SciEn
Journal of Engineering Advancements Vol. 05(02) 2024, pp 27-34 https://doi.org/10.38032/jea.2024.02.001

Depression Intensity Identification using Transformer Ensemble Technique for the Resource-constrained Bengali Language

Md. Nesarul Hoque[1,*] *Umme Salma*[2], *Md. Jamal Uddin*[1], *Sadia Afrin Shampa*[2]

[1]Department of Computer Science and Engineering, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj-8100, Bangladesh
[2]Department of Computer Science & Engineering, Bangladesh University, Dhaka-1207, Bangladesh

Received: December 29, 2023, Revised: March 25, 2024, Accepted: April 21, 2024, Available Online: May 10, 2024

ABSTRACT

Depression is an ordinary mental health-related disorder that hampers people's daily activities, and sometimes, it destroys an individual's life. It is one of the major social issues at present. Since depressed people use various social networking sites for sharing their thoughts and feelings, many scholars have tried to identify depression texts in highly resourced languages like English; however, only a small quantity of papers are detected in the resource-constrained Bengali language. This paper focuses on developing a depression intensity detection system from Bengali text data. In this regard, this study experiments on a 2,596 sample-sized dataset with four levels of depression by utilizing five state-of-the-art transformer models, including multilingual Bidirectional Encoder Representations from Transformers, DistilmBERT, XLM-RoBERTa, Bangla-BERT-Base, and BanglaBERT, and suggests a new ensemble method called MaxOfAvgProb. This method goes beyond the performance of the previous work on the same dataset, scoring 63.47% F1-score and 62.90% accuracy. To increase the reliability of the proposed method, we utilize this approach on another available dataset with 4,897 entries. In this case, our recommended method also surpasses the performance of the existing work on the same dataset, with accuracy at 86.45% and F1-score at 86.35%. Identifying the intensity of depression, depressed people may get proper counseling or treatment from their respected guardians or psychologists according to the victims' level of depression.

Keywords: Bengali, Classification, Depression, Resource-constrained Language, Transformer Ensemble.

1 Introduction

Depression is a prevalent and severe mental condition that exerts a significant effect on individuals. It undermines people's self-worth, positive thinking, and motivation, making it more challenging to make progress towards self-improvement. It is also a chronic illness that disrupts day-to-day functioning. Depression can be influenced by a complex interplay of factors, including genetic predisposition, adverse life experiences (including both traumatic events and childhood trauma), social isolation, and certain personality traits. Worldwide, depression is considered the leading cause of suicide. Furthermore, more than 70 million people lose their lives to suicide every year. Moreover, the fourth most common cause of mortality for people aged 15 to 29 is suicide [1].

The worldwide landscape reveals a concerning occurrence of moderate to severe levels of depression, which stands at 43%, with anxiety affecting 63% and stress impacting 41% of the population [2]. While the discourse on mental health issues may not be as widespread, the growing threat is particularly evident in low- and middle-income countries such as Bangladesh.

In the context of mental health in Bangladesh, as indicated in [3], the range of prevalence for mental illnesses in adults is 6.5% to 31.0%, and for children, it is 13.4% to 22.9%. According to a recent study, over 7 million people in Bangladesh, mostly in urban and metropolitan areas, struggle with anxiety and depression, which highlights the severity of the issue [4]. Numerous studies have demonstrated the high frequency of depression among college and university students, which is noteworthy [5]. Moreover, a survey emphasizes the dual effect, showing that 24% of students experience both anxiety and depression at the same time [6].

At present, people from all around the world share their sorrows and pains on different social media platforms, such as Facebook, Twitter, and others. The people of Bangladesh are no exception to this use. As a result, we concentrate on text data in Bengali that has been collected from well-known social networking platforms like Facebook. Many machine learning (ML), deep learning (DL), and transformer models are exploited to classify Bengali text data in different domains like cyberbullying [7], accident-related news [8], and others. We have also identified numerous scholarly works regarding the detection of depression. However, the majority of these studies used traditional ML and DL classifiers to study binary classification problems (depression or not) [9]–[13]. A few investigations dealt with multi-class classification problems from emotional perspectives, such as anger, fear, disgust, sadness, joy, surprise, and others [14], [15]. Moreover, a few scholars developed depression severity detection systems [15]–[17]. Consequently, this research employs five transformer models, including multilingual Bidirectional Encoder Representations from Transformers (mBERT), DistilmBERT, XLM-RoBERTa (XLM-R), Bangla-BERT-Base, and BanglaBERT, and designs an effective depression intensity recognition system using a transformer ensemble method, MaxOfAvgProb, from the Bengali text data. The key contributions of this paper are outlined below:

- Rigorous experimenting on the Bengali depression severity dataset with four classes.

*Corresponding Author Email Address: mnhshisir@gmail.com

- Leveraging cutting-edge transformer models with parameter tuning for developing a reliable depression classification system.
- Proposing an effecting Transformer-ensemble model, MaxOfAvgProb, for recognizing four severity levels of depression.
- Proposed approach surpasses the existing state-of-the-art works for detecting the intensity of depression.

The remaining parts of the paper are structured as follows. Section 2 explains a comparative analysis of the recent related works. Afterward, the system's working process, including dataset statistics, preprocessing, feature selection, and an overview of the transformer models, is demonstrated in Section 3. Then, Section 4 discusses the working environment and implementation setup of the classification system. Next, the comparative result analysis and detailed discussion of the proposed method are illustrated in Section 5. Finally, Section 6 gives concluding statements and provides plans for this research domain.

2 Related Works

Detecting a depressed individual from a text containing depression material is a recent research topic at present. Even though researchers have made significant efforts to develop depression detection systems for English or other high-resource languages, Bengali is still far behind. This section discusses the recent studies regarding depression identification systems from Bengali content and other than Bengali content.

2.1 Depression Detection from the Not Bengali Content

Most of the research dealt with English text data for recognizing depressive content. Khan and Alqahtani [18] investigated Twitter data for detecting signs of depression. The authors experimented with four approaches, where the Logistic Regression (LR) with Term Frequency and Inverse Document Frequency (TF-IDF) method performed the maximum accuracy of 99.40%. Mustafa et al. [19] utilized the Linguistic Inquiry and Word Count (LIWC) English text analysis tool for classifying three severity levels of depressive posts: high, medium, and low. De et al. [20] investigated a depression dataset collected from Twitter and Reddit social networking sites. The authors proposed a novel profile-based sentiment-aware model that achieved better performance. Chiu et al. [21] focused on sentiment analysis of social media sites like Instagram and paid more attention to diagnosing depression. A multimodal system was suggested for classifying posts as depressed or not by using characteristics from images, language, and behavior and obtained an 83.50% F1-score. Abd et al. [22] demonstrated a hybrid approach consisting of text mining and neural networks for positive and negative sentiment classification. They obtained the highest precision of 83.60%, with a sensitivity of 87.10% and a specificity of 79.30%. Paul et al. [23] proposed a binary classification model for analyzing depressive data. The authors applied different ML and DL methods, in which the combined method of Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU) with the Bag-of-Words feature extraction technique performed the best accuracy score at 87.11%. Soliman et al. [24] worked on Arabic slang comments. They proposed Slang Sentimental Words and Idioms Lexicon (SSWIL), a sentiment analysis approach to unstructured and ungrammatical customers' slang content. They utilized the Support Vector Machine with Gaussian kernel and achieved a better output with 86.86% accuracy.

2.2 Depression Detection from the Bengali Content

Most of the research dealt with a binary classification issue (depression or not depression). Uddin et al. [9] proposed a Long Short-Term Memory (LSTM) based depression identification system from the Bengali social media data. They achieved a higher accuracy score of 86.30%. Mumu et al. [10] used the hybrid CNN-LSTM model to successfully identify a depressive text with an accuracy of 81.05%. Mohammad et al. [11] employed an Extra Tree classifier for feature extraction and utilized the Principal Component Analysis technique to minimize feature dimensionality. Afterward, the authors applied the eXtreme Gradient Boost classifier, showing the highest 92.80% accuracy and 93.61% F1-score. Ghosh et al. [12] proposed an LSTM-GRU-CNN hybrid approach for recognizing depressive texts from social media posts. They showed significant detection performance by achieving an accuracy of 92.25%, sensitivity of 94.46%, and specificity of 91.15%.

A limited study is detected for identifying the level of depression. Ahmed et al. [13] developed two questionnaire-based datasets concerning depression and anxiety-type mental disorders. The authors collected samples from the Bangladeshi people. The authors prepared 30 questions for measuring the level of depression and 35 questions for anxiety. They got the maximum accuracy score of 96.80% for depression and 96.00% for anxiety by utilizing the CNN model. Another questionnaire-based dataset was made by Siddiqui et al. [8] from the people of the same country. The author selected 106 questions to build a 520 sample-sized dataset leveled with Normal, Moderate, and Extreme classes using the voting method of eight recognized depression scales. They employed ten ML and two DL models to classify depression levels. They also investigated the optimal values of five hyperparameters and utilized nine feature selection techniques for better prediction performance. Finally, they interpreted the significant factors regarding the severity of depression using the Local Interpretable Model-Agnostic Explanations (LIME) framework. Besides the questionnaire datasets, some researchers detected the intensity of depression from text data from various online blogging sites and social networking platforms. Hossen et al. [15] worked on binary and two multi-class classification (four classes and six classes) issues for the Facebook posts. The authors proposed LR with the TF-IDF technique for the multi-classification problems, while in binary classification, they suggested LSTM with a word-embedding approach. Kabir et al. [16] developed a depression severity detection system with four classes. They accumulated data from various online blogs and social media groups. The authors recommended a bidirectional GRU (BiGRU)-based method for achieving the best outputs (F1-score and accuracy are 81.00%). Hoque and Salma [17] investigated a depression severity detection system with four classes: Not Depression, Mild, Moderate, and Severe. They experimented on several ML, DL, and transformer models. However, XLM-R presented the highest performance, with an accuracy of 60.89% and an F1-score of 61.11%.

3 Materials and Methods

The research investigates a Bengali depressive dataset of [17] for developing an effective depression severity detection system.

This section covers four points: dataset overview, data preprocessing, feature selection, and discussion of transformer models for explaining the overall procedures. The entire working process is depicted in Fig. 1.

3.1 Dataset Description

The study considers the same Bengali depression corpus [17]. We name this dataset DSD-1 for the later usages throughout the paper. The samples were collected from users' profiles and various Facebook pages. The DSD-1 has two fields: Text and Label. The Text column contains the Facebook posts of the users, and the Label column indicates the four intensity levels of depressive text, including Not Depression, Mild, Moderate, and Severe. The overall statistics regarding this dataset are articulated in Table 1, where *qty* presents the number of instances, *percent* indicates the percentage of samples belonging to the classes, and *minWord*, *maxWord*, and *avgWord*, respectively, point the minimum, maximum, and average number of words in a sample.

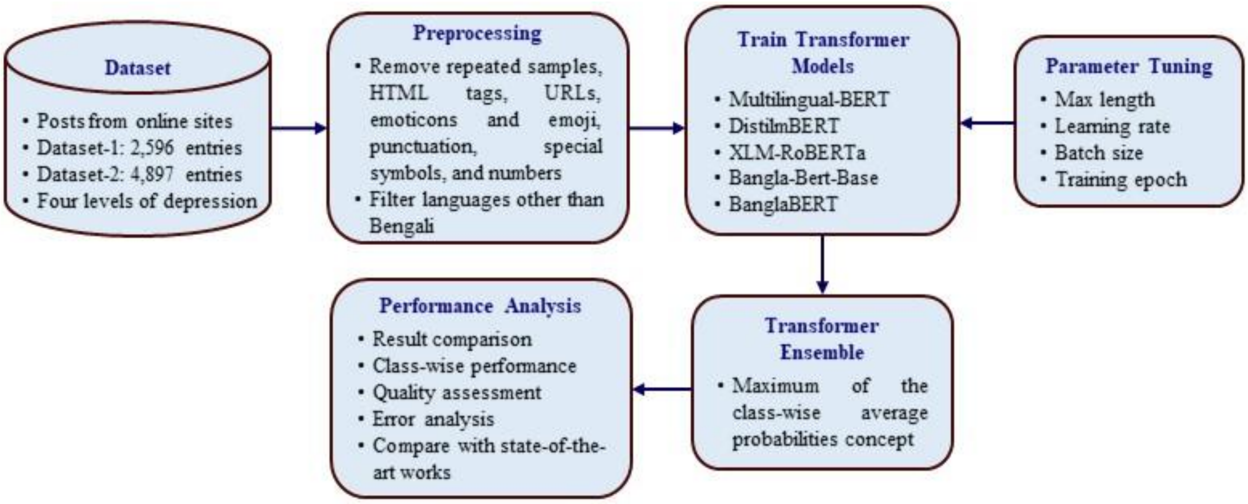

Dataset
• Posts from online sites
• Dataset-1: 2,596 entries
• Dataset-2: 4,897 entries
• Four levels of depression
Preprocessing
• Remove repeated samples, HTML tags, URLs, emoticons and emoji, punctuation, special symbols, and numbers
• Filter languages other than Bengali
Train Transformer Models
• Multilingual-BERT
• DistilmBERT
• XLM-RoBERTa
• Bangla-Bert-Base
• BanglaBERT
Parameter Tuning
• Max length
• Learning rate
• Batch size
• Training epoch
Transformer Ensemble
• Maximum of the class-wise average probabilities concept
Performance Analysis
• Result comparison
• Class-wise performance
• Quality assessment
• Error analysis
• Compare with state-of-the-art works

Fig. 1 Overview of the proposed system

Table 1 Statistical information of the DSD-1 dataset [17].

Class	*qty*	*percent*	*maxWord*	*minWord*	*avgWord*
Not Depression	949	36.56%	115	4	19
Mild	775	29.85%	172	4	23
Moderate	608	23.42%	261	4	34
Severe	264	10.17%	245	4	40
All Classes	**2596**	**100%**	**261**	**4**	**29**

This research also employs another depression severity dataset from [16] to justify our proposed method. We name this corpus DSD-2 for the later usage of the remaining parts of the paper. The DSD-2 consists of 4,897 entries labeled with four levels of depression: Level 1, Level 2, Level 3, and Level 4. Level 1 and Level 4 represent the lowest and the highest intensity of depression, respectively.

3.2 Data Preprocessing

The noisy and redundant elements are filtered from the entire dataset by applying various preprocessing tasks. At first, identical samples are identified and are discarded. After that, other unnecessary content like HTML tags, URLs, emoji and emoticon characters, punctuation, special symbols, and numbers are eliminated from every post. Since this research focuses on Bengali data, languages other than Bengali are removed. Furthermore, excluding stop-word filtering and stemming [25] operations improves the classification performance in many Bengali Natural Language Processing (NLP)-related tasks [7], [17]. Thus, these two preprocessing operations are not considered in this paper.

3.3 Feature Selection

Every transformer model utilizes self-tokenization and embedding techniques for extracting feature vectors. The mBERT, Bangla-BERT-Base, DistilmBERT, and BanglaBERT exploit the WordPiece method, while XLM-R-base applies the Sentence Piece Model (SPM) [26] technique. The SMP consists of two segmentation algorithms: the uni-gram language model [27] and byte-pair-encoding (BPE) [28].

3.4 Transformer Models

The paper prefers five state-of-the-art transformer models, including mBERT, DistilmBERT, Bangla-BERT-Base, XLM-R-base, and BanglaBERT that show better outputs in depression [17] and other Bengali NLP-related downstream tasks [7], [29]. The mBERT [30] follows a similar working principle to the BERT [31]. This model pre-trained over 104 languages instead of just English. Its pre-training and fine-tuning frameworks use multi-head attention layers, where each attention head is calculated by following Eq. (1) and multi-head computed by Eqs. (2) and (3) [32]. This model achieves an intuitive knowledge of a particular language by the Mask Language Model (MLM) and Next Sentence Prediction (NSP) objectives. DistilmBERT is another pre-trained model like mBERT. It uses knowledge distillation during the pre-training phase. Thus, it takes less time and less memory space in the time of model execution. Another pre-trained model, XLM-R-base [33], a cross-lingual model, has been trained over a large corpus of more than 2 TB (terabytes) of multi-lingual data. This model performs well for resource-constrained languages like Bengali. This paper also utilizes two Bengali language-specific transformer models. One is Bangla-BERT-Base [34], which follows the MLM objective of the original BERT model, and the second one is BanglaBERT [35], derived from another transformer model called ELECTRA [36]. The model architecture of these five models is notified in Table 2, where *L*, *E*, *H*, *FF*, *A*, *V*, *G2D*, and *P* points transformer layers, embedding size, hidden size, intermediate feed-forward layer size, attention head, vocabulary size, generator-to-discriminator ratio, and number of parameters, respectively.

$$Attention\ (Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where, *Q*, *K*, and *V* point the query, key, and value matrices, respectively, d_k presents the dimension of *Q* and *K*, and $\frac{1}{\sqrt{d_k}}$ indicates the scaling factor.

$$MultiHead(Q,K,V) = Concat(head_1,\dots,head_h)W^O \quad (2)$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (3)$$

where, *h* is the number of attention head, W^O presents output weight of the attention unit, and W_i^Q, W_i^K, and W_i^V indicates the i^{th} attention weight of *Q*, *K*, and *V* matrices, respectively.

Table 2 Architecture of the transformer models

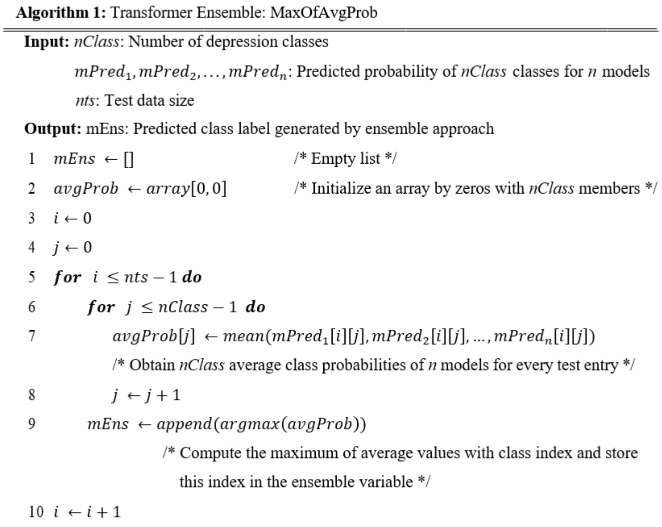
Model	*L*	*E*	*H*	*FF*	*A*	*V*	*G2D*	*P*
mBERT	12	-	768	3072	12	110k	-	172M
DistilmBERT	6	-	768	3072	12	120k	-	134M
XLM-R-Base	12	-	768	3072	12	250k	-	270M
Bangla-BERT-Base	12	-	768	3072	12	102k	-	110M
BanglaBERT	12	768	768	3072	12	32k	1/3	110M

Transformer Ensemble: Generally, the performance of the combined model exhibits better results than the individual one [37]. This research proposes a new ensemble technique named MaxOfAvgProb, shown in Algorithm 1 (see Fig. 2). Let $mPred_k(k = 1,2,\dots,n)$ be the predicted class probabilities of k^{th} model on the *nts* test data, and *n* represents the number of transformer models. Lines 1 to 4 bring the variable initialization statements. In the inner loop, line 7 calculates the average probable values of each depression class for *n* models of an individual sample. Line 9 of the outer loop finds a class index regarding the maximum average value and saves this index value in the ensemble variable (*mEns*). The paper investigates every combination of the five models for obtaining the best output.

4 Experimental Set-up and Implementation

Since this paper leverages the highly computational state-of-the-art transformer models, a high hardware specification is required. Thus, this research uses the Google Colab cloud environment that supports a Tesla T4 NVIDIA Graphics Processing Unit (GPU) with 15 GB RAM. This Colab also facilitates many Python modules and packages that play significant roles in the implementation phase.

The DSD-1 dataset is partitioned into three parts: training of 70%, validation of 20%, and testing of 10%. The splitting ratio is equivalent to the earlier work [17] on the same dataset. Our research utilizes the Ktrain [38] Python library for implementing transformer models. After rigorous experiments on four performance-influencing parameters like the maximum number of tokens of each sample (*maxLen*), learning rate (*lr*), number of training epochs (*epoch*), and training batch size (*batch*), the best combination is settled, which is shown in Table 3.



```
Algorithm 1: Transformer Ensemble: MaxOfAvgProb
Input: $nClass$: Number of depression classes
       $mPred_1, mPred_2, \dots, mPred_n$: Predicted probability of $nClass$ classes for $n$ models
       $nts$: Test data size
Output: mEns: Predicted class label generated by ensemble approach
1  $mEns \leftarrow []$                    /* Empty list */
2  $avgProb \leftarrow array[0,0]$         /* Initialize an array by zeros with $nClass$ members */
3  $i \leftarrow 0$
4  $j \leftarrow 0$
5  for $i \le nts - 1$ do
6      for $j \le nClass - 1$ do
7          $avgProb[j] \leftarrow mean(mPred_1[i][j], mPred_2[i][j], \dots, mPred_n[i][j])$
           /* Obtain $nClass$ average class probabilities of $n$ models for every test entry */
8          $j \leftarrow j + 1$
9      $mEns \leftarrow append(argmax(avgProb))$
                 /* Compute the maximum of average values with class index and store
                    this index in the ensemble variable */
10 $i \leftarrow i + 1$
```

Fig. 2 Algorithm 1: Transformer Ensemble: MaxOfAvgProb

5 Result and Discussion

This paper takes four assessment metrics: precision, recall, F1-score, and accuracy for evaluating the performance of the transformer models. Since the DSD-1 dataset has an imbalanced nature, the weighted average scores are taken into account. We illustrate this section with two subsections. The first subsection demonstrates the comparative analysis of every classification system. The second one focuses on a detailed analysis of the proposed approach.

Table 3 The best coordination of the hyper-parameters values of each transformer model

Model	***maxLen***	***lr***	***epoch***	***batch***
mBERT	120	4e-05	10	12
DistilmBERT	124	4e-05	10	12
XLM-R-base	124	4e-05	10	12
Bangla-BERT-Base	124	4e-05	10	12
BanglaBERT	132	4e-05	10	12

5.1 Result Analysis

Table 4 exhibits the outputs of each transformer model for the DSD-1 dataset. The BanglaBERT obtains the highest scores in terms of recall (62.10%), precision (66.50%), F1-score (62.61%), and accuracy (62.10%) than the other four individual models. The main reason for the better performance is that this is a Bengali language-specific pre-trained model, and its discriminator part plays a significant role in Natural Language Understanding for this language. The XLM-R-base and mBERT show second and third-top-ranked scores, respectively. The remaining two models, Bangla-BERT-Base and DistilmBERT, present lower detection performance. On the other hand, the proposed ensemble technique MaxOfAvgProb [see Fig. 2] surpasses every individual model with a precision of 66.69%, recall of 62.90%, F1-score of 63.47%, and accuracy of 62.90%

for classifying four levels of depression. This research observes that the MaxOfAvgProb achieves the best results when the top three models, BanglaBERT, XLM-R-base, and mBERT, are taken.

5.2 Detailed Discussion

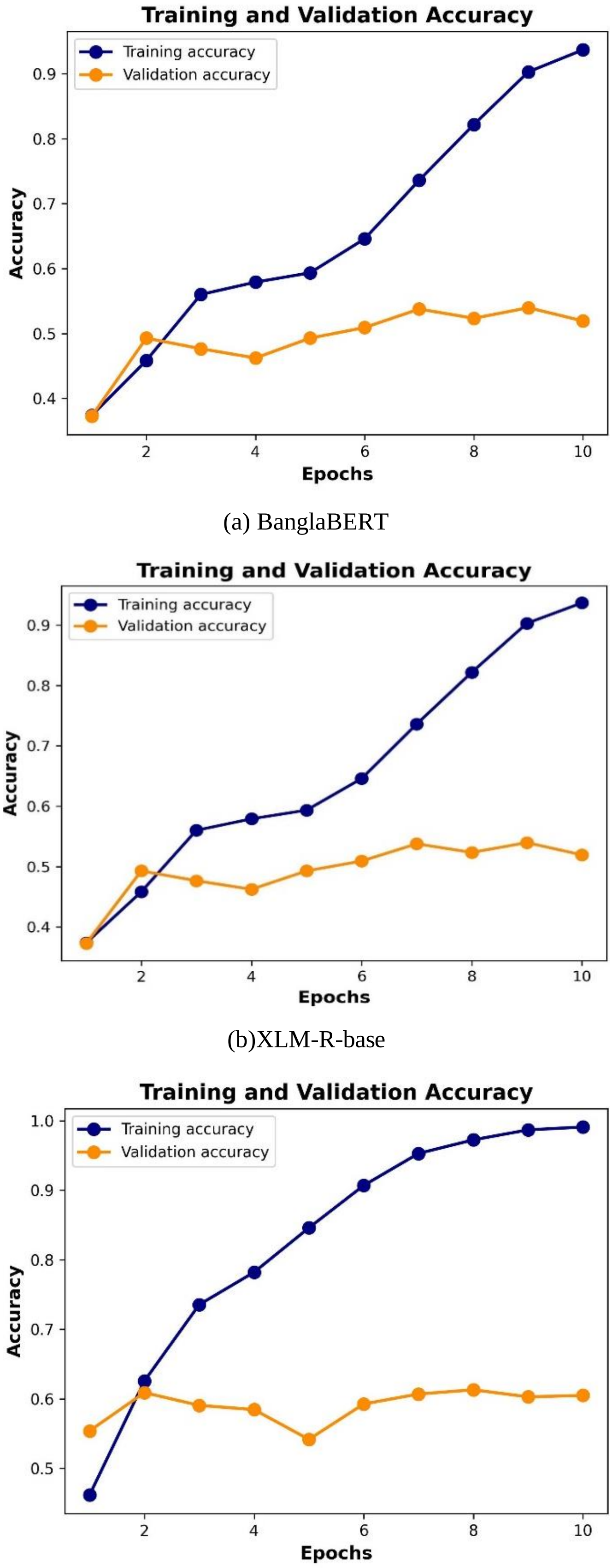
This subsection exhaustively discusses the proposed approach MaxOfAvgProb from various aspects. At first, the training and validation accuracy curves are visually described. Afterward, the paper reveals the performance of every depression class. Next, the model quality is assessed. Subsequently, the system's errors are discussed. Finally, this paper compares the proposed approach with the recent works.

Table 4 Performance comparison of the transformer models for the DSD-1 dataset [**P = Precision, R = Recall, F = F1-score, A = Accuracy**]

Model	P	R	F	A
mBERT	59.45%	56.85%	56.87%	56.85
DistilmBERT	54.82%	54.84%	54.12%	54.84%
XLM-R-base	61.63%	60.89%	61.11%	60.89%
Bangla-BERT-Base	55.06%	54.84%	54.37%	54.84%
BanglaBERT	**66.50%**	**62.10%**	**62.61%**	**62.10%**
MaxOfAvgProb	**66.69%**	**62.90%**	**63.47%**	**62.90%**

Analysis of training-validation accuracy curve: The MaxOfAvgProb comprises BanglaBERT, XLM-R-base, and mBERT models. Therefore, this part visually explains these three models. Fig. 3 delineates the training-validation accuracy curves of BanglaBERT, XLM-R, and mBERT-base for the ten training epochs. The training accuracy of the BanglaBERT model gradually increases up to seven epochs. Then, the accuracy curve slightly rises until at the end, and the ultimate accuracy reaches 0.991. The accuracy curve for the two other models, the XLM-R-base and mBERT, display similar patterns, where the curves moderately increase from the first to the last epoch, and the final accuracy values are shown as 0.895 and 0.937, respectively. On the other hand, the validation accuracy curve for these three models shows the ups and downs pattern with an upward direction. The final accuracy values are counted as 0.605, 0.562, and 0.519 for the BanglaBERT, XLM-R-base, and mBERT, respectively.

Class-wise assessment: Table 5 and the confusion matrix diagram in Fig. 4 delineate the performance of the MaxOfAvgProb approach over the four depression classes. The proposed method shows higher detection scores with precision of 81.93%, recall of 76.40%, and F1-score of 79.07% for the Not Depression class because of the higher percentage (36.56%) of data belonging to this class. The method presents poor results for the Mild (F1-score is 53.80%) and Moderate (F1-score is 52.17%) class data. These two classes are mostly biased by each other, where 23 Moderate test samples are incorrectly labeled as Mild and 20 Mild test samples are misclassified as Moderate class. In the case of the Severe class, the method does not wrongly detect the other classes as the Severe class; consequently, precision shows 100%. However, many Severe test samples are incorrectly classified as the Mild (8 samples) and Moderate (5 samples) class; thus, the recall value is too low (46.15%).



(a) BanglaBERT

(b)XLM-R-base

(c) mBERT

Fig. 3 Training-validation accuracy curves of the transformer models

Table 5 Class-wise performance measurement of the MaxOfAvgProb method for the DSD-1 dataset [**P = Precision, R = Recall, F = F1-score**]

Class	P	R	F
Not Depression	81.93%	76.40%	79.07%
Mild	47.92%	61.33%	53.80%
Moderate	52.63%	51.72%	52.17%
Severe	100.00%	46.15%	63.16%

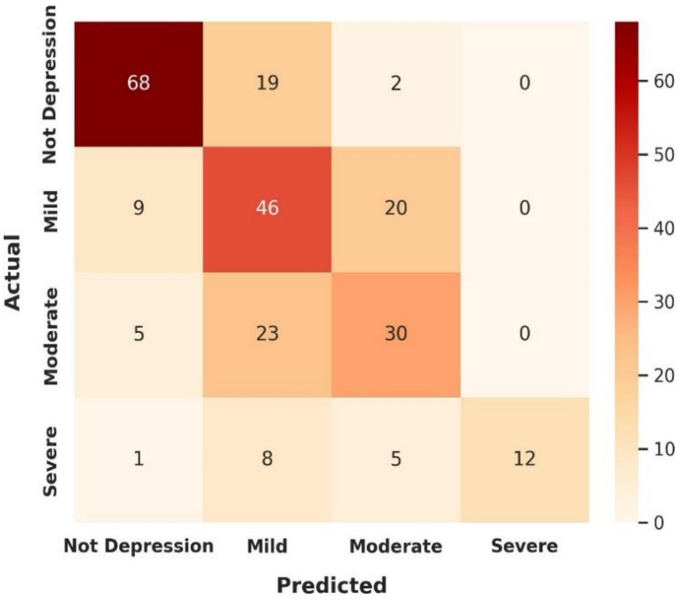


Fig. 4 Confusion matrix of the MaxOfAvgProb method for the DSD-1 dataset

Quality assurance: It is clear from Table 4 that the proposed MaxOfAvgProb-based system exhibits better results than every single transformer model. Additionally, this part discusses two sample examples for increasing the reliability of the proposed system. Let us consider the first example, "আগে যে বিষয়গুলো আমাকে বিরক্ত করেনি তাতে আমি সহজেই বিরক্ত হয়ে যাই" (I get bored easily with things that didn't bother me before), labeled with the Mild class. The mBERT and Bangla-BERT-Base misclassified this text as the Moderate class and the DistilmBERT as the Not Depression class. For the second example, "পালিয়ে যাবো কোথায়? চারিদিকে সেই একই কারাগার" (Where to run away? It's the same prison all around), labeled as Mild class, every model other than the mBERT wrongly detects this sample. However, the proposed approach perfectly recognizes both the samples as the Mild class. Furthermore, we apply our suggested MaxOfAvgProb-based approach to a new depression severity-related dataset proposed by Kabir et al. [16] to justify the quality of the method. During the implementation, the same dataset splitting ratio is retained like [16]. Our proposed approach shows more than 5% better outputs than [16] with precision of 86.30%, recall of 86.45%, F1-score of 86.35%, and accuracy of 86.45%. The class-wise performance and the confusion matrix diagram are articulated in Table 6 and Fig. 5, respectively. Thus, the MaxOfAvgProb-based method ensures a more reliable depression severity detection system for the resourced-constrained Bengali language.

Error analysis: In the case of the DSD-1 dataset, the proposed method is sometimes confused in recognizing the Mild and Moderate class samples because of the overlapping feature attributes. Furthermore, these two classes are influenced by the Severe class text. Since the Severe class has a limited number of instances (10.17%), the method faces difficulty in extracting unique features belonging to this class. On the other hand, the Level 3 class is highly biased by the Level 1 and Level 2 class samples for the DSD-2 dataset because of overlapping tendencies. Additionally, both datasets have grammatical and spelling mistakes that hinder the proposed method for obtaining proper contextual embedding and class-specific knowledge.

Table 6 Class-wise performance measurement of the MaxOfAvgProb method for the DSD-2 dataset [**P = Precision, R = Recall, F = F1-score**]

Class	P	R	F
Level 1	89.30%	92.19%	90.72%
Level 2	88.65%	87.27%	87.96%
Level 3	71.26%	67.23%	69.19%
Level 4	87.59%	86.96%	87.27%

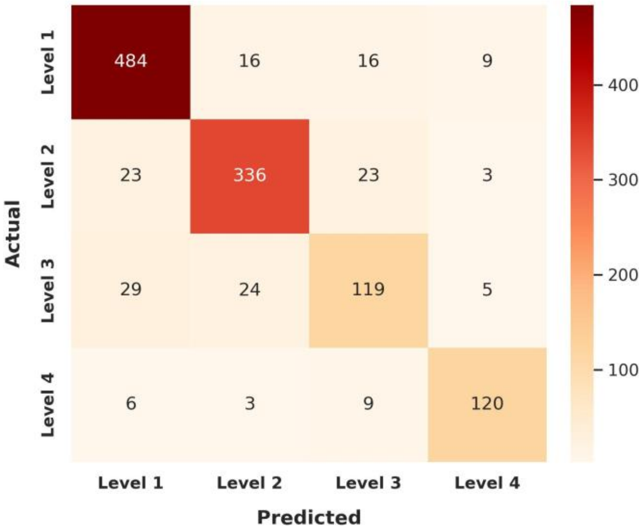


Fig. 5 Confusion matrix of the MaxOfAvgProb method for the DSD-2 dataset

Compare with related works: The proposed method, MaxOfAvgProb, identifies the severity level of depression from the Bengali text data. To the best of our knowledge, we detected three research works [15]–[17] that dealt with recognizing the level of depression in text for the low-resource Bengali language. Table 7 articulates the comparative overview between the proposed method and the recent related works. Hossen et al. [15] achieved better outputs (F1-score is 46.64% and accuracy is 48.00%) using the LR with TF-IDF technique, while Hoque and Salma [17] obtained maximum result with F1-score at 61.11% and accuracy at 60.89% by applying XLM-R-base transformer model for the DSD-1 corpus. However, our proposed approach achieves better results than [17] for the same dataset with 63.47% F1-score and 62.90% accuracy. On the other side, Kabir et al. [16] worked on the DSD-2 dataset and obtained promising performance with 81.00% F1-score and accuracy by applying the BiGRU classifier. However, our suggested method, MaxOfAvgProb, for this dataset obtains over 5% better results (F1-score is 86.35% and accuracy is 86.45%).

Table 7 Performance comparison with the state-of-the-art works [**P = Precision, R = Recall, F = F1-score, A = Accuracy**]

Approach	**Dataset Source**	**Dataset Size**	**Depression Class**	**P**	**R**	**F**	**A**
LR with TF-IDF (Hossen et al. [15])	Facebook	3,000	4 classes: No, Mild, Moderate, and Severe	-	-	46.64%	48.00%
XLM-R-base (Hoque and Salma [17])	Facebook	2,596	4 classes: Not Depression, Mild, Moderate, and Severe	61.63%	60.89%	61.11%	60.89%
BiGRU (Kabir et al. [16])	Social media Groups and blogs	4,897	4 classes: Level 1, Level 2, Level 3, and Level 4	81.00%	81.00%	81.00%	81.00%
MaxOfAvgProb (Proposed)	Facebook	2,596	4 classes: Not Depression, Mild, Moderate, and Severe	**66.69%**	**62.90%**	**63.47%**	**62.90%**
MaxOfAvgProb (Proposed)	Social media Groups and blogs	4,897	4 classes: Level 1, Level 2, Level 3, and Level 4	**86.30%**	**86.45%**	**86.35%**	**86.45%**

6 Conclusion

This paper develops a depression text classification system that effectively recognizes the intensity level of depression in the resource-constrained Bengali language. In this regard, a rigorous experiment is conducted with five cutting-edge transformer models, and the research proposes a new transformer ensemble method, MaxOfAvgProb. The proposed approach surpasses the existing works with an F1-score of 63.47% and an accuracy of 62.90% for the DSD-1 dataset and an F1-score of 86.35% and an accuracy of 86.45% for the DSD-2 dataset. By identifying the intensity of the depressed text, this system may help to detect depressed people, and then an affected individual can get proper counseling or treatment according to the level of depression. Since the classification performance is still behind, we will experiment on a large dataset, including multi-modal data like text, images, and others, with a robust transformer-based hybrid model in future research work.

References

[1] Depressive disorder (depression), World Health Organization, Mar. 2023, accessed: 20 December 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression

[2] Kulsoom, B. and Afsar, N.A., 2015. Stress, anxiety, and depression among medical students in a multiethnic setting. *Neuropsychiatric Disease and Treatment*, pp.1713-1722.

[3] Hossain, M.D., Ahmed, H.U., Chowdhury, W.A., Niessen, L.W. and Alam, D.S., 2014. Mental disorders in Bangladesh: a systematic review. *BMC psychiatry*, *14*, pp.1-8.

[4] Arusha, A.R. and Biswas, R.K., 2020. Prevalence of stress, anxiety and depression due to examination in Bangladeshi youths: A pilot study. *Children and youth services review*, *116*, p.105254.

[5] Choudhury, A.A., Khan, M.R.H., Nahim, N.Z., Tulon, S.R., Islam, S. and Chakrabarty, A., 2019, June. Predicting depression in Bangladeshi undergraduates using machine learning. In *2019 IEEE Region 10 Symposium (TENSYMP)* (pp. 789-794). IEEE.

[6] Hoque, R., 2015. Major mental health problems of undergraduate students in a private university of Dhaka, Bangladesh. *European Psychiatry*, *30*, p.1880.

[7] Hoque, M.N. and Seddiqui, M.H., 2024. Detecting cyberbullying text using the approaches with machine learning models for the low-resource Bengali language. *Int J Artif Intell ISSN*, *2252*(8938), p.358-367.

[8] Tanjim, K.F.H., Hoque, M.N. and Seddiqui, M.H., 2023. A Benchmark Dataset with Developing a Strong Baseline Accident Text Classification System for the Low-resource Bengali Language. In *2023 26th International Conference on Computer and Information Technology (ICCIT)* (pp. 1-6). IEEE.

[9] Uddin, A.H., Bapery, D. and Arif, A.S.M., 2019. Depression analysis from social media data in Bangla language using long short term memory (LSTM) recurrent neural network technique. In *2019 international conference on computer, communication, chemical, materials and electronic engineering (IC4ME2)* (pp. 1-4). IEEE.

[10] Mumu, T.F., Munni, I.J. and Das, A.K., 2021. Depressed people detection from bangla social media status using lstm and cnn approach. *Journal of Engineering Advancements*, *2*(01), pp.41-47.

[11] Mohammed, M.B., Abir, A.S.M., Salsabil, L., Shahriar, M. and Fahmin, A., 2021, December. Depression Analysis from Social Media Data in Bangla Language: An Ensemble Approach. In *2021 Emerging Technology in Computing, Communication and Electronics (ETCCE)* (pp. 1-6). IEEE.

[12] Ghosh, T. and Kaiser, M.S., 2022, February. Bangla depressive social media text detection using hybrid deep learning approach. In *Proceedings of the Third International Conference on Trends in Computational and Cognitive Engineering: TCCE 2021* (pp. 111-120). Singapore: Springer Nature Singapore.

[13] Ahmed, A., Sultana, R., Ullas, M.T.R., Begom, M., Rahi, M.M.I. and Alam, M.A., 2020, December. A machine learning approach to detect depression and anxiety using supervised learning. In *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)* (pp. 1-6). IEEE.

[14] Das, A., Sharif, O., Hoque, M.M. and Sarker, I.H., 2021. Emotion classification in a resource constrained language using transformer-based approach. *arXiv preprint arXiv:2104.08613*.

[15] Hossen, I., Islam, T., Rashed, M.G. and Das, D., 2022, October. Early Suicide Prevention: Depression Level Prediction Using Machine Learning and Deep Learning Techniques for Bangladeshi Facebook Users. In *Proceedings of International Conference on Fourth Industrial Revolution and Beyond 2021* (pp. 735-747). Singapore: Springer Nature Singapore.

[16] Kabir, M.K., Islam, M., Kabir, A.N.B., Haque, A. and Rhaman, M.K., 2022. Detection of Depression Severity Using Bengali Social Media Posts on Mental Health: Study Using Natural Language Processing Techniques. *JMIR Formative Research*, *6*(9), p.e36118.

[17] Hoque, M.N. and Salma, U., 2023. Detecting level of depression from social media posts for the low-resource bengali language. *Journal of Engineering Advancements*, *4*(02), pp.49-56.

[18] Khan, S. and Alqahtani, S., 2023. Hybrid machine learning models to detect signs of depression. *Multimedia Tools and Applications*, pp.1-19.

[19] Mustafa, R.U., Ashraf, N., Ahmed, F.S., Ferzund, J., Shahzad, B. and Gelbukh, A., 2020. A multiclass depression detection in social media based on sentiment analysis. In *17th International Conference on Information Technology–New Generations (ITNG 2020)* (pp. 659-662). Springer International Publishing.

[20] de Jesús Titla-Tlatelpa, J., Ortega-Mendoza, R.M., Montes-y-Gómez, M. and Villaseñor-Pineda, L., 2021. A profile-based sentiment-aware approach for depression detection in social media. *EPJ data science*, *10*(1), p.54.

[21] Chiu, C.Y., Lane, H.Y., Koh, J.L. and Chen, A.L., 2021. Multimodal depression detection on instagram considering time interval of posts. *Journal of Intelligent Information Systems*, *56*(1), pp.25-47.

[22] Abd El-Jawad, M.H., Hodhod, R. and Omar, Y.M., 2018, December. Sentiment analysis of social media networks using machine learning. In *2018 14th international computer engineering conference (ICENCO)* (pp. 174-176). IEEE.

[23] Paul, P. C., Ahmed, M. T., Hasan, M. R., Rajee, A., and Sultana, K., 2023. Analyzing depression on social media utilizing machine learning and deep learning methods. *Indian Journal of Computer Science and Engineering*, *14*(5), pp.740– 746.

[24] Soliman, T.H., Elmasry, M.A., Hedar, A. and Doss, M.M., 2014. Sentiment analysis of Arabic slang comments on facebook. *International Journal of Computers & Technology*, *12*(5), pp.3470-3478.

[25] Seddiqui, M.H., Maruf, A.A.M. and Chy, A.N., 2016. Recursive suffix stripping to augment bangla stemmer. In *International Conference Advanced Information and Communication Technology (ICAICT)*.

[26] Kudo, T. and Richardson, J., 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

[27] Kudo, T., 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.

[28] Sennrich, R., Haddow, B. and Birch, A., 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

[29] Hoque, M.N. and Seddiqui, M.H., 2023, December. Leveraging Transformer Models in the Cyberbullying Text Classification System for the Low-resource Bengali Language. In *2023 26th International Conference on Computer and Information Technology (ICCIT)* (pp. 1-6). IEEE.

[30] Pires, T., Schlinger, E. and Garrette, D., 2019. How multilingual is multilingual BERT?. *arXiv preprint arXiv:1906.01502*.

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

[32] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, *30*.

[33] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. and Stoyanov, V., 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

[34] Sarker, S., 2022. BanglaBERT: Bengali mask language model for Bengali language understanding (2020). *URL: https://github. com/sagorbrur/bangla-bert*.

[35] Bhattacharjee, A., Hasan, T., Ahmad, W.U., Samin, K., Islam, M.S., Iqbal, A., Rahman, M.S. and Shahriyar, R., 2021. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. *arXiv preprint arXiv:2101.00204*.

[36] Clark, K., Luong, M.T., Le, Q.V. and Manning, C.D., 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

[37] Rafi-Ur-Rashid, M., Mahbub, M. and Adnan, M.A., 2022. Breaking the curse of class imbalance: Bangla text classification. *Transactions on Asian and Low-Resource Language Information Processing*, *21*(5), pp.1-21.

[38] Maiya, A.S., 2022. ktrain: A low-code library for augmented machine learning. *Journal of Machine Learning Research*, *23*(158), pp.1-6.