Ahmed et al. (2020a).

4.3.1 Lemmatization based on Levenshtein distance

Bangla is a shallow orthographic language with lots of regional Regional dialect differences. There are diversities in Bangla words; even sometimes, a single word can exist in different appearances. For example, the word 'করা' can be used as 'করো', 'করে', 'করেন', 'করেছেন', 'করবেন', 'করছে', 'করবে', 'করব', etc. in different situations. For the attainment of effective results, we need to lemmatize these words in their root. In that scheme, we determine the Levenshtein distance of words and lemmatize them into root words. Levenshtein distance stipulates how dissimilar two words are from one another, which is the number of operations (insert, delete, edit) needed to effectuate for transforming one string to another. The higher value indicates higher dissimilarity.

The function mentioned in Eq. 9 is the function used to obtain Levenshtein distance. Thie Levenshtein distance function is written to designate the diversity in two words length of $|w_1|$ and $|w_2|$. We also add examples in Tables 4 and 5. The root word for 'পরিচিত' is 'পরিচয়'. 'পরিচিত' has fewer edit distances from 'পরিচিত' than word 'পরিচালনা'.

Table 4 Example of edit distance পরিচিত-পরিচয়

		প	র	ি	চ	য়
	0	1	2	3	4	5
প	1	0	1	2	3	4
র	2	1	0	1	2	3
ি	3	2	1	0	1	2
চ	4	3	2	1	0	1
ি	5	4	3	2	1	1
ত	6	5	4	3	2	2

Table 5 Example of edit distance পরিচিত-পরিচালন

		প	র	ি	চ	া	ল	া	ন
	0	1	2	3	4	5	6	7	8
প	1	0	1	2	3	4	5	6	7
র	2	1	0	1	2	3	4	5	6
ি	3	2	1	0	1	2	3	4	5
চ	4	3	2	1	0	1	2	3	4
ি	5	4	3	2	1	1	2	3	4
ত	6	5	4	3	2	2	2	3	4

$$\text{lev}_{w_1,w_2}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) \\ \min\begin{cases} \text{lev}_{w_1,w_2}(i-1,j)+1 \\ \text{lev}_{w_1,w_2}(i,j-1)+1 \\ \text{lev}_{w_1,w_2}(i-1,j-1)+1_{(w_1 \neq w_2)} \end{cases} & \text{otherwise} \end{cases} \tag{9}$$

4.4 Performance evaluation metrics

We have used different metrics to evaluate the perfection of our work. Confusion metrics visualize correct and incorrect predictions of signs, and Evolution metrics verify the model's performance. True positive (TP) and true negative (TN) indicate the correct prediction of a model. On the other hand, false positive (FP) and false negative (FN) detect wrong predictions Ahmed et al. (2021c). Using these four types of confusion metrics, we can generate numerous evolution metrics. To verify our model performance, we have to determine the accuracy, recall, precision, and $F1$ score of the model. The formulas of these evolution metrics are given here:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{10}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

$$F1\ \text{score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

4.5 Results and discussions

The results of our experiments are given in this segment. Firstly, we present the class-wise classification results for the both BERT and ELECTRA model in Tables 6 and 7 . We also demonstrate the results in terms of different learning rates. We find that, among five learning rates, both models perform better for a specific learning rate of $2e-04$. The highest scores for the metrics (precision, recall, and $F1$ score) are reported for this learning rate.

We also present the normalized confusion matrix for the BERT model in Fig. 5a. It can be seen from the confusion matrix that the highest true positive rate is 0.92, which is seen for the Troll class. Among other classes, the Sexual and Threat classes reported higher true positive rates. However, the other two classes, which are Not Bully and Religious, showed lower TP rates. For the normalized confusion matrix of the ELECTRA model in Fig. 5b, the Troll and Threat classes achieved the highest TP rates of 0.95 and 0.90, respectively. A lower TP rate is reported for Sexual class. When comparing both models, it can be seen that the Troll class has the highest TP rate for both models. Nevertheless, the Not Bully class reported the lower TP rate in the Bert model, but it is the Sexual class in the Electra model.

In Fig. 6, we showed accuracy and loss for both BERT and ELECTRA models over each epoch. From Fig. 6a, it can be seen that training accuracy over epochs for the ELECTRA model is increasing with test accuracy. After some iteration, the line graph becomes much stable. When it comes to loss, both train and test loss are gradually decreasing. When the model reaches its stable state, train and test loss is closer to zero, and it constitutes a model which is good in generalization. The highest training accuracy is 97.87%, while the test accuracy is 84.92%. In Fig. 6b, the train and test progression of the BERT model is presented for each epoch. Here we also get a stable state of both accuracy and loss. The reported maximum training accuracy is 98.09%, where the maximum test accuracy is 85.00%.

There are other variants of both BERT and ELECTRA models. To compare the performance among these variants, in Table 8 we present the performance of different variants of BERT and ELECTRA in terms of precision, recall, and

Table 6 Average classification results for BERT-Base model with different learning rate

Model	Class	Learning rate	Precision (%)	Recall (%)	*F*1-score (%)
BERT-Base	Sexual	5e−04	82.00	72.74	77.09
		4e−04	82.90	72.11	77.13
		3e−04	81.68	71.19	76.08
		2e−04	86.32	83.36	84.81
		1e−04	80.52	72.50	76.30
	Not Bully	5e−04	86.34	78.04	81.98
		4e−04	85.51	80.48	82.92
		3e−04	86.36	81.12	83.66
		2e−04	76.24	68.69	72.27
		1e−04	84.13	79.59	81.80
	Troll	5e−04	84.12	87.10	85.58
		4e−04	82.04	86.17	84.05
		3e−04	83.96	88.19	86.02
		2e−04	91.31	92.39	91.85
		1e−04	84.27	85.47	84.87
	Religious	5e−04	75.51	78.28	76.87
		4e−04	74.25	76.16	75.19
		3e−04	76.11	78.16	77.12
		2e−04	80.12	79.06	79.59
		1e−04	75.85	76.67	76.26
	Threat	5e−04	94.87	86.25	90.35
		4e−04	93.95	85.46	89.50
		3e−04	89.77	88.09	88.92
		2e−04	84.98	87.98	86.45
		1e−04	93.14	88.13	90.57

*f*1-score. We tested the models for different learning rates and got good results for the learning rate of 2e−04. From this tabular representation, it can be seen that the BERT-Base and ELECTRA-base model outperforms others. ELECTRA-large model also performed well for learning rate 2e−04.

We have also compared the results of the BERT and ELECTRA model on our dataset with other deep learning-based approaches in Fig. 7. LSTM, Bi-LSTM, LSTM-GRU, Graph CN is implemented alongside the BERT and ELECTRA model variants. This representation shows that the LSTM model performed with the lowest test accuracy of 76.9%. The highest test accuracy is 85% which is observed for the BERT-base model. In the meantime, the ELECTRA-Base model reached the second-highest test accuracy of 84.92%. We present the results for the learning rate of 2e−04 since we find that all the models performed well compared to other learning rates on this learning rate.

4.6 Observation of BERT and ELECTRA architecture

In this subsection, we will investigate the performance of the BERT and ELECTRA model on datasets containing different classes of text or comments, or sentences. We used three open-source dataset for this purpose, named Prothom-Alo[4], BARD Alam and Islam (2018) and OSBC[5] dataset. These datasets include sentences of different classes like sports, entertainment, politics, economy, technology, crime, art, opinion, education, etc.

In Tables 9 and 10, we showed the obtained results of our BERT-base and ELECTRA-base models. It can be seen that both the models performed well while classifying different classes of Bangla texts. For the ProthomAlo dataset, BERT and ELECTRA models achieved a classification accuracy of 97.23% and 95.82%. On the other hand, for BARD and OSBC datasets, the ELECTRA model performed slightly better than the BERT model. Other metrics like precision, recall, and *f*1-score are also significant for NLP tasks like text classification. The signification is that both BERT and ELECTRA models have excellent generalization capabilities while classifying Bengali texts.

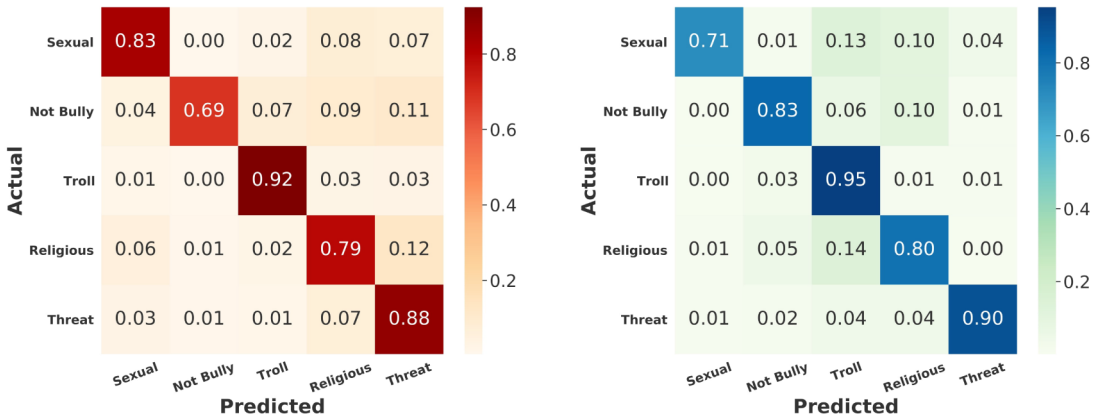
Transformer-based models avoid recursion, process the texts as a whole. It automatically extracts the relationship

[4] https://www.kaggle.com/twintyone/prothomalo.

[5] https://scdnlab.com/corpus/.

Table 7 Class-wise classification result of ELECTRA-Base with learning rate

Model	Class	Learning rate	Precision (%)	Recall (%)	*F*1-score (%)
ELECTRA-Base	Sexual	5e−04	83.14	72.40	77.39
		4e−04	81.26	73.14	76.98
		3e−04	84.35	71.33	77.29
		2e−04	84.56	71.30	77.36
		1e−04	82.58	70.69	76.17
	Not Bully	5e−04	86.32	79.91	82.99
		4e−04	87.01	80.09	83.40
		3e−04	86.97	81.61	84.20
		2e−04	87.17	82.63	84.83
		1e−04	85.43	79.82	82.52
	Troll	5e−04	83.66	87.05	85.32
		4e−04	84.72	86.17	85.43
		3e−04	80.83	88.68	84.57
		2e−04	84.35	88.95	86.58
		1e−04	83.15	85.70	84.40
	Religious	5e−04	72.27	78.28	75.15
		4e−04	71.67	76.39	73.95
		3e−04	73.86	78.21	75.97
		2e−04	76.87	79.78	78.29
		1e−04	75.28	74.93	75.10
	Threat	5e−04	93.55	87.66	90.50
		4e−04	94.63	88.70	91.56
		3e−04	91.50	86.74	89.05
		2e−04	95.96	89.63	92.68
		1e−04	92.97	88.37	90.61



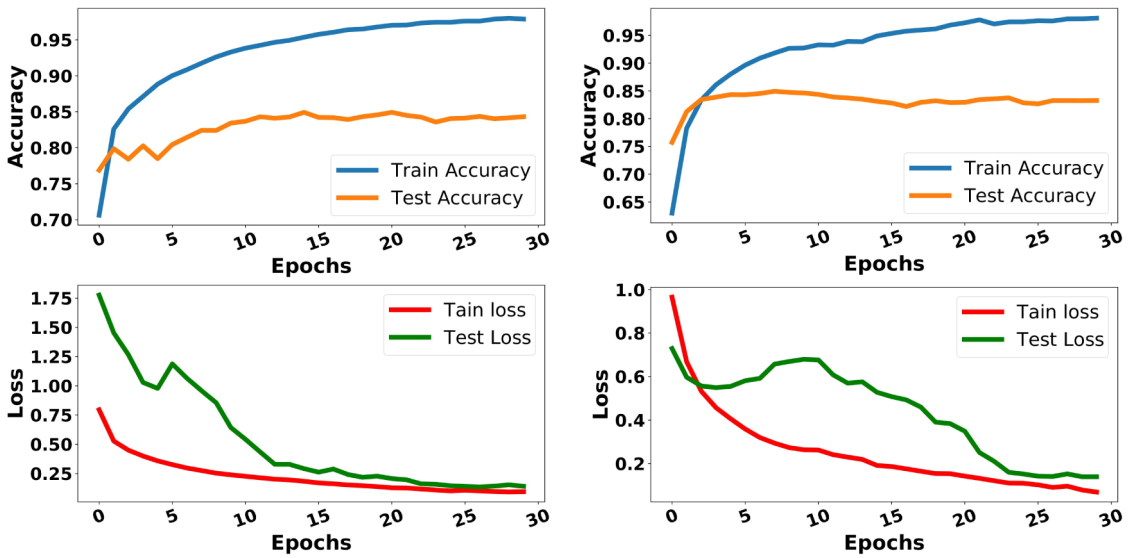
(a) BERT-Base (b) ELECTRA-base

Fig. 5 Confusion matrix for BERT-Multilingual and ELECTRA-Base models in a normalized form

between words by employing techniques like multi-head attention and positional embeddings. BERT is multilingual and is trained on Wikipedia corpus. In the data sampling phase, weights are adjusted exponentially, and words from low resource language like Bangla are represented better than other models like CNN or RNN. On the other hand, ELECTRA utilizes the replaced token detection (RTD) method, which works more efficiently than BERT in terms of computation performance. A generator network is employed to replace the tokens with alternative samples. Unlike the

Table 8 Comparison between different variants of BERT and ELECTRA models with different learning rates

Models	Learning rate	Precision (%)	Recall (%)	*F*1-score (%)
BERT-Base	5e−04	**84.57**	80.48	82.47
	4e−04	83.73	80.08	81.86
	3e−04	83.58	81.35	82.45
	2e−04	**83.79**	**82.3**	**83.04**
	1e−04	83.58	80.47	82.00
BERT-Large	5e−04	81.55	81.03	81.29
	4e−04	80.08	80.22	80.15
	3e−04	79.16	79.57	79.36
	2e−04	82.29	81.66	81.97
	1e−04	78.86	79.04	78.95
ELECTRA-Small	5e−04	77.07	76.8	76.93
	4e−04	76.61	76.8	76.70
	3e−04	75.89	75.95	75.92
	2e−04	77.6	76.95	77.27
	1e−04	75.4	75.25	75.32
ELECTRA-Base	5e−04	83.79	81.06	82.40
	4e−04	83.86	80.9	82.35
	3e−04	83.5	81.31	82.39
	2e−04	**85.78**	**82.46**	**84.09**
	1e−04	83.88	79.9	81.84
ELECTRA-Large	5e−04	83.59	**82.18**	82.88
	4e−04	83.52	81.59	82.54
	3e−04	83.94	80.81	82.35
	2e−04	84.09	**83.01**	**83.55**
	1e−04	83.48	80.4	81.91



(a) ELECTRA-Base (b) BERT-Base

Fig. 6 Accuracy and Loss for ELECTRA and BERT model over epochs (We run our models for 30 epochs. For each epoch, we get the training progression of our models in terms of accuracy and loss, which is represented).

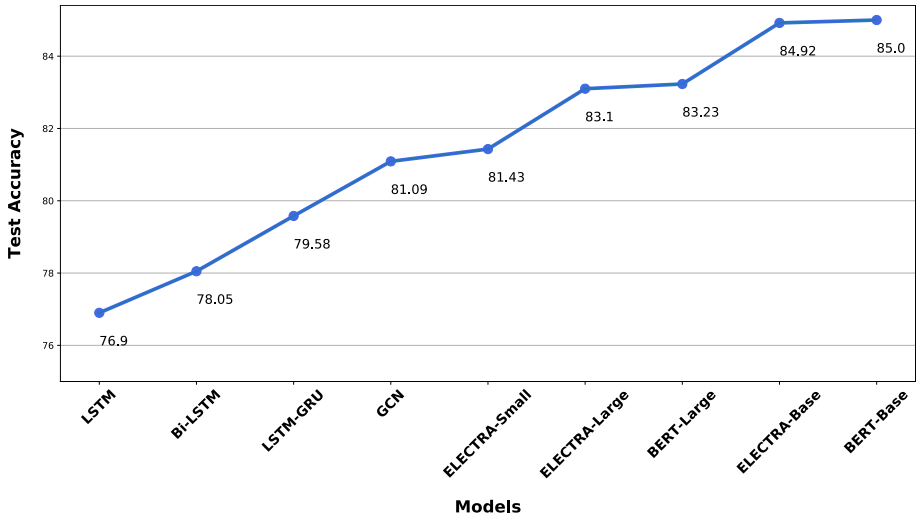


Fig. 7 Test accuracy comparison between different models.

Table 9 Performance of BERT-base model on datasets containing differnt classes of Bengali texts

Dataset	Accuracy (%)	Precision (%)	Recall (%)	*F*1-Score (%)
Prothom-Alo	97.23	93.96	92.78	93.36
BARD	93.56	89.10	90.72	89.90
OSBC	79.80	75.06	71.41	73.18

Table 10 Performance of ELECTRA-base model on datasets containing differnt classes of Bengali texts

Dataset	Accuracy (%)	Precision (%)	Recall (%)	*F*1-Score (%)
Prothom-Alo	95.82	95.03	93.55	94.28
BARD	94.25	93.29	91.84	92.55
OSBC	80.67	77.17	74.23	75.67

BERT model, which uses MASK modeling, the ELECTRA model replaces tokens with plausible or fake samples. This strategy helps the network to learn a better representation of words.

To the best of our knowledge, the research effort we presented in this paper is unique. No works are focusing on Bangla comments on Facebook. BERT and ELECTRA models were used previously for languages like English, German, Arabic, etc. However, our study emphasizes Facebook data for a specific Bangla language, which is novel. Besides, our experiment is conducted on a more structured dataset with an immense collection of data. We also presented the effects of the different learning rates for the pre-training of transformer-based models. Considering all our findings and results gives us the confidence that our proposed approach can accurately detect Bangla abusive comments on Facebook and other social media platform.

5 Conclusion and future work

We attempt to bring out an automated intelligent solution to the latest increasing cyberbullying issues in Bangladesh. We proposed a transformer-based system that is capable of classifying abusive comments written in the Bangla language. Two latest transformer-based architectures, BERT and ELECTRA, are implemented for the Bangla language here. These two efficient architectures bring out remarkable accuracy in our experiment. Furthermore, we conducted our experiments on real-world abusive comments taken from social media (Facebook). To justify our classifier, we mention some evolution processes. We also determine the related confusion matrix and evolution matrix based on our classifier's predictions. The value of precision, recall, and $f1$-score for different classes indicates how correctly our model classifying the abusive comments. We also show the loss of BERT and ELECTRA over each epoch for our experiment. Then shows the comparison of different deep learning architecture accuracy to bring out the performance of our proposed models.

In the future, to increase the efficiency of our classifier, we want to train it with other regional language forms of Bangla. Our additional focus is to identify abusive comments at an initial stage for any application with the power of REST API and GraphQL. We also plan to develop an automated apparatus to detect spam users and block or report those users.

References

Adhikari A, Ram A, Tang R, et al (2019) Docbert: bert for document classification. arXiv preprint arXiv:1904.08398

Ahmed MF, Mahmud Z, Biash ZT, et al (2021a) Bangla text dataset and exploratory analysis for online harassment detection. arXiv preprint arXiv:2102.02478

Ahmed MS, Aurpa TT, Anwar MM (2020a) Online topical clusters detection for top-k trending topics in twitter. In: 2020 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), IEEE, pp 573–577

Ahmed MS, Aurpa TT, Anwar MM (2020b) Query oriented topical clusters detection for top-k trending topics in twitter. In: 2020 IEEE 8th R10 humanitarian technology conference (R10-HTC), IEEE, pp 1–6

Ahmed MS, Aurpa TT, Anwar MM (2021) Detecting sentiment dynamics and clusters of twitter users for trending topics in covid-19 pandemic. Plos one 16(8):e0253300

Ahmed MS, Aurpa TT, Azad MAK (2021c) Fish disease detection using image based machine learning technique in aquaculture. J King Saud Univ-Comput Inf Sci

Al-Twairesh N (2021) The evolution of language models applied to emotion analysis of Arabic tweets. Information 12(2):84

Alam MT, Islam MM (2018) Bard: Bangla article classification using a new comprehensive dataset. In: 2018 international conference on bangla speech and language processing (ICBSLP), IEEE, pp 1–5

Alzamzami F, Hoda M, El Saddik A (2020) Light gradient boosting machine for general sentiment classification on short texts: a comparative evaluation. IEEE Access 8:101840–101858

Ashrafi I, Mohammad M, Mauree AS et al (2020) Banner: a cost-sensitive contextualized model for Bangla named entity recognition. IEEE Access 8:58206–58226

Awal MA, Rahman MS, Rabbi J (2018) Detecting abusive comments in discussion threads using Naïve Bayes. In 2018 international conference on innovations in science, engineering and technology (ICISET), IEEE, pp 163–167

Bauer T, Devrim E, Glazunov M, et al (2019) # metoomaastricht: building a chatbot to assist survivors of sexual harassment. In: Joint European conference on machine learning and knowledge discovery in databases, Springer, pp 503–521

Carneiro T, Da Nóbrega RVM, Nepomuceno T et al (2018) Performance analysis of google colaboratory as a tool for accelerating deep learning applications. IEEE Access 6:61677–61685

Chia YK, Witteveen S, Andrews M (2019) Transformer to cnn: Label-scarce distillation for efficient text classification. arXiv preprint arXiv:1909.03508

Clark K, Luong MT, Le QV, et al (2020) Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555

Das KA, Baruah A, Barbhuiya FA, et al (2020) Ensemble of electra for profiling fake news spreaders. In: CLEF

Devlin J, Chang MW, Lee K, et al (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

Emon EA, Rahman S, Banarjee J, et al (2019) A deep learning approach to detect abusive bengali text. In: 2019 7th international conference on smart computing and communications (ICSCC), IEEE, pp 1–5

Farha IA, Magdy W (2021) Benchmarking transformer-based language models for arabic sentiment and sarcasm detection. In: Proceedings of the sixth Arabic natural language processing workshop, pp 21–31

Iwendi C, Srivastava G, Khan S, et al (2020) Cyberbullying detection solutions based on deep learning architectures. Multimed Syst:1–14

Janardhana D, Shetty AB, Hegde MN, et al (2021) Abusive comments classification in social media using neural networks. In: International conference on innovative computing and communications, Springer, pp 439–444

Kurnia R, Tangkuman Y, Girsang A (2020) Classification of user comment using word2vec and SVM classifier. Int J Adv Trends Comput Sci Eng 9:643–648

Li X, Bing L, Zhang W, et al (2019) Exploiting bert for end-to-end aspect-based sentiment analysis. arXiv preprint arXiv:1910.00883

Libovický J, Rosa R, Fraser A (2019) How language-neutral is multilingual bert? arXiv preprint arXiv:1911.03310

Nobata C, Tetreault J, Thomas A, et al (2016) Abusive language detection in online user content. In: Proceedings of the 25th international conference on world wide web, pp 145–153

Ostendorff M, Bourgonje P, Berger M, et al (2019) Enriching bert with knowledge graph embeddings for document classification. arXiv preprint arXiv:1909.08402

Ozyurt IB (2020) On the effectiveness of small, discriminatively pre-trained language representation models for biomedical text mining. In: Proceedings of the first workshop on scholarly document processing, pp 104–112

Park JH, Fung P (2017) One-step and two-step classification for abusive language detection on twitter. arXiv preprint arXiv:1706.01206

Pericherla S, Ilavarasan E (20218) Performance analysis of word embeddings for cyberbullying detection. In: IOP conference series: materials science and engineering, IEEEOP Publishing, p 012008

Pires T, Schlinger E, Garrette D (2019) How multilingual is multilingual bert? arXiv preprint arXiv:1906.01502

Rosenberg A, Hirschberg J (2007) V-measure: A conditional entropy-based external cluster evaluation measure. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), pp 410–420

Salminen J, Hopf M, Chowdhury SA et al (2020) Developing an online hate classifier for multiple social media platforms. Human-Cent Comput Inf Sci 10(1):1–34

Salur MU, Aydin I (2020) A novel hybrid deep learning model for sentiment classification. IEEE Access 8:58080–58093

Samad MD, Khounviengxay ND, Witherow MA (2020) Effect of text processing steps on twitter sentiment classification using word embedding. arXiv preprint arXiv:2007.13027

Shukla S, Mittal G, Arya KV, et al (2021) Detecting hostile posts using relational graph convolutional network. arXiv preprint arXiv:2101.03485

Souza F, Nogueira R, Lotufo R (2019) Portuguese named entity recognition using bert-crf. arXiv preprint arXiv:1909.10649

Su J, Yu S, Luo D (2020) Enhancing aspect-based sentiment analysis with capsule network. IEEE Access 8:100551–100561

Tripto NI, Ali ME (2018) Detecting multilabel sentiment and emotions from bangla youtube comments. In: 2018 international conference on Bangla speech and language processing (ICBSLP), IEEE, pp 1–6

Xu H, Liu B, Shu L, et al (2020) Dombert: Domain-oriented language model for aspect-based sentiment analysis. arXiv preprint arXiv:2004.138167

Xue K, Zhou Y, Ma Z, et al (2019) Fine-tuning bert for joint entity and relation extraction in chinese medical text. In: 2019 IEEE International conference on bioinformatics and biomedicine (BIBM), IEEE, pp 892–897

Yadav J, Kumar D, Chauhan D (2020) Cyberbullying detection using pre-trained bert model. In: 2020 International conference on electronics and sustainable communication systems (ICESC), IEEE, pp 1096–1100

Yu J, Jiang J (2019) Adapting bert for target-oriented multimodal sentiment classification. IJCAI

Yu S, Su J, Luo D (2019) Improving bert-based text classification with auxiliary sentence and domain knowledge. IEEE Access 7:176600–176612

Yuan C (2019) Bb-kbqa: Bert-based knowledge base question answering. In: Chinese computational linguistics: 18th China national conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings, Springer Nature, p 81

Zhang H, Sun S, Hu Y et al (2020) Sentiment classification for Chinese text based on interactive multitask learning. IEEE Access 8:129626–129635

Zhu Y, Kiros R, Zemel R, et al (2015) Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE international conference on computer vision, pp 19–27