Natural Language Processing Journal 7 (2024) 100075

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.elsevier.com/locate/nlp

Harnessing large language models over transformer models for detecting Bengali depressive social media text: A comprehensive study

Ahmadul Karim Chowdhury [a,*], Saidur Rahman Sujon [a], Md. Shirajus Salekin Shafi [a], Tasin Ahmmad [a], Sifat Ahmed [b], Khan Md Hasib [c], Faisal Muhammad Shah [a]

[a] *Department of Computer Science and Engineering Ahsanullah University of Science and Technology, Bangladesh*
[b] *Dolami Inc., United States of America*
[c] *Department of Computer Science & Engineering Bangladesh University of Business and Technology, Bangladesh*

ARTICLE INFO

Keywords:
Depression
Social media
Bengali
Large language model
Transformer model
Deep learning
Natural language processing

ABSTRACT

In an era where the silent struggle of underdiagnosed depression pervades globally, our research delves into the crucial link between mental health and social media. This work focuses on early detection of depression, particularly in extroverted social media users, using LLMs such as GPT 3.5, GPT 4 and our proposed GPT 3.5 fine-tuned model DepGPT, as well as advanced Deep learning models(LSTM, Bi-LSTM, GRU, BiGRU) and Transformer models(BERT, BanglaBERT, SahajBERT, BanglaBERT-Base). The study categorized Reddit and X datasets into "Depressive" and "Non-Depressive" segments, translated into Bengali by native speakers with expertise in mental health, resulting in the creation of the Bengali Social Media Depressive Dataset (BSMDD). Our work provides full architecture details for each model and a methodical way to assess their performance in Bengali depressive text categorization using zero-shot and few-shot learning techniques. Our work demonstrates the superiority of SahajBERT and Bi-LSTM with FastText embeddings in their respective domains also tackles explainability issues with transformer models and emphasizes the effectiveness of LLMs, especially DepGPT (GPT 3.5 fine-tuned), demonstrating flexibility and competence in a range of learning contexts. According to the experiment results, the proposed model, DepGPT, outperformed not only Alpaca Lora 7B in zero-shot and few-shot scenarios but also every other model, achieving a near-perfect accuracy of 0.9796 and an F1-score of 0.9804, high recall, and exceptional precision. Although competitive, GPT-3.5 Turbo and Alpaca Lora 7B show relatively poorer effectiveness in zero-shot and few-shot situations. The work emphasizes the effectiveness and flexibility of LLMs in a variety of linguistic circumstances, providing insightful information about the complex field of depression detection models.

1. Introduction

Despite being the most common mental disease in the world, depression remains an under-diagnosed disorder with potentially fatal consequences. It is an epidemic of mental illness that affects a large number of people globally. It creates a persistent sense of sorrow, a loss of interest or pleasure in activities, and can disrupt everyday tasks such as sleeping, eating, or working. Food or weight changes, difficulty sleeping, loss of energy, feelings of worthlessness or guilt, difficulty thinking, focusing, or making decisions, and thoughts of death or suicide are all symptoms of depression (MayoClinic2022, 2022; World Health Organization, 2023). Early identification and intervention are critical for effective depression management and therapy. Depression, if left untreated, can cause substantial impairments in personal, social, and professional functioning. According to the World Health Organization (WHO), the frequency of mental disease patients has increased by 13% in the last decade. 2.8 billion of them are affected by depression, which is one of the primary causes of disability and a considerable contributor to the global sickness burden (Cha et al., 2022).

Social media platforms are very popular as a way of self-expression in today's digital era. Individuals feel free to express their emotions and views due to the anonymity given by these platforms (Maloney et al., 2020). Many users often find it useful for dealing with the stresses of daily life. Platforms such as Facebook, X, and Instagram allow users to express sentiments that they would not be able to express in person. It can also have an influence on self-esteem, sleep habits, and the risk of depression, as well as anxiety symptoms. The Institute for Health Metrics and Evaluation (IHME) conducted a study from 2010 to 2019 (GBDResults), which detected depression in some age groups, revealing a significant increase in death or injury cases related to depressive disorder in Bangladesh. Approximately 4.5 million people

* Corresponding author.
E-mail addresses: ahmadulkc@gmail.com (A.K. Chowdhury), sr.sujon.cyb@gmail.com (S.R. Sujon), salekin68@gmail.com (M.S.S. Shafi), tasinahmed411@gmail.com (T. Ahmmad), sifat.austech@outlook.com (S. Ahmed), khanmdhasib.aust@gmail.com (K.M. Hasib), faisal.cse@aust.edu (F.M. Shah).

https://doi.org/10.1016/j.nlp.2024.100075
Received 14 January 2024; Received in revised form 6 April 2024; Accepted 18 April 2024

between the ages of 20 and 54 experience depression, with the rate increasing by 1.9% each year. Moreover, an average of 1.4 million people over the age of 50 were affected by it, with the rate increasing by 5.2% annually.

Researchers have been experimenting with several ways to identify depressive posts on social media using machine learning and text analysis techniques, as these posts can provide insightful information about users' mental health (Biradar and Totad, 2019; Bokolo and Liu, 2023). Massive amounts of textual data can be automatically evaluated using these techniques, which have the potential to produce insightful results (Aggarwal, 2013). Specifically, NLP approaches have been used to build computer models that can identify symptoms of depression in user-generated information, like Facebook posts. These models offer a practical way to enhance current diagnostic practices and offer a cost-effective, flexible, and efficient way to screen for depression across a large population (Gamon et al., 2013).

Our primary objective is to precisely recognize depressive symptoms in social media profiles so that the condition can be identified early on and also determine which of the three types of language models – deep learning, conventional transformer models, and large language models – is best, given these consequences. The goal is to predict the presence of depression in Bangla social media posts with more accuracy, reduced error, and less training time. This kind of analysis has never been done on an enriched and distinct Bangla dataset because the dataset is new. This project uses large language models and advanced deep learning techniques to analyze people's social media posts to identify depression in them. Our study aims to classify individuals into two groups: those with depression and those without it, using standardized depression ratings as its basis. Several state-of-the-art (SOTA) deep learning techniques, including LSTM (Hochreiter and Schmidhuber, 1997), BiLSTM (Graves and Schmidhuber, 2005), GRU (Cho et al., 2014), BiGRU (Chung et al., 2014), conventional transformer models like BERT Multilingual Base Model (Devlin et al., 2018), BanglaBERT (Bhattacharjee et al., 2021), sahajBERT (sahajBERT, 2023) and Bangla BERT Base (Sarker, 2020), and large language models like GPT 3.5, GPT 3.5 turbo fine-tuned, GPT 4, and Alpaca Lora 7B have been used (Ilham and Maharani, 2022; Hasan, 2023).

The following are the prime objectives of our study:

- Precisely recognize depressive symptoms in Bangla social media posts with more accuracy and Identify depression early on
- Determine the best language model among deep learning, conventional transformer models, and large language models for recognizing depressive symptoms
- Reduce error and training time in predicting depression
- Use standardized depression ratings as the basis for classification into two groups: those with depression and those without it
- Utilize state-of-the-art deep learning techniques including LSTM, BiLSTM, GRU, BiGRU, conventional transformer models like BERT Multilingual Base Model, BanglaBERT, sahajBERT, and Bangla BERT Base, as well as large language models like GPT 3.5, GPT 3.5 turbo fine-tuned, GPT 4, and Alpaca Lora 7B

The serious implications of untreated depression, which include mortality, underscore the importance of early care. The emphasis is on extroverted social media users in particular, as it is acknowledged that many people who struggle with depression might not publicly share their despair online. While acknowledging the limitations of using social media data, the study attempts to use the information that is now accessible to identify depressive symptoms to notify the person's immediate social circle of their emotional status. By fine-tuning LLMs with custom data and implementing strategic prompt engineering, we can achieve higher accuracy with less time and expense compared to other deep learning and transformer models in detecting depression from Bangla text.

2. Literature review

In this section, we delve into the existing literature pertinent to our research focus. Our survey encompassed works dedicated to detecting depressive text in Bangla and various other languages. We meticulously examined and categorized these works into three primary classifications: conventional machine learning approaches, deep learning approaches, hybrid models incorporating transformers, and the integration of explainable AI with transformers, alongside extensive language-based approaches. Each of these categories is comprehensively discussed in the following sections.

2.1. Conventional machine learning approaches

Studies using a variety of approaches have made major contributions to the field of study on depression detection from Bengali text in social media.

The study by Bhowmik et al. (2021) achieved 82.21% accuracy using Support Vector Machines (SVM) and BTSC algorithm for Bengali text rating, incorporating speech tagging and special characters. The study suggests larger datasets could improve machine learning accuracy.

Shah et al. (2020b) investigated several classifiers, including Naive Bayes, SVM, K-Nearest Neighbor, and Random Forest, and integrated genetic and linear approaches for feature extraction; Naive Bayes achieved an accuracy of 73.6%. Many factors were used, including N-gram (Uni-gram, Bi-gram, and Tri-gram), TF-IDF, and linguistic features from LI-WC 2015. The study lacked detailed evaluation metrics, numerical results for the proposed method, and information on the specific parameters and configurations used for feature extraction or classifiers.

Hasan et al. (2018) applied a hybrid approach with Naive Bayes and SVM for analyzing political sentiment and achieved good results with uni-gram data. The paper failed to address limitations and challenges in analyzing political views on social media, potentially affecting the findings' generalizability.

Gautam and Yadav (2014) used machine learning classifiers and WordNet for synonym extraction Naive Bayes achieved 88.2%, Maximum Entropy reached 83.8%, and Support Vector Machine achieved 85.5% accuracy, according to the study's excellent classifier performance measures. The authors failed to address potential biases and limitations of machine learning algorithms for sentiment analysis, including labeled datasets and algorithmic bias, and the impact of sarcasm or irony in tweets.

2.2. Deep learning approaches

Deep learning methods are better at recognizing important features and understanding the semantic context of textual input than traditional machine learning methods. Extensive research efforts have utilized state-of-the-art methods, including LSTM-CNN, RNN, hybrid models using transformers, and explainable AI combined with transformer models.

Basri et al. (2021) explored two model variations the first involves an attention layer fed from the output of the pooling layer through an RNN layer before reaching the output layer. In contrast, the second, named Plain CNN, excludes LSTM and attention layers, directly forwarding CNN layer outputs to the output layer. The paper incorporates a convolution layer using the soft-max function as the activation function for convolution operations. The combined CNN and RNN aim to efficiently capture and predict sentiments in Banglish texts. The models exhibit notable performance, particularly in binary classification, where CNN with LSTM and CNN with concatenated Attention and LSTM outperform multi-class classification. While the model performs well on sad and help-labeled data, challenges arise in categories like help or sarcasm due to ambiguous expressions. Research shows surprising success in identifying humor with machine learning models trained on multi-class Bengali datasets. This overcomes previous limitations in detecting humor in text.

Using Bangla social media data, Uddin et al. (2019) compared both LSTM and GRU recurrent neural networks for depression analysis. On this little dataset, GRU models outperformed LSTMs, according to the study. This provides insightful information about the relative efficacy of different models for Bengali depression analysis. The study fails dive how well the conclusions apply to larger or more diverse datasets, why certain hyperparameters were chosen, or how those parameters should be tuned.

A hybrid CNN-LSTM model was employed by Mumu et al. (2021) to detect depression in Bangla social media status updates. The accuracy of the SVM, DT, and KNN classifiers on Facebook data ranged from 60%–80%, whereas the accuracy of the SVM model on a separate dataset was 82.2%. Better results were obtained by the GRU and LSTM models, with accuracy rates of 75.7% and 88.6%, respectively. There may be concerns about the dataset's representativeness and reliability due to the study's lack of specific information about the data collection methods utilized to assess the Bangla status of depression and non-depression.

Shah et al. (2020a) evaluated model performance for early depression detection using metrics like Early Risk Detection Error (ERDE), Latency, and Latency-weighted F1. With an F1 score of 0.81, precision of 0.78, and recall of 0.86, researchers discovered that the "Word2VecEmbed+Meta" feature set performed the best. This demonstrates how well it can detect depression in its early stages. The use of a Reddit dataset, which could not adequately reflect the variety of social media sites and user demographics, limits the study's generalizability. The research lacks a comprehensive analysis of the effectiveness of the several word embedding techniques and metadata components integrated into the model. Greater investigation and comparison of these techniques could provide greater insight into how they impact depression diagnosis accuracy.

2.3. Hybrid and transformer models with explainable AI

Ahmed et al. (2020) introduced an attention-based model for emotion detection in tweets. Two distinct types of embeddings, including Word2Vec, GloVe, FastText, and Sentiment Specific Word Embedding (SSWE), are employed in different components and later concatenated. The proposed model achieves a notable 79% accuracy in emotion detection, evaluated through K-fold cross-validation with 5 folds.

(Haque et al., 2020) investigated the identification of suicidal ideation in social media posts using pre-trained language models such as BERT, ALBERT, ROBERTa, and XLNET. It is emphasized that Transformer models—in particular, ROBERTa—have an edge over traditional deep learning architectures like BiLSTM. The study faced difficulties due to the small amount of data and annotation bias, which may have limited the applicability of the findings. The study acknowledges that further annotated data tagged with criteria developed by mental health professionals are needed to improve the accuracy and reliability of the detection method.

Using Arabic text data, Abdelwahab et al. (2022) used an attention-based LSTM model for sentiment analysis, outperforming other deep learning models in terms of accuracy. The Explainable AI (XAI) approach, more especially LIME, is used in this study to give explanations for the sentiment classifications the LSTM model produced. The study looks at postings with positive, negative, and neutral labels found in the Arabic text data collection to gain an understanding of customers' general opinions of LASIK procedures on social media. The study's exclusive focus on sentiment analysis on Twitter in connection with LASIK operations may limit the findings' generalizability to other domains or social media platforms. The study disregards potential limitations and difficulties, such as dialect variations and complex morphology of the Arabic language, when using XAI algorithms in Arabic text.

2.4. Large language models

The advent of Large Language Models (LLMs) such as GPT-3, Alpaca, and FLAN-T5 has brought about a profound change in the field of mental health diagnosis. Large volumes of online content may be analyzed by these AI-powered technologies, which may open up new possibilities for diagnosing and treating depression. Reviewing the most recent studies on LLMs in this field, this analysis focuses on an extensive study by Xu et al. (2023). Through a thorough comparison of different LLMs, such as Alpaca-LoRA and GPT-4, across a range of tasks, their work shows that optimized models, such as Mental-Alpaca and Mental-FLAN-T5, perform significantly better than their general counterparts. Because the evaluation is based on a small number of datasets and types of LLMs, the conclusions of the research could not hold for different datasets and models.

These findings showcase the immense potential of LLMs for early depression detection, but further research is crucial to address challenges like data bias and ethical considerations before full-fledged integration into clinical practice. The future promises exciting possibilities for AI-powered mental health interventions, but careful navigation is required to ensure responsible and effective implementation.

In zero- and few-shot prompting for Bangla sentiment analysis, Arid Hasan et al. (2023) examined the efficacy of linguistic models (LLMs) such as Flan-T5, GPT-4, and Bloomz. Traditional models, such as SVM and Random Forest, function as reference points for comparison, whereas extensive pre-trained transformer models like Sentiment Analysis fine-tuning are applied to BanglaBERT, mBERT, XLM-RoBERTa, and Bloomz. BanglaBERT in particular, a monolingual transformer-based approach, regularly shows better results in sentiment analysis, surpassing other algorithms in few-shot and zero-shot situations.

Kabir et al. (2023) conducted a detailed analysis of large language models (LLMs) to see how well they performed on various Bengali natural language processing (NLP) tasks. The tasks include sentiment analysis, natural language inference, text categorization, question answering, abstractive summarization, and paraphrasing. LLMs such as Claude-2, Chat-GPT (GPT-3.5), and LLaMA-2-13b-Chat are assessed in a fine-tuning-free zero-shot learning environment. The study carefully crafts prompts for every NLP task and evaluates LLMs against cutting-edge supervised models on benchmark datasets particular to each task. Notably, ChatGPT is very good at abstractive summarization, but Claude-2 is good at answering questions.

Fu et al. (2023) proposed the LLM-Counselors Support System, which supports non-professionals in providing online psychological therapies by utilizing large language models. An iterative method is used by the system to assess and potentially enhance counselor replies using AI models to improve communication with those who are depressed. While avoiding suggestions that were damaging or useless, the AI model concentrated on providing users with personalized guidance and help within the parameters of its assigned function.

BERT, BART-MNLI, GPT-3 Ada, and GPT-3 Davinci are the four large language models (LLMs) based on transformer architectures that Chae and Davidson (2023) have thoroughly analyzed. Relative to conventional machine-learning techniques, the results show that LLMs are more accurate in recognizing stances in social media texts.

Chen et al. (2023) investigated ChatGPT's application for chatbot simulations between psychiatrists and patients. The discoveries demonstrate that chatbots that use prompts with empathy components perform better in engagement measures and higher in empathy metrics. The chatbot with the least amount of speech shows a need for a more thorough diagnosis because of their weaker symptom recall but greater professional skills. Longer patient responses from the chatbot yield the best symptom recall, suggesting effectiveness in a comprehensive investigation of symptoms.

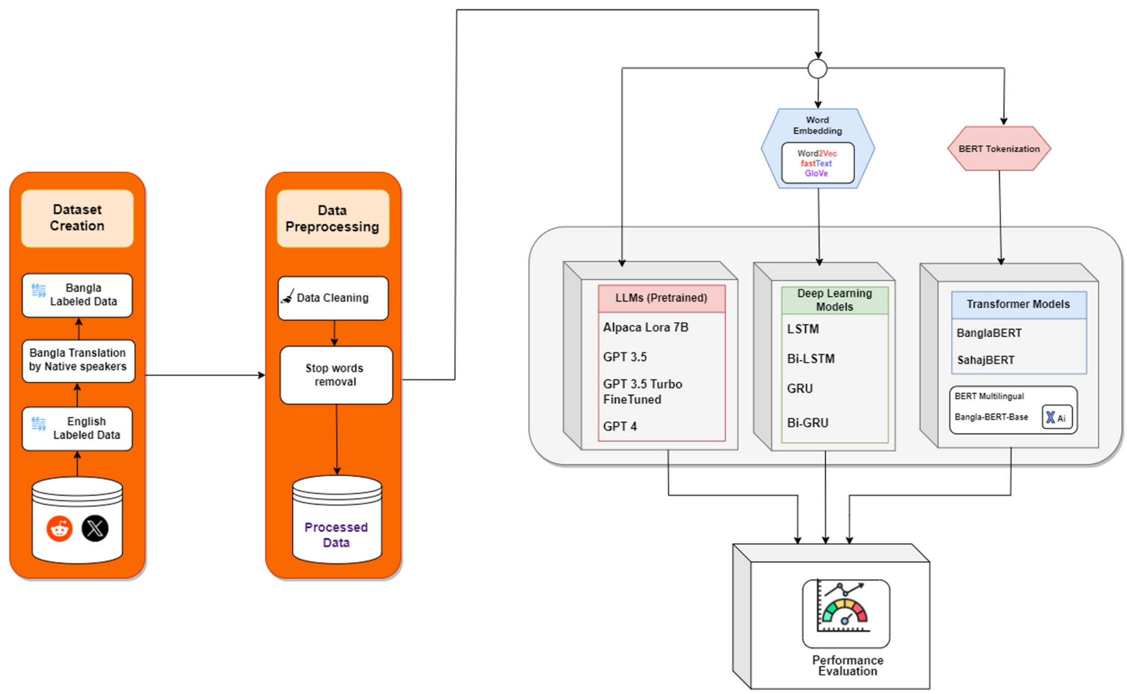

Fig. 1. Proposed methodology.

Table 1
Dataset characterization table.

Language type	Bangla
Platforms	Reddit, Kaggle
Gender biases	Male and female
Total amassed samples	31,695
Total Words	59,23,168
Average no. of words in a sample	187
Total characters	3,38,13,794
Average characters in a sample	1069

Ji et al. (2023) investigated, with theoretical and empirical backing, the unpredictable nature of generative models in mental health prediction. The work cites a case study by Yang et al. (2023), which underlines difficulties in comprehending complicated mental states from self-reported posts and emphasizes the value of interpretable methodologies versus potentially misleading LLM-generated explanations.

3. Problem description and our study

Categorizing the social media published content in Bengali between two categories – depressive and non-depressive – is the root of the problem at hand. The goal is to develop a system that can conclude if a given text x_i taken from a collection of n Bengali social media texts $\mathbf{X} = \{x_1, x_2, x_3, \ldots ..x_n\}$ belongs to one of the two predefined categories: $\mathbf{C} = \{c_1, c_2\}$. The system's main goal is to classify each text as $\boldsymbol{C}_i$, where depressed and non-depressive texts are indicated by the labels c_1 and c_2, respectively.

In this research effort, we created the Bengali Social Media Depressive Dataset (BSMDD), a comprehensive dataset of depressive content sourced from Reddit and X (formerly Twitter). Strenuous efforts were taken to ensure the cleanliness of the dataset by eliminating duplicates and maintaining high translation and annotation quality standards. This dataset is now publicly accessible to the wider research community.

Our approach to experimentation goes beyond the conventional approaches of deep learning and transformer models that are commonly used. Specifically, we delved into the efficacy of sophisticated models, including GPT-3.5 Turbo, our customized model DepGPT (fine-tuned on GPT-3.5), GPT-4, and Alpaca Lora 7B, exploring their performance in both zero-shot and few-shot settings. The results of this experiment contribute valuable insights to understanding the detection of depressing content in social media.

4. Dataset

4.1. Dataset collection

Our experiment is based on two well-established datasets that are commonly employed for mental health analysis collected from social media platforms. A total of 31,695 English textual samples were amassed from reputable sources, namely Reddit (Reddit, 2023) and X (formerly Twitter) (Sentiment140, 2023) verified datasets, which native speakers had meticulously scrutinized (Table 1). Notably, we intentionally avoid using texts that contain less than 30 words as it would be hard to understand emotions in these texts.

Reddit dataset (Reddit, 2023): We utilized a dataset sourced from the Reddit community, specifically focusing on submissions from two relevant subreddits: r/depression and r/SuicideWatch. The dataset comprises a total of 20,000 submissions. We employed this dataset for the depressive texts in our datasets.

Sentiment140 dataset (Sentiment140, 2023): For the non-depressive component of our dataset, we leveraged the Sentiment140 dataset, a comprehensive collection of 1,600,000 tweets extracted using the Twitter API. We have collected approximately 11,000 datasets from this.

4.2. Annotations, translations & validation

The datasets we gathered from Reddit (Reddit, 2023) and X (formerly Twitter) (Sentiment140, 2023) were categorized into depressive and non-depressive segments. To ensure the precision of translations from English to Bengali, seven (7) persons participated in the translation and annotation process. These people are native Bengali speakers and are proficient in both Bengali and English. Among them, three (3) are currently pursuing studies in Bengali, and four (4) are studying computer science.

In addition to their academic qualifications, they have also gained experience in the domain of mental health by participating in several psychological groups on social media. Participating in this group has helped them develop a deeper understanding of the different emotions and experiences associated with mental health. This knowledge has been valuable in their work on the translation and annotation project, as it has helped them to accurately identify and label the emotions expressed in the text.

The final validation assigned to each post was determined through a collaborative approach, with consensus among annotators being the determining factor. In instances of disagreement, a consensus meeting was convened to address disparities and reach a conclusive decision. In this way, we compiled a comprehensive dataset consisting of content related to depression, known as the Bengali Social Media Depressive Dataset (BSMDD).

It is worth noting that participants in the annotation process were remunerated for their contributions, acknowledging the effort and expertise invested in ensuring the quality of the annotated dataset.

5. Proposed methodology

In this section, we have impersonated our proposed model for depression detection that is shown in Fig. 1.

The deep learning and transformer models' algorithm for our proposed model can be inscribed in Algorithm 1. The large language models' algorithm for our proposed model can be inscribed in Algorithm 2.

Algorithm 1: Deep Learning and Transformer model process for depression detection

```
01. LoadRawData();
02. PreprocessData(data);
03. ShuffleData(data);
04. SplitData(data, train_ratio);
05. GenerateWordEmbeddings(data);
06. for epoch in range(num_epochs):
07. ___for batch in get_batches(train_data, batch_size):
08. _______logits = model(features);
09. _______loss = compute_loss(logits, labels,
"binary_crossentropy");
10. _______backward_pass(loss);
11. _______train_accuracy = evaluate(train_data, model);
12. _______test_accuracy = evaluate(test_data, model);
13. _______update_best_accuracy(train_accuracy,
best_train_accuracy);
14. _______update_best_accuracy(test_accuracy, best_test_accuracy);
16. ___end
17. end
18. def evaluate(data, model):
19. _correct = 0;
20. _for batch in get_batches(data, batch_size):
21. ___logits = model(features);
22. ___loss = compute_loss(logits, labels, "binary_crossentropy");
22. ___correct += calculate_correct_predictions(logits, labels);
23. ___accuracy = (correct /total_data(data)) * 100;
24. ___precision, recall, f1 = calculate_metrics(logits, labels);
25. _end
25. return accuracy, precision, recall, f1;
```

Table 2
Dataset summary after preprocessing.

Total processed samples	28,000
Depressive samples (1)	14,000
Non-Depressive samples (0)	14,000
Training samples	22,400
Testing samples	5,600

Table 3
Dataset summary for large language model.

Class	Train	Test	Text Length
Depressive	43.837 K	19.6 K	511+-74
Non Depressive	43.836 K	16.1 K	525+-60
Total	87.675 K	35.7 K	

Algorithm 2: Large language model process for depression detection

```
01. LoadRawData();
02. train_data, test_data = trainTestSplitData();
03. prompts = designPrompt();
04. trainingTokens = convertToJsonFormat(train_data, prompts);
05. loadLLMs() //GPT-3.5-Turbo, GPT-4, Alpaca-LoRA
06. fineTuneModel(model="gpt-3.5-turbo", trainingTokens);
07. def evaluate(data, model):
08. _for model in models:
08. ___predictedData = predict(model, prompts, testData)
09. ___saveResults[model] = evaluateModel(model, testData,
predictedData)
10. comparePerformance(saveResults);
```

5.1. Data preprocessing

The texts obtained after translation from Reddit and Twitter frequently contain elements that introduce noise, such as usernames, hashtags, URLs, English characters, and symbols. To enhance the data quality, we applied several preprocessing filters to the collected texts. As a result of these filters, we were able to remove 3695 texts, leaving 28,000 texts that were then sent to embedding and experimental evaluations. The following steps were taken to process the texts:

- Duplicate texts were removed.
- Repeated punctuations were removed.
- Short texts (less than 30 words) were excluded as it is difficult to grasp emotions from such brief texts.
- English characters and numbers were removed using nltk (NLTK, 2023).
- Bengali text was segmented into tokens and stop words were removed using BNLTK (bnltk, 2023).
- The Bengali texts were reduced to their root forms using bangla-stemmer (bangla-stemmer, 2023).

We divided our dataset of 28,000 texts into two categories: 14,000 "Depressive" and 14,000 "Non-Depressive" for the evaluation of deep learning and large-scale pre-trained transformer models (PLMs) (see Table 2).

For assessing large language models, we further categorized the data into training and test subsets based on tokens. The "Depressive" class had approximately 43,837K training tokens and 19.6K testing tokens, with an average token length of 511 characters. The "Non Depressive" class had around 43,836K training tokens and 16.1K testing tokens, with an average token length of 525 characters. In total, the dataset contained 87.675K tokens, distributed across training and test sets for both classes. This division provides insights into token sample distribution and average lengths for further analysis, as illustrated in Table 3.

5.2. Models

We conducted our experiments using deep learning models, large-scale pre-trained transformer models (PLMs), and large language models (LLMs). For the performance measure for all different experimental settings, we compute accuracy, precision, recall, and F1 score. We split the dataset into 80:20 and got 22,400 texts for training and 5600 texts for training (show in Table 2).

5.2.1. Deep learning models

While deep learning models such as LSTM (Hochreiter and Schmidhuber, 1997), BiLSTM (Graves and Schmidhuber, 2005), GRU (Cho et al., 2014), BiGRU (Chung et al., 2014) have been widely used in prior studies and remain in use of sentiment analysis, we also wanted to assess their performance on our dataset. We trained the models using the training set, fine-tuned the parameters with the development set, and assessed the model's performance on the test set To prepare the data for these models, we transformed the text into a word embedding representation, using Word2vec, GloVe, and fastText. We present the architecture details of all the deep learning models in Table 4.

(a) LSTM: Long Short-Term Memory is a foundational deep learning architecture designed to overcome the vanishing gradient problem in recurrent neural networks. LSTMs maintain a cell state that can selectively store or remove information, allowing them to capture long-range dependencies in sequential data. This makes them well-suited for tasks such as language modeling and sentiment analysis, where understanding the context of words across extended sequences is crucial. LSTMs consist of memory cells and gates that regulate the flow of information, enabling effective learning of temporal patterns. Here are the main equations along with explanations of the variables:

$$i_t = \sigma(W_{ii} \cdot x_t + b_{ii} + W_{hi} \cdot h_{t-1} + b_{hi}) \tag{1}$$

$$f_t = \sigma(W_{if} \cdot x_t + b_{if} + W_{hf} \cdot h_{t-1} + b_{hf}) \tag{2}$$

$$o_t = \sigma(W_{io} \cdot x_t + b_{io} + W_{ho} \cdot h_{t-1} + b_{ho}) \tag{3}$$

In these equations:

x_t is the input at time step t,

f_t represents the forget gate output at time step t,

h_{t-1} is the hidden state from the previous time step,

W and b are the weights and biases,

σ is the sigmoid activation function

(b) BiLSTM: Bidirectional Long Short-Term Memory extends the LSTM architecture by processing input sequences in both forward and backward directions simultaneously. This bidirectional approach enhances the model's capability to understand the context by considering information from past and future time steps. BiLSTMs, like LSTMs, leverage memory cells and gates, allowing them to capture and update information across different time steps bidirectionally. They are particularly effective for text classification tasks where word meanings depend on the surrounding words in both directions, providing a comprehensive understanding of sequential data. Here are the forward and backward equations, along with explanations of the variables.

(c) GRU: Gated Recurrent Unit stands as a robust architecture in the realm of sentiment analysis. GRUs, an enhancement of traditional recurrent neural networks, are adept at capturing dependencies in sequential data. Unlike standard RNNs, GRUs introduce gating mechanisms that enable the model to selectively update its memory, facilitating better long-range dependency handling. In the context of sentiment analysis, GRUs prove valuable for capturing nuanced contextual information and understanding the emotional flow within a given text. The model's ability to selectively retain relevant information ensures efficient sentiment representation. Equations are as follows:

$$z_t = \sigma(W_z \cdot x_t + U_z \cdot h_{t-1} + b_z) \tag{4}$$

$$r_t = \sigma(W_r \cdot x_t + U_r \cdot h_{t-1} + b_r) \tag{5}$$

$$\tilde{h}_t = \tanh(W_h \cdot x_t + U_h \cdot (r_t \odot h_{t-1}) + b_h) \tag{6}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{7}$$

In these equations:

tanh is the hyperbolic tangent activation function.

(d) BiGRU: Building upon the foundation of GRUs, a Bidirectional Gated Recurrent Unit emerges as a formidable architecture for sentiment analysis tasks. BiGRUs process input sequences in both forward and backward directions concurrently, allowing the model to assimilate information from past and future contexts. This bidirectional approach enhances the model's comprehension of sentiment nuances by considering the entire context surrounding each word. The memory cells and gating mechanisms in BiGRUs contribute to a more comprehensive understanding of sentiment, making them particularly effective for discerning subtle emotional cues in text data. Here are the forward and backward equations, along with explanations of the variables:

Table 4
Information on deep learning models' architecture.

Architecture	LSTM	BiLSTM	GRU	BiGRU
Input length	512			
Embedding Dimension	300			
Bidirectional	No	Yes	No	Yes
Dropout	–	0.5	–	0.5
Total Parameters	10,288,202	7,799,998	8,777,751	8,049,198
Activation Function	ReLU			
Activation (Output Layer)	Softmax			
Loss Function	Categorical Crossentropy			
Optimizer	Adam			

5.2.2. Large-scale pre-trained transformer models (PLMs)

Large-scale pre-trained transformer models (PLMs) have achieved state-of-the-art performance across numerous NLP tasks. In our study, we fine-tuned several of these models. These included the BERT Multilingual Base Model (Devlin et al., 2018), the monolingual transformer model BanglaBERT (Bhattacharjee et al., 2021), sahajBERT (sahajBERT, 2023) and Bangla BERT Base (Sarker, 2020). We fine-tuned each model using the default settings over 10 to 50 epochs and 8 to 128 batch sizes.

(a) BERT Base: A deep learning framework called Bidirectional Encoder Representations from Transformers (BERT) links input and output elements and adaptively assigns weights based on their relationship. One of BERT's unique selling points is its capacity for bidirectional training, which allows the language model to understand a word's context by taking into account words that are close to it instead of just concentrating on the word that comes before or after it. BanglaBERT Base (Sarker, 2020) adheres to the same architecture as the original BERT model. The techniques used in BERT are as follows:
Self-Attention Mechanism:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{8}$$

Multi-Head Self-Attention:

Table 5
Information on transformer models' architecture.

Architecture	Used model	Layer	Hyperparameter	Parameter
BERT	BERT Base	12	12	110M
	BanglaBERT Base	12	12	110M
ELECTRA	BanglaBERT	12	12	110M
ALBERT	sahajBERT	24	16	18M

$$\text{MultiHead}(X) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O \tag{9}$$

$$\text{head}_i = \text{Attention}(XW_{Qi}, XW_{Ki}, XW_{Vi}) \tag{10}$$

Layer Normalization and Feedforward Network:

$$\text{Output} = \text{FeedForward}(\text{LayerNorm}(\text{MultiHead}(X))) \tag{11}$$

$$\text{FeedForward}(X) = \text{ReLU}(XW_{FF1} + b_{FF1})W_{FF2} + b_{FF2} \tag{12}$$

Here,

X is the input sequence of embeddings,

Q, K, V are query, key, and value matrices,

d_k is the dimensionality of key vectors,

W and b are the weights and biases,

(b) ELECTRA based: Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA) (Clark et al., 2020) identifies replaced tokens in the input sequence. To do this, a generator model must be trained to predict the original tokens for the masked-out ones, and a discriminator model must be trained to distinguish between original and substituted tokens. ELECTRA discriminator model is BanglaBERT (Bhattacharjee et al., 2021).

(c) ALBERT based: Larger models are not always necessary to achieve improved language models, as shown by ALBERT (A Lite BERT) (Lan et al., 2019). Three major changes are made to the original Transformer's encoder segment architecture in order to do this: factorized embedding parameters, cross-layer parameter sharing, and the use of Sentence-order prediction (SOP) rather than Next Sentence Prediction (NSP). Using masked language modeling (MLM) and Sentence Order Prediction (SOP) aims, sahajBERT (sahajBERT, 2023) is a cooperatively pre-trained ALBERT model in the Bengali language context (see Table 5).

5.2.3. Large language models (LLMs)

Based on the quantity of parameters that each model was trained on, the large language models (LLMs) from Google, Meta, and OpenAI were compared in this study. Particularly when it comes to LLMs, parameters play a crucial role as they are often correlated with the model's comprehension and production of language that is human-like. A model with more parameters is often better able to learn from large volumes of data, leading to more accurate and sophisticated language creation and interpretation.

GPT Models

(a) GPT-3.5: It is OpenAI's large language model (LLM). It is an improved version of GPT-3, which underwent extensive training on a large text and code dataset. The tasks that GPT-3.5 can complete include text creation, translation, summarization, answering questions, and creative writing (Ye, 2023).

(b) GPT-4: It is a multi-modal model, it can process both text and image input, produce text, translate between languages, create various types of creative content, and provide answers to questions. More complicated tasks can be handled by GPT-4 than by earlier GPT models. The model performs at human levels across a wide range of academic and professional benchmarks (OpenAI, 2023).

(c) DepGPT: It is a refined version of GPT-3.5 Turbo fine-tuned model which is designed for specific tasks and trained to follow commands thoughtfully. Fine-tuning is crucial for maximizing performance in big language models, allowing developers to personalize the model with domain-specific data. This tailoring enhances relevance, accuracy, and performance for niche applications, enabling the creation of personalized AI applications (Latif, 2023). Fine-tuning also increases steer-ability and dependability, making the model more predictable in production. Additionally, it significantly improves performance, allowing DepGPT to potentially surpass GPT-4 capabilities in specialized tasks.

GPT-4 and GPT-3.5 were also tested using typical machine learning benchmarks. GPT-4 beats existing big language models as well as the majority of state-of-the-art (SOTA) models, which may incorporate benchmark-specific programming or extra training methods.

LLaMA Models

Stanford University's Alpaca 7B model is a small, efficient instruction-following language model that mimics OpenAI's text-davinci-003 but is more cost-effective for academic research. It was fine-tuned using Text-davinci-003 and performed better than OpenAI's text-based model. However, Alpaca 7B has limitations like generating erroneous information and perpetuating societal preconceptions. Low-Rank Adaptation (LoRA) approaches, such as the Stanford Alpaca cleaned version dataset, can help overcome these limitations without sacrificing performance. (Andermatt and Fankhauser, 2023).

LoRA Low-Rank Adaptation (LoRA) was utilized in the Alpaca 7B model to enhance efficiency in natural language processing, reducing trainable parameters to 16 million while maintaining model performance and reducing processing needs (Hu et al., 2021) (see Table 6).

We used OpenAI's playground to execute our tasks, which has mainly 3 sections: System, User, and Assistant. The 'System' section specifies how the model will respond. It is essentially a collection of exact instructions for specific formatting that is communicated to the model with each prompt. The 'User' section contains the main prompt input into the model, while the 'Assistant' section contains the model's response. Mode, Model, Top P, Frequency penalty, Presence penalty, and Stop sequence are the other controls.

Prompt Design

FTo provide accurate instructional prompting to perform a specific task, the model makes use of system prompts. Regarding our categorization of depression, the system prompt was instructed to function as an expert psychologist with extensive knowledge of recognizing depression in Bangla literature. It will examine every text and evaluate it word for word before categorizing it as either "Non-Depressive" or "Depressive". Direct instructions combined with user prompts are used to categorize the provided Bangla text. The prompt begins with a multiple-choice binary question, and an equal symbol indicates the content that follows. It is best to either completely remove or set to null the input portion prompting from the interface. The assistant provided the classification result in response to the model.

Three models from OpenAI were utilized: GPT-3.5 Turbo, GPT-4, and a fine-tuned version of GPT-3.5 Turbo using our dataset that we have named **DepGPT**. The models were chosen above alternative models due to their cost-effectiveness and efficient tokenization technique, which reduces the number of tokens for Bangla words. As a result, API costs are also minimized. Instead of detecting many tokens in English, the default GPT-3 tokenizer (see Figure 5.4) finds many tokens in Bangla. As a result, the token size grows with each command. The price of the API rises as a result, which, given our limited means, is neither practical nor economical.

Table 6
Information of large language models' architecture.

Features/Models	Parameter size	Structure	Architecture
GPT-3.5	175B	Decoder-only, 12 layers	Autoregressive, masked language modeling
GPT-3.5 Finetuned	Same as GPT-3.5		Fine-tuned for specific tasks
GPT-4	1.7T	120 layers	Multimodal (text and image), likely autoregressive
Alpaca 7B	7B	Encoder–decoder, 6 layers each	Sequence to sequence, masked & factual language modeling

Instruction	You are an expert psychologist who has profound experience on detecting depression from Bangla text. You will be given Bangla text as input and you have to analyze the text very carefully, word by word, and classify it in two classes either: 'Depressive' or 'Non Depressive'.
Task	Is the following Bangla [TEXT] Depressive or Non Depressive?\n-[TEXT] =সুখ আর দুঃখ মিলিয়েই আমাদের এই জীবন। আজ এমন একটি দিন...এই দিনে এই দুনিয়া আলো করে আমার ঘরে এসেছে আমার ছেলে রাফি। শুভ জন্মদিন বাবা। দু-বছরের রাফি আজ হাটতে পারে, দৌড়াতে পারে, দুষ্টুমি করতে পারে। কিন্তু এখনো কথা বলা শিখতে পাড়েনি। (Our life is full of happiness and sadness. Today is such a day...on this day my son Rafi came to my house with the light of this world. Happy birthday dad. Today two years old Rafi can walk, run and play mischief, but has not yet learned to speak)
Assistant	Non Depressive (Output generated by LLM)

Fig. 2. Design of zero shot example.

Table 7
Design of a zero shot sample.

Instruction	System
Task	User prompt
Assistant	Output will be generated by LLM

Table 8
Design of a few shot sample.

Instruction	: {System prompt}
Task-1	: {User prompt [Depressive]}
Assistant	: {Depressive}
Task-2	: {User prompt [Non-Depressive]}
Assistant	: {Non Depressive}
Task-3	: {User prompt}
Assistant	: {Output will be generated by LLM}

Table 9
Control of large language model.

Temperature	1.0
Top P	1.0
Maximum tokens	256
Frequency penalty	0.0
Presence penalty	0.0

Zero-Shot

We used a zero-shot learning technique that operates on classes or data that are not explicitly present in the model's training set. This method allows the model to anticipate or perform tasks associated with novel classes or tasks, demonstrating its capacity to generalize beyond specific training cases. We used the format shown in Table 7 to organize the zero-shot example prompt in our technique. We gave the model an "instruction" and a Task as inputs, and we expected the "Assistant" to produce an output categorizing the text as either "Depressive" or "Non-Depressive". Importantly, the model completed this job without any prior contextual knowledge, demonstrating its capacity to classify objects based only on the task and instruction given, thus capturing the spirit of zero-shot learning shown in Fig. 2.

Few-Shot

When compared to the zero-shot learning scenario, few-shot learning performs better, as demonstrated by the groundbreaking work of Brown et al. (2020). Numerous benchmarking studies (Ahuja et al., 2023) have also confirmed this. In our methodology, we have employed few-shot learning to tackle the challenge of learning from limited labeled data. Unlike traditional machine learning paradigms that often necessitate substantial labeled datasets for training, few-shot learning focuses on training models capable of generalizing and making accurate predictions even when presented with only a small number of examples per class. This approach proves particularly advantageous in scenarios where acquiring extensive labeled data is either cost-prohibitive or unfeasible. The format of our few-shot prompt structure is shown in Table 8.

The example prompt was set up as depicted in Fig. 3 above. It consisted of an "Instruction", Tasks-1 and Task-2, and their corresponding responses from the model via "Assistant". The model received input from both samples, one of which was "Depressive" and the other "Non Depressive". Next, the last task was included after the prompt and forwarded to the model. 'Assistant' is the final solution that categorizes Task 3 using the provided examples and previous knowledge.

Other controls

Temperature, Top P, Maximum Tokens, Frequency penalty, Presence penalty, and stop sequence are respectively used to control the output's randomness, diversity, generative token limits, and the amount of penalization on new tokens based on their existing frequency and appearance in the texts. The configurations used are given in Table 9.

Instruction	You are an expert psychologist who has profound experience on detecting depression from Bangla text. You will be given Bangla text as input and you have to analyze the text very carefully, word by word, and classify it in two classes either: 'Depressive' or 'Non Depressive'.
Task-1	Is the following Bangla [TEXT] Depressive or Non Depressive?\n-[TEXT] = উপরে সুবিশাল আকাশের ছাউনি, নীচে ধূলিময় ধরণীর স্বপ্ন, মাঝখানে ভেলা ভেসে চলেছে মিথ্যে আশায়, মরুভূমি আবর্জনায় ঢেকে আছে হৃদয়ে ভালো নেই, বিদ্রুপের খেলা নেই, নেই আশা, প্রাণে ভালবাসা নেই, তবুও কেন তুমি বেঁচে থাকার ভান করো না? , আমি এখনও জানি না ভবিষ্যত কি, আমি বেঁচে আছি এবং বেঁচে থাকার বিষয়ে আমার সন্দেহ আছে। (Big sky above canopy, dreams of dusty fields below, rafts float amid false hopes, no good in the desert, no mockery, no hope, no love in the soul, yet why don't you pretend to live? I still don't know what the future holds, I'm alive and I doubt my survival)
Assistant	Depressive
Task-2	Is the following Bangla [TEXT] Depressive or Non Depressive?\n-[TEXT] = সুখ আর দুঃখ মিলিয়েই আমাদের এই জীবন। আজ এমন একটি দিন...এই দিনে এই দুনিয়া আলো করে আমার ঘরে এসেছে আমার ছেলে রাফি। শুভ জন্মদিন বাবা। দু-বছরের রাফি আজ হাটতে পারে, দৌড়াতে পারে, দুষ্টুমি করতে পারে। কিন্তু এখনো কথা বলা শিখতে পাড়েনি। (Our life is full of happiness and sadness. Today is such a day...on this day my son Rafi came to my house with the light of this world. happy birthday dad. Two-year-old Rafi can walk, run and play mischief today. But has not yet learned to speak)
Assistant	Non Depressive
Task-3	Is the following Bangla [TEXT] Depressive or Non Depressive?\n-[TEXT] =এটি অনুমোদিত কিনা জানি না তবে কেউ কি জানেন কিভাবে একটি ভাল সুইসাইড নোট তৈরি করতে হয় যেখানে আমি জিনিসগুলিকে সঠিকভাবে ব্যাখ্যা করতে পারি এবং অন্যদের গিল্টট্রিপ না করতে পারি। আমি বুঝতে পারি এটি ট্রিগার হতে পারে এবং কেউ ট্রিগার হলে আমি দুঃখিত। (Don't know if this is allowed but does anyone know how to make a good suicide note where I can explain things properly and not guilt trip others. I realize this can be triggering and I'm sorry if anyone is triggered)
Assistant	Non Depressive (Output generated by LLM)

Fig. 3. Design of few shot example.

6. Result and discussion

In this summary, we present the outcomes of our experiment employing various deep learning models (Section 6.1), transformer models (Section 6.2), and large language models (Section 6.3). While our primary focus revolves around tasks related to detecting depression in this study, we also implemented explainable AI to transformer models in Section 6.2.1.

In essence, our findings reveal promising performance in zero-shot and few-shot scenarios using large language models (LLMs) for depression detection tasks. However, this performance is still constrained. Notably, fine-tuning the models with instructions on the dataset significantly enhances their performance across all tasks concurrently. Our detailed examination highlights the robust reasoning capabilities of specific LLMs, particularly GPT-3.5 and GPT-4. It is crucial to note that these results do not imply deployability. In Section 6, we underscore important ethical considerations and identify limitations in our research.

6.1. Evaluation of deep learning models

GRU, BiGRU, LSTM, and BiLSTM are fundamental deep-learning models used in sentiment analysis due to their ability to handle sequential data and capture long-term dependencies. These models use gating mechanisms to update their hidden state, retaining past information relevant to the current time step. BiGRU and BiLSTM can capture data from both forward and backward time series and handle the vanishing gradient problem well. This study systematically assesses the performance of three-word embeddings Word2Vec, Glove, FastText and four deep learning architectures LSTM, Bi-LSTM, GRU, and Bi-GRU in depression detection tasks.

Deep learning models were used to predict text depressive and non-depressive, with data split into 80:20 sets for training and testing, and experimented with different layer counts(0–3), drop out(0.2-0.5), weight decay(0.001 to 0.01) and batch size (8,16,32,64). The accuracy results are summarized in Table 10.

As depicted in Table 10, the models are evaluated under varying hyperparameters. Key findings include:

Bi-GRU with fastText embeddings achieved the highest accuracy of 90.36%, outperforming other models, including LSTM and Bi-directional LSTM, across Word2Vec, fastText, and GloVe embeddings. The full table of 10 is shown in 13.

6.2. Evaluation of large-scale pre-trained transformer models (PLMs)

In this series of experimental setups, Large-scale pre-trained transformer models (PLMs) like the BERT Multilingual Base Model, monolingual transformer model BanglaBERT, sahajBERT, and BERT-Base-Bangla were assessed, each involving a distinct set of critical hyperparameters. These hyperparameters included batch size (8-128), learning rate (0.01 - 0.0005), number of epochs (10-50), number of folds (5), momentum (0.9), and dropout (0.1), all of which had significant implications for model performance.

The metrics include accuracy, precision, recall, and F1 score. SahajBERT is the model that performs the best according to all criteria; it has an F1 score of 0.8662, an accuracy of 0.867, and a precision of 0.8718. This shows that sahajBERT is better at capturing the subtleties of emotion in Bangla language. The pre-training data or architecture of SahajBERT may be very suitable for this job. Strong outcomes were also demonstrated by BanglaBERT, maybe as a result of its specialized training in Bangla text. The BERT Multilingual Base Model performed better than the Bangla-focused models, perhaps because it was pre-trained. This emphasizes how crucial domain-specific pre-training is to the outcomes of sentiment analysis.

The maximum performance of the large-scale pre-trained transformer models is illustrated in the Table 11. The full table of 11 is shown in 14, 15, 16, 17.

6.2.1. Feature-level explainability of transformer models

This study also investigated the efficacy of Explainable Artificial Intelligence (XAI) techniques in elucidating the decision-making processes of transformer models for sentiment analysis. Given the inherent subtleties of human language, the transformers exhibited limitations in achieving consistently accurate classifications. To address this challenge and gain insights into model reasoning, XAI strategies were implemented. These strategies focused on analyzing word usage patterns and frequencies within the text data (see Figs. 5 and 6).

Despite achieving some concordance between predicted and actual sentiment labels in Fig. 4, the XAI explanations revealed limitations in the model's understanding of nuanced language. For instance, the XAI identified terms such as "restaurant" and "birthday" as contributing to the depressive classification. However, these words are not inherently depressive in isolation, highlighting the model's difficulty in capturing the broader context of human language. Similar limitations were observed in Fig. 7. Words like "happy" and "picture", which typically convey positive sentiment, were also identified by the XAI explanations as contributing to depressive classifications.

Given the limitations identified, we explored the potential of large language models (LLMs) to further analyze depressive text. LLMs, with

Table 10
Performance of the deep learning architectures.

Models & embeddings	Batch size	Epoch	Learning rate	Dropout	Weight decay	Layers	Accuracy (Maximum)	ROC-AUC score
LSTM (Word2Vec)	32	20	0.001	0.2	0.001	3	0.8722	0.851
LSTM (fastText)	32	25	0.01	0.2	0.001	4	0.887	0.879
LSTM (GloVe)	16	10	0.001	0.2	0.001	2	0.88	0.879
Bi LSTM (Word2Vec)	32	20	0.001	0.2	0.005	3	0.885	0.874
Bi LSTM (fastText)	64	20	0.001	0.2	0.001	3	0.8987	0.860
Bi LSTM (GloVe)	16	10	0.001	0.3	0.001	2	0.882	0.880
GRU (Word2Vec)	16	15	0.001	0.3	0.001	2	0.8948	0.881
GRU (fastText)	64	25	0.001	0.4	0.01	2	0.9002	0.882
GRU (GloVe)	16	20	0.001	0.2	0.001	3	0.8777	0.884
Bi GRU (Word2Vec)	16	15	0.001	0.3	0.001	3	**0.8627**	0.853
Bi GRU (fastText)	16	15	0.001	0.3	0.005	4	**0.9036**	0.901
Bi GRU (GloVe)	16	15	0.001	0.2	0.001	2	0.8791	0.853

Table 11
Performance of the large-scale pre-trained transformer models.

Models	Batch size	Learning rate	Num of epoch	Accuracy (Max)	Precision	Recall	F1 Score
BERT Multilingual Base Model (Cased)	32		10	0.8233	0.7935	0.8986	0.8544
Bangla-Bert-Base	128	0.01	40	0.8535	0.8405	0.8967	0.8582
BanglaBERT	64		20	0.8604	0.8692	0.8932	0.8625
sahajBERT	128		13	**0.867**	0.8718	0.8816	0.8662

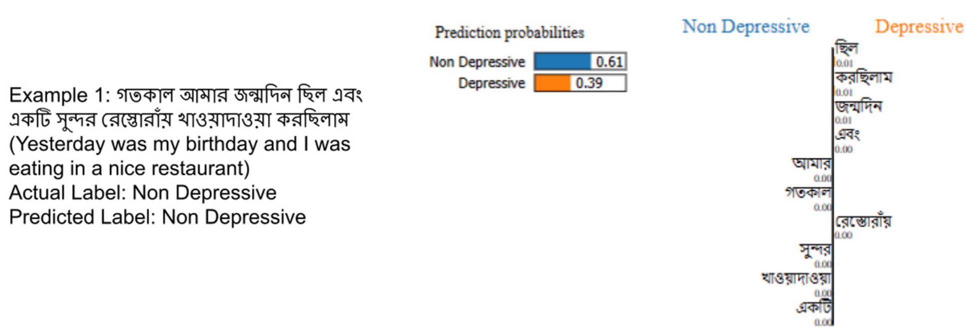

Fig. 4. A true positive sample example.

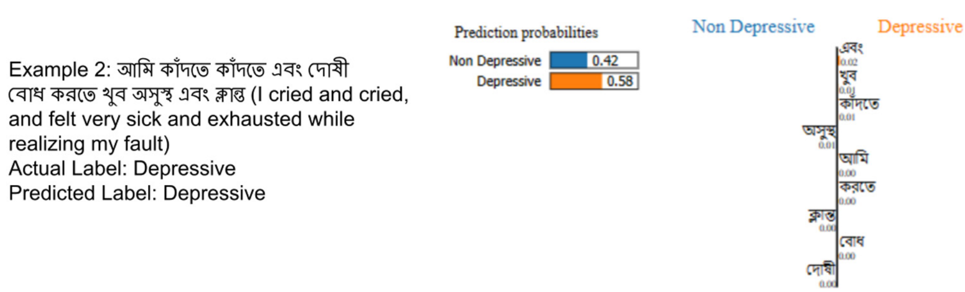

Fig. 5. A true negative sample example.

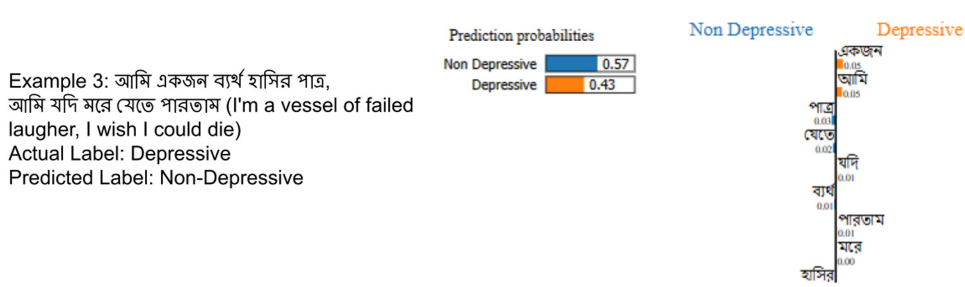

Fig. 6. A false negative sample example.

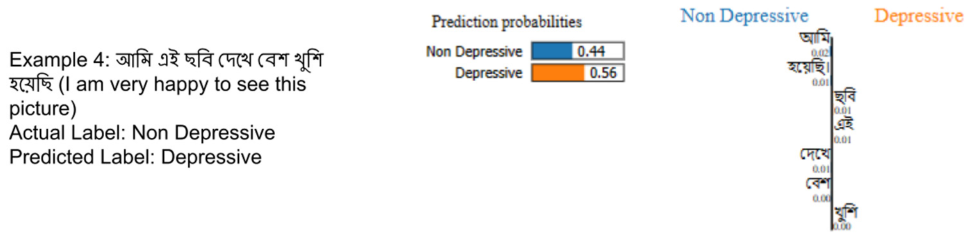

Fig. 7. A false positive sample example.

Table 12
Performance of the large language models.

Category	Models	Performance metrics			
		Accuracy	Precision	Recall	F1-score
Zero-shot	GPT-3.5 Turbo	0.8608	**0.8931**	0.8477	0.8698
	DepGPT	**0.9248**	0.8899	**0.9848**	**0.9349**
	GPT-4	0.8747	0.8452	0.9442	0.8921
	Alpaca Lora 7B	0.7549	0.7121	0.9291	0.8062
Few-shot	GPT-3.5 Turbo	0.8981	0.8846	0.9205	0.9022
	DepGPT	**0.9796**	**0.9615**	**0.9998**	**0.9804**
	GPT-4	0.9388	0.9231	0.9611	0.9412
	Alpaca Lora 7B	0.8571	0.8752	0.8409	0.8571

their superior capabilities in capturing contextual information and semantic relationships, may offer a more robust approach to sentiment analysis, especially in complex domains like depressive language.

6.3. Evaluation of large language models

With excellent recall and accuracy, our suggested model, DepGPT (GPT 3.5 Finetuned), performs better in zero-shot and few-shot learning circumstances than the Alpaca Lora 7B. DepGPT's enormous 175 billion parameters and task-specific fine-tuning based on a decoder-only configuration with 12 layers, which excels in diverse language tasks and offers wide comprehension across domains, are the reasons for its amazing performance. On the other hand, Alpaca Lora is more suitable for some applications that need factual language modeling than DepGPT since it has 7 billion parameters and focuses on sequence-to-sequence tasks. This is because it favors efficiency and targeted performance above versatility in language problem solvers. Our model was able to identify more accurately than Alpaca Lora 7B because of the fine-tuning of DepGPT with our unique dataset. This might also be the cause.

In comparison to zero-shot settings, the GPT-4 model exhibits greater accuracy and worse recall in few-shot scenarios. One explanation might be that, based on the training set provided by OpenAI, the model has a little bias towards the non-depressive setting.

The maximum performance of the Large Language Models is illustrated in the Table 12. The full table of 12 is shown in 18.

DepGPT has an F1-score of 0.9804, near-perfect recall, and exceptional precision in few-shot scenarios with the highest accuracy of 0.9796. Narrowing down its scope by giving multiple test examples with the correct classification helped the model perform excellently well in classifying depressive symptoms. It also performed well in zero-shot scenarios as well. The edge where DepGPT outperformed GPT-4 is identifying ambiguous contexts where a thorough analysis is required with common sense and logical reasoning. See Table R1 for a detailed explanation of the prompting.

Despite prompt and test dataset improvements, GPT-3.5 Turbo and Alpaca Lora 7B perform poorly overall. Lack of fine-tuning and fewer parameters might be the reasons behind the lower accuracy. GPT-4 hallucinates while identifying implicitly indicated depressive texts. On the other hand, DepGPT can identify the hidden depressive meaning of the context with its fine-tuning pattern recognition capability. That is where GPT-4 failed and DepGPT succeeded. We conducted further common sense and logical reasoning experiments on both models. The results are shown below with a detailed analysis in Fig. 8.

From the upper prompt-response analysis in Fig. 8, it is very evident that our fine-tuned model, DepGPT, is more efficient in classifying depressive and non-depressive vibes from ambiguous contexts than GPT-4 by providing detailed common sense reasoning and logical explanation.

6.4. Comparison of all approaches

Our study explored different deep learning, transformer, and large language models for depression detection in text data. Deep learning models, particularly Bi-GRU with fastText embeddings, achieved the best results. This is likely because Bi-GRUs can capture the sequential nature of depression-related sentiment, and fastText embeddings can identify informal language cues in depressive texts. Transformer models, like the BERT Multilingual Base Model, underperformed potentially due to a mismatch between their pre-trained data and the specific language used in depression detection (see Fig. 9).

LLMs have superior accuracy compared to deep learning and transformer models. Transformer models have lower accuracy than deep learning models. The analysis reveals that our proposed model, DepGPT (GPT-3.5 Turbo FineTuned), outperforms all models in adaptability and proficiency, excelling in zero-shot and few-shot tasks for large language models. DepGPT's vast number of parameters (175 billion) allows it to learn complex patterns. Additionally, fine-tuning specifically for depression detection likely enhanced its ability to identify depressive language. Despite being competitive, GPT-3.5 Turbo and Alpaca Lora 7B show poorer efficacy across both learning scenarios.

6.5. Expected vs unexpected behaviors

The Importance of Proper Training and Guidance for LLMs

When employing LLMs like GPT-3.5 Turbo and Alpaca Lora 7B, the expected behavior is to provide predictions about whether the text is 'Depressive' or 'Non-depressive'. However, LLMs started analyzing and performing text completion in GPT-3.5 Turbo and Alpaca Lora 7B, and sometimes they also suggested approaching health professionals for treatment, which was not expected. These models, when functioning correctly, should demonstrate an understanding of the context and offer only appropriate classifications. However, one possible reason for this phenomenon is that whenever there is a missing punctuation '|' (full stop in Bengali) at the end of a statement, their performance might be compromised. This lack of punctuation could confuse the model, leading to unexpected outcomes like performing text completion instead of classification. Furthermore, if the models are not adequately trained or guided to focus on specific tasks, like classification, their responses may lack precision and relevance. (Mohanty, 2024). As DepGPT was fine-tuned on our custom dataset, it did not show any unexpected behavior for any given prompt. The adaptability of our model to our custom dataset might be the prime reason behind the expected outcomes.

Challenges Faced by XAI in Bengali Text Classification

Despite the expectation of accurate depression detection in Bengali text from XAI models, reality often presents unexpected challenges.

Instruction		You are an expert psychologist who has profound experience detecting depression from Bangla text. You will be given Bangla text as input, and you have to analyze the text very carefully, word by word, and classify it into two classes: 'Depressive' or 'Non-Depressive'. Also, explain why you chose this answer with common sense and logical reasoning. Your response should be in the following two formats: [Classification] = Depressive / Non-depressive [Reasoning] = Explanation with common sense and logical reasoning behind selecting the answer.
Prompt-1	True Label: Depressive	Is the following Bangla [TEXT] Depressive or Non-depressive? [TEXT] = আমি খুবই দুঃখিত সে আমাকে প্রশংসিত করেছে সে আমাকে খুশি করেছে সে আমাকে অনুভব করেছে যে এটার মূল্য ছিল আমি জানি না সে এটা করার সিদ্ধান্ত নিয়েছে কিন্তু আমি এটা পরিবর্তন করতে পারি না সবথেকে ভালো জিনিস মনে করিয়ে দেয় যে সে আমার কাছে কি ছিল তার চারপাশে থাকাটা আমাকে ১০গুণ বেশি সুখী করেছে এটা বেশিদিন ছিল না কিন্তু প্রেমে জড়িয়ে থাকার জন্য যথেষ্ট ছিল হয়তো পরের বছর একজনকে খুঁজে পাব কিন্তু এখন কি ভুল হয়েছে তা নিয়ে ভাবছি। **[Classification] =** **[Reasoning] =**
Response-1A	Model: GPT-4	**[Classification] = Non-depressive** **[Reasoning] =** The text does not express hopelessness, lethargy, sadness, or prolonged despair, which are common indicators of depression. Instead, it talks about past good times and hope for future happiness ("সে আমাকে অনুভব করেছে যে এটার মূল্য ছিল", "সে আমার কাছে কি ছিল তার চারপাশে থাকাটা আমাকে ১০গুণ বেশি সুখী করেছে", "হয়তো পরের বছর একজনকে খুঁজে পাব"). Though the speaker is questioning past mistakes, this is not indicative of depression necessarily, as mistakes and their contemplation can be part of normal life.
Response-1B	Model: DepGPT	**[Classification] = Depressive** **[Reasoning] =** Although the text initially mentions positive emotions such as feeling praised and happy, it quickly shifts to expressing sadness and regret. The text mentions feeling very sad, not being able to change something, and questioning what went wrong. The mention of feeling "10 times happier" can be interpreted as an exaggeration to contrast the current feeling of sadness. Overall, the text indicates a depressive tone due to the focus on negative emotions and regret.
Prompt-2	True Label: Non-depressive	Is the following Bangla [TEXT] Depressive or Non-depressive? [TEXT] = হতাশাগ্রস্ত এবং কিছুটা হৃদয়ভঙ্গ তাই আমি এই মেয়েটিকে পছন্দ করি সে সত্যিই স্মার্ট এবং আমি কিছু কারণে এর প্রতি আকৃষ্ট হয়েছি এবং আজ আমরা মাঠে সিটিএফ খেলছিলাম এবং আমার দল শুধু আমাকে ডিফেন্ডার হিসাবে তৈরি করেছিল এবং সে অন্য দিকে ছিল যখন সে দলে এলো তখন সে আমাকে জিজ্ঞেস করলো কেন তুমি সবসময় ডিফেন্ডার হয়ে থাকো আমার থেকে দূরে। আমি মনে করি না সে আমাকে পছন্দ করে হয়তো এমনকি আমাকে ঘৃণা করেও ছেড়ে দিতে পারে। **[Classification] =** **[Reasoning] =**
Response-2A	Model: GPT-4	**[Classification] = Depressive** **[Reasoning]** = The text uses expressions like "হতাশাগ্রস্ত" and "হৃদয়ভঙ্গ", which indicate feelings of despair and heartbreak. The person also perceives that they are unwanted and detested ("আমাকে ঘৃণা করেও ছেড়ে দিতে পারে"), which adds to the depressive tone of the text. Therefore, this text can be classified as depressive.
Response-2B	Model: DepGPT	**[Classification] = Non-depressive** **[Reasoning]** = Although the text mentions a moment of despair and heartbreak, it doesn't entirely present a depressive state. The person is interested in a girl and is playing a game, which indicates engagement in activities. The overall tone seems to be more of a teenager's normal ups and downs, rather than a depressive state. The sentence "হয়তো এমনকি আমাকে ঘৃণা করেও ছেড়ে দিতে পারে" could perhaps be a sign of insecurity or fear of rejection rather than depression. Please note that while this text may not seem depressive on its own, recurring sentiments of this nature may suggest a lower mood state.

Fig. 8. Common sense and logical reasoning behind prediction.

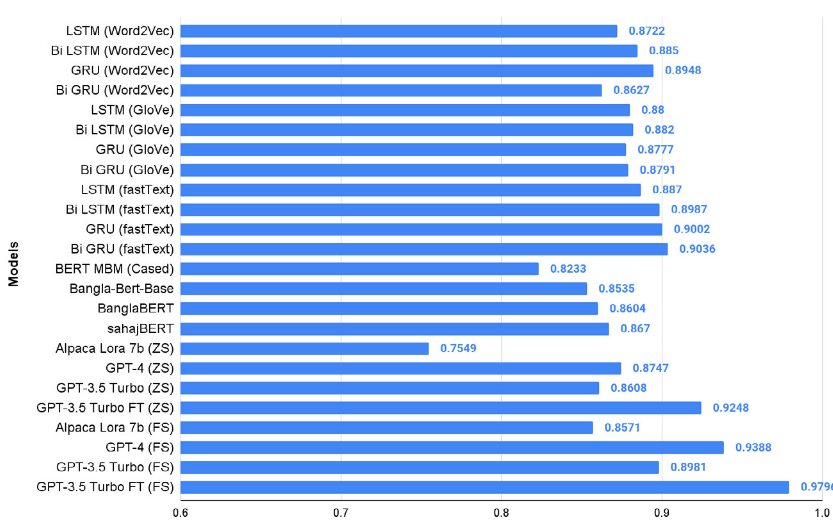

Fig. 9. Comparison of performance of all the models (Here MBM — BERT Multilingual Base (Cased), ZS — Zero shot, FS — Few shot, FT - Fine-tuned).

Table 13
Performance comparison of different sets of experiments in deep learning architectures.

Model Name	Embedding	Batch size	Epoch	Learning rate	Dropout	Weght decay	Layers	Accuracy	Precision	Recall	F1 score
LSTM	Glove	16	10	0.001	0.2	0.001	2	0.88	0.88	0.88	0.88
		16	20	0.001	0.4	0.01	3	0.864	0.866	0.864	0.863
		32	15	0.001	0.2	0.001	3	0.859	0.86	0.859	0.858
		32	25	0.001	0.3	0.005	4	0.8705	0.8706	0.87	8,705
	Word2Vec	16	15	0.001	0.3	0.001	3	0.8568	0.8589	0.8568	0.8566
		32	20	0.001	0.2	0.001	3	0.8722	0.8748	0.8722	0.8719
		32	25	0.001	0.4	0.01	2	0.8713	0.8753	0.8713	0.871
		64	15	0.001	0.2	0.001	3	0.87193	0.87194	0.8719	0.8719
	fastText	16	20	0.01	0.2	0.005	2	0.8775	0.8782	0.8775	0.8774
		16	15	0.001	0.2	0.0001	2	0.8814	0.8834	0.8814	0.88126
		32	25	0.01	0.2	0.001	4	0.887	0.888	0.887	0.887
		64	20	0.001	0.2	0.001	3	0.8828	0.884	0.8828	0.882719
BiLSTM	Glove	16	10	0.001	0.3	0.001	2	0.882	0.88	0.88	0.882
		32	20	0.001	0.2	0.005	3	0.862	0,86	0.862	0.862
		64	20	0.001	0.4	0.001	2	0.866	0.87	0.866	0.865
		64	25	0.001	0.3	0.01	4	0.8814	0.8816	0.8814	0.8813
	Word2Vec	16	15	0.001	0.3	0.001	4	0.8786	0.8793	0.8786	0.8785
		32	20	0.001	0.2	0.005	3	0.885	0.8857	0.885	0.8849
		32	25	0.001	0.3	0.01	2	0.8763	0.8768	0.8764	0.8763
		64	30	0.001	0.2	0.001	3	0.8722	0.8735	0.8722	0.872
	fastText	16	20	0.001	0.2	0.005	2	0.8889	0.8889	0.88895	0.88894
		32	20	0.001	0.2	0.01	2	0.886	0.8909	0.886	0.886
		32	25	0.001	0.4	0.001	4	0.886	0.8903	0.886	0.885
		64	20	0.001	0.2	0.001	3	0.8987	0.89874	0.8987	0.89871
		64	30	0.00001	0.2	0.001	3	0.88895	0.8889	0.88895	0.88894
GRU	Glove	16	15	0.001	0.2	0.001	2	0.8716	0.8729	0.8716	0.8715
		16	20	0.001	0.2	0.001	3	0.8777	0.878	0.8777	0.87777
		32	25	0.001	0.2	0.001	3	0.8738	0.8748	0.8738	0.8738
		64	30	0.001	0.3	0.001	3	0.8725	0.87408	0.8724	0.8723
	Word2Vec	16	15	0.001	0.3	0.001	2	0.8948	0.8959	0.8948	0.8947
		32	20	0.001	0.2	0.005	4	0.8856	0.8862	0.8856	0.8855
		32	25	0.001	0.4	0.01	2	0.8752	0.8753	0.8752	0.8752
		64	30	0.001	0.3	0.001	3	0.8563	0.8594	0.8563	0.8559
	fastText	16	15	0.001	0.2	0.05	4	0.8984	0.8987	0.8984	0.8984
		32	20	0.001	0.3	0.001	2	0.8891	0.8891	0.8891	0.8891
		64	25	0.001	0.4	0.01	2	0.9002	0.9024	0.9002	0.9001
		64	30	0.001	0.3	0.001	3	0.8882	0.8901	0.8882	0.888
BiGRU	Glove	16	15	0.001	0.2	0.001	2	0.8791	0.8834	0.8792	0.8788
		32	20	0.001	0.3	0.001	3	0.8513	0.8517	0.8512	0.8512
		32	25	0.001	0.3	0.001	2	0.8769	0.8784	0.8769	0.8768
		64	30	0.001	0.2	0.001	2	0.847	0.8479	0.84709	0.847
	Word2Vec	16	15	0.001	0.3	0.001	3	0.8627	0.8636	0.8627	0.8626
		32	20	0.001	0.2	0.001	3	0.8574	0.8595	0.8574	0.8572
		32	25	0.001	0.4	0.01	2	0.8574	0.8575	0.8574	0.8574
		64	30	0.001	0.2	0.001	3	0.8513	0.8554	0.8512	0.8508
	fastText	16	15	0.001	0.3	0.005	4	0.9036	0.9037	0.9036	0.9036
		32	20	0.001	0.2	0.01	3	0.8991	0.8991	0.8991	0.8991
		32	25	0.001	0.4	0.001	2	0.887	0.8952	0.887	0.8865
		64	30	0.001	0.2	0.001	2	0.8965	0.8967	0.8965	0.8965

XAI's poor performance can be attributed to the complexity and richness of the Bengali language. The scarcity of Bengali training data limits the effectiveness of XAI models, hindering their ability to understand and explain the nuances of Bengali text. Moreover, the intricate structure of Bengali words and the variations in dialects pose further difficulties for accurate classification. These unexpected outcomes highlight the need for more comprehensive training data and tailored approaches for Bengali text analysis. (Aporna et al., 2022).

Comparative Analysis of Multilingual Models

In the comparative analysis between the BERT Multilingual Base Model (Cased) and specialized Bengali models like Bangla-BERT and sahajBERT, the possible reason behind the variation in performance becomes evident. While BERT Multilingual is trained in multiple languages, it may not capture the subtleties of any particular language as effectively as specialized models. Bangla-BERT and sahajBERT, on the other hand, benefit from pretraining that allows them to capture the nuances and intricacies of the Bengali language more comprehensively. This underscores the importance of tailored training approaches for achieving optimal performance in language-specific tasks. (Bhattacharjee et al., 2021).

Transformers vs. Deep Learning: Understanding Model Performance

The expectation that transformers should outperform deep learning models is challenged by the observed results [Reference to Transformers result table], where transformers performed worse. This unexpected outcome can be attributed to the different strengths of each model architecture. While deep learning models excel in capturing long-term dependencies, transformers are better suited for parallel processing tasks. The discrepancy in performance underscores the importance of understanding the inherent capabilities and limitations of each model type when selecting the most suitable approach for a given task. (Wang et al., 2023).

Table 14
Performance of BERT Multilingual Base (Cased).

Batch size	Learning rate	Num of epoch	Num of folds	Momentum	Dropout	Accuracy	Precision	Recall	F1 Score
8	0.01	10	5	0.9	0.1	0.799	0.755	0.869	0.8079
		20				0.8111	0.7897	0.8481	0.8178
		30				0.8136	0.791	0.8526	0.8206
		40				0.8216	0.7794	0.8971	0.8341
		50				0.8163	0.7935	0.8476	0.8196
16	0.01	10				0.8153	0.7738	0.8605	0.8149
		20				0.8088	0.7618	0.8986	0.8245
		30				0.8198	0.7792	0.8924	0.832
	0.001	20				0.7886	0.7507	0.8642	0.8544
32	0.1	50				0.8149	0.7923	0.8534	0.8217
	0.09	20				0.8031	0.7791	0.8462	0.8113
	0.07	50				0.8181	0.7813	0.8834	0.8292
	0.05	50				0.8199	0.7792	0.8928	0.8321
	0.03	50				0.8233	0.7821	0.8964	0.8354
	0.01	10				0.7954	0.7652	0.8523	0.8064
		20				0.8155	0.7722	0.8952	0.8291
		30				0.8109	0.7729	0.8806	0.8233
		40				0.8201	0.7808	0.8901	0.8319
		50				0.8206	0.7799	0.8934	0.8328
	0.007	50				0.8174	0.7796	0.8851	0.829
	0.005	50				0.788	0.7555	0.8518	0.8007
		50				0.7897	0.7572	0.8528	0.8022
		50				0.8186	0.7814	0.8849	0.8299
	0.003	50				0.8113	0.77	0.8877	0.8247
	0.001	50				0.7934	0.7575	0.8631	0.8068
	0.0005	50				0.7775	0.745	0.8437	0.7913
64	0.01	10				0.7996	0.7576	0.8812	0.8147
		20				0.8049	0.7697	0.8702	0.8035
		30				0.802	0.763	0.8761	0.8156
		40				0.8201	0.7858	0.8802	0.8303
		50				0.8189	0.7785	0.8452	0.8104

Table 15
Performance of BanglaBERT.

Batch size	Learning rate	Num of epoch	Num of folds	Momentum	Dropout	Accuracy	Precision	Recall	F1 Score
8	0.01	10	5	0.9	0.1	0.8443	**0.8692**	0.8106	0.8389
		20				0.8456	0.8607	0.8246	0.8423
		30				0.8338	0.8382	0.8274	0.8328
		40				0.8362	0.8304	0.845	0.8376
		50				0.8411	0.8397	0.8433	0.8415
16		10				0.8474	0.8227	0.8858	0.853
		20				0.8381	0.846	0.8266	0.8362
		30				0.8538	0.8342	0.8831	0.8579
		50				0.853	0.8673	0.8335	0.8501
32		10				0.8381	0.8094	0.8844	0.8453
		20				0.8494	0.8373	0.8673	0.8521
		30				0.8567	0.8325	**0.8932**	0.8618
		40				0.8551	0.8426	0.8734	0.8577
		50				0.8599	0.8435	0.8714	0.8572
64		10				0.8515	0.8367	0.8735	0.8547
		20				**0.8604**	0.8575	0.8646	0.861
		30				0.8593	0.8477	0.8761	0.8617
		40				0.8554	0.8618	0.8465	0.8541
		50				0.8546	0.85	0.8613	0.8556
128		10				0.8458	0.8276	0.8736	0.85
		20				0.8504	0.842	0.8626	0.8522
		30				0.86	0.8472	0.8784	**0.8625**
		40				0.8574	0.8493	0.8688	0.859
		50				0.858	0.8514	0.8673	0.8593

Table 16
Performance of sahajBERT.

Batch size	Learning rate	Num of epoch	Num of folds	Momentum	Dropout	Accuracy	Precision	Recall	F1 Score
8	0.01	10	5	0.9	0.1	0.7171	0.7175	0.7162	0.7169
		15				0.7357	0.7337	0.74	0.7368
		17				0.7205	0.7078	0.751	0.7288
		20				0.7335	0.7332	0.7362	0.7342
16		7				0.7229	0.7172	0.7359	0.7264
		10				0.7411	0.7471	0.729	0.7379
		13				0.7508	0.7401	0.7729	0.7562
		17				0.756	0.7619	0.7448	0.7532
32		7				0.7451	0.7417	0.752	0.7468
		10				0.7405	0.7567	0.7089	0.732
		13				0.7548	0.7524	0.7594	0.7559
		17				0.7422	0.7419	0.7429	0.7424
64		7				0.7342	0.7345	0.7337	0.7341
		10				0.7285	0.7296	0.7261	0.7278
		13				0.7249	0.7286	0.7167	0.7226
		17				0.7539	0.7555	0.7509	0.7532
128		7				0.8506	0.8405	0.8655	0.8528
		10				0.863	0.85	0.8816	0.8655
		13				0.867	0.8718	0.8606	0.8662
		17				0.8618	0.8595	0.865	0.8622

Table 17
Performance of Bangla-BERT-Base.

Batch size	Learning rate	Num of epoch	Num of folds	Momentum	Dropout	Accuracy	Precision	Recall	F1 Score
8	0.01	10	5	0.9	0.1	0.7933	0.7492	0.8819	0.8101
		20				0.8177	0.7914	0.8626	0.8255
		30				0.8326	0.8254	0.8436	0.8344
		40				0.8063	0.816	0.791	0.8033
		50				0.8051	0.7896	0.8319	0.8102
16		10				0.842	0.8083	0.8967	0.8502
		20				0.8372	0.8227	0.8596	0.8407
		30				0.7994	0.7824	0.8295	0.8053
		40				0.8046	0.8061	0.8022	0.8042
		50				0.8011	0.788	0.824	0.8056
32		10				0.8449	0.8405	0.8514	0.8459
		20				0.8437	0.8212	0.8787	0.849
		30				0.8042	0.7886	0.8313	0.8094
		40				0.825	0.7951	0.8757	0.8334
		50				0.832	0.8093	0.8687	0.838
64		10				0.837	0.803	0.8931	0.8456
		20				0.8493	0.8362	0.8689	0.8522
		30				0.8356	0.8221	0.8565	0.8389
		40				0.832	0.8217	0.8479	0.8346
		50				0.824	0.815	0.8383	0.8265
128		10				0.8462	0.8373	0.8595	0.8482
		20				0.8417	0.8173	0.8803	0.8476
		30				0.8469	0.8243	0.8816	0.852
		40				0.8535	0.8315	0.8866	0.8582
		50				0.8503	0.8219	0.8943	0.8566

6.6. Impact on scientific community

DepGPT, a variant of the Generative Pre-trained Transformer (GPT) model, has a significant impact on the scientific community in detecting depression from social media posts by leveraging advanced natural language processing techniques. It is designed to analyze text data related to mental health conditions like depression, offers a powerful tool for identifying depressive symptoms and patterns in social media content, contributing to the early detection and intervention of mental health issues within online communities (Xu et al., 2023; Salas-Zárate et al., 2022).

DepGPT outperforms existing state-of-the-art models in tasks where fine-tuning for specific applications is crucial for maximizing performance and relevance, allowing developers to personalize the model with domain-specific data. This tailoring enhances the accuracy, performance, and relevance of DepGPT for niche applications, enabling the creation of personalized AI applications and potentially surpassing the capabilities of more generalized models like GPT-4 in specialized tasks (Latif and Zhai, 2024). The refinement and fine-tuning of DepGPT make it particularly effective in scenarios where model customization and adaptation to specific domains are essential for achieving superior performance compared to existing state-of-the-art models. By utilizing the extensive capability of LLMs over existing transformers and deep learning models, DepGPT provides a more accurate, efficient, and fast response with the minimum training time required to predict depressive symptoms in the native Bengali language.

7. Conclusion

Deep learning models, transformer models, and large language models all have limits when it comes to sentiment analysis. These include data bias, context understanding, out-of-domain performance, language and dialect issues, long-term dependencies, resource reliance, interpretability, adaptability, generalization, dynamic language, emotional nuance, ethical considerations, and reliance on annotations. Transformer models understand context better, but they struggle with verbal

Table 18
Performance comparison of different sets of experiments in LLM (where D — Depressive, ND — Non Depressive).

Category	Models	Performance metrics				Class
		Accuracy	Precision	Recall	F1-score	
Zero-shot Example	GPT-3.5 Turbo	0.8608	0.8931	0.8477	0.8698	D
			0.8256	0.8765	0.8503	ND
	DepGPT	0.9248	0.8899	0.9848	0.9349	D
			0.9787	0.8518	0.9109	ND
	GPT-4	0.8747	0.8452	0.9442	0.8921	D
			0.9207	0.7901	0.8505	ND
	Alpaca Lora 7B	0.7549	0.7121	0.9291	0.8062	D
			0.8628	0.5432	0.6667	ND
Few-shot Example	GPT-3.5 Turbo	0.8981	0.8846	0.9205	0.9022	D
			0.9131	0.8751	0.8936	ND
	DepGPT	0.9796	0.9615	0.9998	0.9804	D
			0.9981	0.9583	0.9787	ND
	GPT-4	0.9388	0.9231	0.9611	0.9412	D
			0.9565	0.9167	0.9362	ND
	Alpaca Lora 7B	0.8571	0.8752	0.8409	0.8571	D
			0.8401	0.8751	0.8571	ND

nuances like sarcasm, irony, and inferred meanings. Large language models are computationally costly to train and fine-tune, limiting the capabilities of smaller groups of researchers. The dynamic nature of language evolution presents a risk since new idioms and slang can cause models to become out of date very fast. Complexity is increased when emotional nuances are expressed on a continuum as opposed to in binary terms. Two more important considerations in the assessment of these large-scale models are the ethical implications and the dependence on human annotations.

Our research focuses on efficiently identifying early depression from a person's social media posts, especially those written in Bangla. Day by day, depression is spreading worldwide and silently taking precious lives without any prior indication. Our model, DepGPT, outperformed all other models with excellent efficiency and can help social media algorithms identify users' life situations so that necessary measures can be taken to provide mental support to the individual and save precious lives. Moreover, our model can also help a therapist identify his/her patient's mental state by providing their life story to the model. In addition, this model can also be used in automated chatbots to detect users' mental states during conversations. Combining advanced sentiment analysis skills with online counseling platforms could revolutionize mental health support.

CRediT authorship contribution statement

Ahmadul Karim Chowdhury: Writing – review & editing, Writing – original draft, Software, Resources, Methodology, Data curation, Conceptualization. **Saidur Rahman Sujon:** Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Methodology, Data curation. **Md. Shirajus Salekin Shafi:** Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Methodology, Data curation. **Tasin Ahmmad:** Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Methodology, Data curation, Conceptualization. **Sifat Ahmed:** Software, Resources, Formal analysis. **Khan Md Hasib:** Validation, Supervision, Project administration, Investigation. **Faisal Muhammad Shah:** Validation, Supervision, Project administration, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Result tables of deep learning models

See Table 13.

Appendix B. Result tables of large-scale pre-trained transformer models (PLMs)

See Tables 14–17.

Appendix C. Result tables of large language models

See Table 18.

Appendix D. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.nlp.2024.100075.

References

Abdelwahab, Youmna, Kholief, Mohamed, Sedky, Ahmed Ahmed Hesham, 2022. Justifying Arabic text sentiment analysis using explainable AI (XAI): LASIK surgeries case study. Information 13 (11), 536.

Aggarwal, Charu C., 2013. Neural networks and deep learning. URL https://link.springer.com/book/10.1007/978-3-319-94463-0. Accessed: 2023-12-18.

Ahmed, Sifat, Reyadh, Abdus Sayef, Sithil, Fatima Tabsun, Shah, Faisal Muhammad, Shaafi, Asif Imtiaz, 2020. An attention-based approach to detect emotion from tweets. In: 2020 3rd International Conference on Computer and Informatics Engineering. IC2IE, IEEE, pp. 182–187.

Ahuja, Kabir, Hada, Rishav, Ochieng, Millicent, Jain, Prachi, Diddee, Harshita, Maina, Samuel, Ganu, Tanuja, Segal, Sameer, Axmed, Maxamed, Bali, Kalika, Sitaram, Sunayana, 2023. MEGA: Multilingual evaluation of generative AI. arXiv.org, URL https://arxiv.org/abs/2303.12528.

Andermatt, Pascal Severin, Fankhauser, Tobias, 2023. UZH_Pandas at SimpleText@CLEF-2023: Alpaca LoRA 7B and LENS model selection for scientific literature simplification. arXiv.org.

Aporna, Amena Akter, Azad, Istinub, Amlan, Nibraj Safwan, Mehedi, Md Humaion Kabir, Mahbub, Mohammed Julfikar Ali, Rasel, Annajiat Alim, 2022. Classifying offensive speech of bangla text and analysis using explainable ai. In: International Conference on Advances in Computing and Data Sciences. Springer, pp. 133–144.

Arid Hasan, Md, Das, Shudipta, Anjum, Afiyat, Alam, Firoj, Anjum, Anika, Sarker, Avijit, Rashed Haider Noori, Sheak, 2023. Zero-and few-shot prompting with LLMs: A comparative study with fine-tuned models for bangla sentiment analysis. arXiv e-prints. arXiv–2308.

bangla-stemmer, 2023. Bangla-stemmer. URL https://pypi.org/project/bangla-stemmer/. Accessed: 2023-12-18.

Basri, Rabeya, Mridha, MF, Hamid, Md Abdul, Monowar, Muhammad Mostafa, 2021. A deep learning based sentiment analysis on Bang-Lish disclosure. In: 2021 National Computing Colleges Conference. NCCC, IEEE, pp. 1–6.

Bhattacharjee, Abhik, Hasan, Tahmid, Ahmad, Wasi Uddin, Samin, Kazi, Islam, Md Saiful, Iqbal, Anindya, Rahman, M Sohel, Shahriyar, Rifat, 2021. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. arXiv preprint arXiv:2101.00204.

Bhowmik, Nitish Ranjan, Arifuzzaman, Mohammad, Mondal, M Rubaiyat Hossain, Islam, MS, 2021. Bangla text sentiment analysis using supervised machine learning with extended lexicon dictionary. Natural Lang. Process. Res. 1 (3–4), 34–45.

Biradar, Abhilash, Totad, S.G., 2019. Detecting depression in social media posts using machine learning. In: Santosh, K.C., Hegadi, Ravindra S. (Eds.), Recent Trends in Image Processing and Pattern Recognition. Springer Singapore, Singapore, ISBN: 978-981-13-9187-3, pp. 716–725.

bnltk, 2023. Bnltk. URL https://pypi.org/project/bnltk/. Accessed: 2023-12-18.

Bokolo, Biodoumoye George, Liu, Qingzhong, 2023. Deep learning-based depression detection from social media: Comparative evaluation of ML and transformer techniques. Electronics (ISSN: 2079-9292) 12 (21), http://dx.doi.org/10.3390/electronics12214396, URL https://www.mdpi.com/2079-9292/12/21/4396. Accessed: 2023-12-18.

Brown, Tom, Mann, Benjamin, Ryder, Nick, Subbiah, Melanie, Kaplan, Jared D, Dhariwal, Prafulla, Neelakantan, Arvind, Shyam, Pranav, Sastry, Girish, Askell, Amanda, et al., 2020. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901, Accessed: 2023-12-18.

Cha, Junyeop, Kim, Seoyun, Park, Eunil, 2022. A lexicon-based approach to examine depression detection in social media: the case of Twitter and university community. Humanit. Soc. Sci. Commun. 9, http://dx.doi.org/10.1057/s41599-022-01313-2.

Chae, Youngjin, Davidson, Thomas, 2023. Large language models for text classification: From zero-shot learning to fine-tuning. Open Sci. Found..

Chen, Siyuan, Wu, Mengyue, Zhu, Kenny Q, Lan, Kunyao, Zhang, Zhiling, Cui, Lyuchun, 2023. LLM-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. arXiv preprint arXiv:2305.13614.

Cho, Kyunghyun, Van Merriënboer, Bart, Gulcehre, Caglar, Bahdanau, Dzmitry, Bougares, Fethi, Schwenk, Holger, Bengio, Yoshua, 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Chung, Junyoung, Gulcehre, Caglar, Cho, KyungHyun, Bengio, Yoshua, 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Clark, Kevin, Luong, Minh-Thang, Le, Quoc V, Manning, Christopher D, 2020. Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, Toutanova, Kristina, 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Fu, Guanghui, Zhao, Qing, Li, Jianqiang, Luo, Dan, Song, Changwei, Zhai, Wei, Liu, Shuo, Wang, Fan, Wang, Yan, Cheng, Lijuan, et al., 2023. Enhancing psychological counseling with large language model: A multifaceted decision-support system for non-professionals. arXiv preprint arXiv:2308.15192.

Gamon, Michael, Choudhury, Munmun, Counts, Scott, Horvitz, Eric, 2013. Predicting depression via social media. In: Proceedings of the Association for the Advancement of Artificial Intelligence. AAAI, Association for the Advancement of Artificial Intelligence, URL https://www.semanticscholar.org/paper/Predicting-Depression-via-Social-Media-Choudhury-Gamon/e992de7c53579bd1f23c80be17ca3f6692593ea6.

Gautam, Geetika, Yadav, Divakar, 2014. Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In: 2014 Seventh International Conference on Contemporary Computing. IC3, IEEE, pp. 437–442.

GBDResults, 2023. Institute for health metrics and evaluation. URL https://vizhub.healthdata.org/gbd-results/. Accessed: 2023-12-18.

Graves, Alex, Schmidhuber, Jürgen, 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Net. 18 (5–6), 602–610.

Haque, Farsheed, Nur, Ragib Un, Al Jahan, Shaeekh, Mahmud, Zarar, Shah, Faisal Muhammad, 2020. A transformer based approach to detect suicidal ideation using pre-trained language models. In: 2020 23rd International Conference on Computer and Information Technology. ICCIT, IEEE, pp. 1–5.

Hasan, Md. Arid, 2023. Zero- and few-shot prompting with LLMs: A comparative study with fine-tuned models for bangla sentiment analysis. arXiv.org URL https://arxiv.org/abs/2308.10783. Accessed: 2023-12-18.

Hasan, Ali, Moin, Sana, Karim, Ahmad, Shamshirband, Shahaboddin, 2018. Machine learning-based sentiment analysis for twitter accounts. Math. Comput. Appl. 23 (1), 11.

Hochreiter, Sepp, Schmidhuber, Jürgen, 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Hu, Edward J, Shen, Yelong, Wallis, Phillip, Allen-Zhu, Zeyuan, Li, Yuanzhi, Wang, Shean, Wang, Lu, Chen, Weizhu, 2021. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

Ilham, Fikri, Maharani, Warih, 2022. Analyze detection depression in social media Twitter using bidirectional encoder representations from transformers. J. Inf. Syst. Res. 3 (4), 476–482. http://dx.doi.org/10.47065/josh.v3i4.1885, Accessed: 2023-12-18.

Ji, Shaoxiong, Zhang, Tianlin, Yang, Kailai, Ananiadou, Sophia, Cambria, Erik, 2023. Rethinking large language models in mental health applications. arXiv preprint arXiv:2311.11267.

Kabir, Mohsinul, Islam, Mohammed Saidul, Laskar, Md Tahmid Rahman, Nayeem, Mir Tafseer, Bari, M Saiful, Hoque, Enamul, 2023. BenLLMEval: A comprehensive evaluation into the potentials and pitfalls of large language models on Bengali NLP. arXiv preprint arXiv:2309.13173.

Lan, Zhenzhong, Chen, Mingda, Goodman, Sebastian, Gimpel, Kevin, Sharma, Piyush, Soricut, Radu, 2019. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.

Latif, Ehsan, 2023. Fine-tuning ChatGPT for automatic scoring. arXiv.org URL https://arxiv.org/abs/2310.10072. Accessed: 2023-12-18.

Latif, Ehsan, Zhai, Xiaoming, 2024. Fine-tuning chatgpt for automatic scoring. Comput. Educ.: Artif. Intell. 100210.

Maloney, Divine, Zamanifard, Samaneh, Freeman, Guo, 2020. Anonymity vs. Familiarity: Self-disclosure and privacy in social virtual reality. In: Proceedings of the 26th ACM Symposium on Virtual Reality Software and Technology. VRST '20, Association for Computing Machinery, New York, NY, USA, ISBN: 9781450376198, http://dx.doi.org/10.1145/3385956.3418967.

MayoClinic2022, 2022. Depression (major depressive disorder). URL https://www.mayoclinic.org/diseases-conditions/depression/symptoms-causes/syc-20356007. Accessed: 2023-6-26.

Mohanty, Shikha, 2024. Unveiling the significance of response relevance and completeness in LLMs. UpTrain AI URL https://blog.uptrain.ai/unveiling-the-significance-of-response-relevance-and-completeness-in-llms/.

Mumu, Tabassum Ferdous, Munni, Ishrat Jahan, Das, Amit Kumar, 2021. Depressed people detection from bangla social media status using lstm and cnn approach. J. Eng. Adv. 2 (01), 41–47.

NLTK, 2023. NLTK :: nltk.tokenize package. URL https://www.nltk.org/api/nltk.tokenize.html. Accessed: 2023-12-18.

OpenAI, 2023. GPT-4 technical report. arXiv.org URL https://arxiv.org/abs/2303.08774. Accessed: 2023-12-18.

Reddit, 2023. Reddit dataset: r/depression and r/SuicideWatch. URL https://www.kaggle.com/datasets/xavrig/reddit-dataset-rdepression-and-rsuicidewatch. Accessed: 2023-12-18.

sahajBERT, 2023. Bangla-stemmer. URL https://huggingface.co/neuropark/sahajBERT. Accessed: 2023-12-18.

Salas-Zárate, Rafael, Alor-Hernández, Giner, Salas-Zárate, María del Pilar, Paredes-Valverde, Mario Andrés, Bustos-López, Maritza, Sánchez-Cervantes, José Luis, 2022. Detecting depression signs on social media: a systematic literature review. In: Healthcare. 10, (2), MDPI, p. 291.

Sarker, Sagor, 2020. Banglabert: Bengali mask language model for bengali language understading. textsIGitHub.

Sentiment140, 2023. Sentiment140 dataset with 1.6 million tweets. URL https://www.kaggle.com/datasets/kazanova/sentiment140. Accessed: 2023-12-18.

Shah, Faisal Muhammad, Ahmed, Farzad, Joy, Sajib Kumar Saha, Ahmed, Sifat, Sadek, Samir, Shil, Rimon, Kabir, Md Hasanul, 2020a. Early depression detection from social network using deep learning techniques. In: 2020 IEEE Region 10 Symposium. TENSYMP, IEEE, pp. 823–826.

Shah, Faisal Muhammad, Haque, Farsheed, Nur, Ragib Un, Al Jahan, Shaeekh, Mamud, Zarar, 2020b. A hybridized feature extraction approach to suicidal ideation detection from social media post. In: 2020 IEEE Region 10 Symposium. TENSYMP, IEEE, pp. 985–988.

Uddin, Abdul Hasib, Bapery, Durjoy, Arif, Abu Shamim Mohammad, 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, IEEE, pp. 1–4.

Wang, Qianying, Tian, Jie, Li, Meng, Lu, Ming, 2023. Text classification based on CNN-BIGRU and its application in telephone comments recognition. Int. J. Comput. Intell. Appl. 22 (04), http://dx.doi.org/10.1142/s1469026823500219, URL https://typeset.io/papers/text-classification-based-on-cnn-bigru-and-its-application-2icr0kev.

World Health Organization, 2023. Depressive disorder (depression). URL https://www.who.int/news-room/fact-sheets/detail/depression.

Xu, Xuhai, Yao, Bingsheng, Dong, Yuanzhe, Gabriel, Saadia, Yu, Hong, Hendler, James, Ghassemi, Marzyeh, Dey, Anind K, Wang, Dakuo, 2023. Mental-LLM: Leveraging large language models for mental health prediction via online text data. arXiv preprint arXiv:2307.14385.

Yang, Kailai, Zhang, Tianlin, Kuang, Ziyan, Xie, Qianqian, Ananiadou, Sophia, 2023. Mentalllama: Interpretable mental health analysis on social media with large language models. arXiv preprint arXiv:2309.13567.

Ye, Junjie, 2023. A comprehensive capability analysis of GPT-3 and GPT-3.5 series models. arXiv.org URL https://arxiv.org/abs/2303.10420. Accessed: 2023-12-18.