A BERT-based sentiment analysis model for depressive text

Shuo Wang*
Hubei University of Technology
Wuhan, Hubei, China
18206375906@163.com

Yuqi Lu
Hubei University of Technology
Wuhan, Hubei, China
18575505915@163.com

Ruijue Luo
Karlsruhe Institute of Technology
Karlsruhe, Baden-Württemberg
Germany
ruijue.luo@student.kit.edu

Boxuan Li
Hubei University of Technology
Wuhan, Hubei, China
xuanzi427dd@163.com

Yingjie Zhang
The Hong Kong Polytechnic
University
Hong Kong, China
yingjieyj.zhang@connect.polyu.hk

Liang Xiao†
Hubei University of Technology
Wuhan, Hubei, China
lx@mail.hbut.edu.cn

Abstract

Depression is a big public health problem around the world. Like the traditional methods of screening, clinical assessments, and self-rating scales are among the limits. These methods are based on the subjectivity of patient reports and the experience of doctors and clearly can have bias. In addition, they do not permit real-time monitoring of changes in mood. For this reason, this field of mental health is challenged with accurately and efficiently detecting probable mood changes in individuals at risk for depression. To resolve these challenges, we then propose a BERT-based sentimental analysis model adapted for depression-related text. Current methods for screening for traditional depression and established sentiment analysis models are hindered by data scarcity, data imbalance, and the problems of deriving emotional changes in a dynamic and clean manner. To leverage the strengths of BERTs in our model, we train the last layer of BERTs and then construct a classification layer tailored to the depressive emotion, which is more sensitive to the semantic information of the mood. Moreover, we also present data augmentation in its simplest form, which is the rearrangement of positive samples to increase the diversity of samples while still preserving the semantic meaning. To tackle data imbalance, we introduce a dynamic weight mechanism to the BCEWithLogitsLoss function by adjusting the weight of the positive samples with data distribution. The model effectively captures the complex emotional changes by combining global and local semantic dynamic representations. In our empirical study, we analyze the open-source dataset of interview transcripts of 162 volunteers. Our model is evaluated using the experimental results, and the precision, the recall rate, and the F1 score obtained are 0.667, 0.667, and 0.667, respectively. Our model has achieved 2.62 higher in terms of precision and F1 score as compared to the existing models. The results indicate that our model provides a big benefit in identifying screened individuals at risk for depression.

CCS Concepts

• **Computing methodologies** → Artificial intelligence; Natural language processing.

Keywords

Depression screening, BERT model, sentiment analysis, text classification, medical text processing

ACM Reference Format:
Shuo Wang, Yuqi Lu, Ruijue Luo, Boxuan Li, Yingjie Zhang, and Liang Xiao. 2025. A BERT-based sentiment analysis model for depressive text. In *2025 International Conference on Generative Artificial Intelligence and Digital Media (GADM 2025), March 14–16, 2025, Dali, China.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3734921.3734934

1 Introduction

Psychological disorder is known to be depression and one of the major afflicting problems all over the world. Traditional depression screening methods make use of clinical assessments and self-rating scales solely. However, such methods are to a considerable extent subject to influence by the reports of patients and their doctors. In addition, these methods are not able to monitor emotional changes in real-time. This has compelled the area of mental health to focus on the challenge of detecting the emotional states of individuals who are at risk of depression efficiently and accurately.

Increasingly in recent years, much attention has been drawn to the development of methods for depression screening based on natural language processing (NLP) that would make use of patient text data. These provide clues to seeking out possible emotional problems as well as serving as some supplemental help to evaluate clinical cases. But conventional ways of emotional analysis like dictionary-based sentiment analysis, support vector machine (SVM), and naive Bayes classifier tend to struggle with complex terminologies and blended emotions in medical texts. The problem is especially apparent when one is trying to grasp the global context semantics of the text.

During the time of deep learning technologies, long short-term memory (LSTM) networks and convolutional neural network (CNN) models have been used in sentiment analysis tasks. LSTM networks

*Shuo Wang and Yuqi Lu: Equal contribution.
†Corresponding author

GADM 2025, Dali, China
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1346-0/2025/03
https://doi.org/10.1145/3734921.3734934

are good at capturing long-term dependencies with their gating mechanisms, while CNNs help them extract local features and extract slight emotional changes. Nevertheless, the use of these methods is limited in addressing global semantics, mainly in the case of medical texts where the understanding of emotions relies on local semantic expressions. The limitation, however, enables only minor interpretation of complex medical information.

In recent years, one of the pre-trained language models, BERT, based on transformer architecture, has improved the understanding of global semantics. BERT's unsupervised pre-training on large-scale corpora enables the capturing of bidirectional contextual information, and one can improve both global and local semantic understanding. Furthermore, in the medical field, specialized derivatives of BERT are also highly successful for the medical text sentiment analysis task, such as Bio-BERT and Med-BERT. Our model also uses the pre-traiing–fine-tuning paradigm that is inspired by these models.

However, transformer-based models have a few problems in the context of medical applications. Imbalanced datasets that yield majority class bias to the medical data are typical, due to which medical data is often scarce and expensive to annotate. In addition, even these models have difficulty with generalizing sentiment analysis in small sample, domain-specific tasks. However, traditional static embedding methods are unable to address the patient's change in emotionality on time. Therefore, there have been some optimization techniques proposed, including positive sample weighting and data augmentation. Nevertheless, they suffer from high algorithmic complexity and cannot avoid numerous noises.

Thus, we propose an improved depression screening framework using the pre-trained BERT model to overcome the above challenges. However, applying BERT to these problems, namely, data scarcity, imbalance, and dynamic emotional changes in medical sentiment analysis, requires some adjustments, and this framework is adapted to achieve that. The rest of this paper will discuss the technical details of this approach.

2 Related Works

2.1 Depression Detection Methods

Traditional Depression Screening Methods: Application of LSTM and CNN

In the depression screening task, long short-term memory networks and convolutional neural networks have been popular models. One characteristic of LSTM is that it can successfully learn to capture long-range dependencies in the information, something necessary when dealing with sequential data. LSTM is able to better understand the relationships between different sections of the text in depression screening than the sequence of text is. On the other hand, CNNs excel at picking up local features from text and learn to pick up patterns and structures in relatively short time by using different-sized convolution kernels.

It is known that people with depression use certain speech patterns, including selecting some words differently, using different amounts of emotional intensity, and putting sentences together in a different way. As such, depression screening will help in determining emotional states and offer good support for early intervention and treatment.

Poria et al. [1] presented a depression detection scheme on multitasking multidimensional data where an efficient combination of LSTM and CNN is made to model emotional tendency in text in 2020. It provides an effective and synergistic approach to understanding text emotions for which both models are exploited. Nevertheless, there are disadvantages to traditional methods.In fact, most of them use a fixed window size to extract local features, thus limiting flexibility in representing various complex contextual semantics. Besides, these methods of local text feature extraction tend to ignore the global semantic relationship of the document, which inevitably brings in an incomplete understanding of emotions[2].

These traditional methods inspired us to look for other procedures of depression screening that are more effective. Therefore, we had to introduce things such as pre-trained models like BERT, removing the need for using traditional feature extraction techniques. Unsupervised learning, a pre-trained language model BERT is a language model based on Transformer that learns semantic and syntactic information automatically from a large-scale corpus. In contrast to LSTM and CNN, BERT can all-inclusively capture the semantic context in the text by considering global and local information[3]. Furthermore, we utilized dropout and the use of random initialization to alleviate the overfitting issue as well as make the domain adaptable. This greatly increases the accuracy of depression screening such that the model can analyze the ups and downs of the emotions and learn about the potential propensity to become depressed in such people.

2.2 Data augmentation and imbalanced data processing

A common problem in sentiment analysis tasks also is small dataset sizes, which can be addressed by data augmentation. When data sets are unbalanced, data imbalance can adversely affect the performance of a model since the model strengthens favoring the majority class. In 2021, Feng et al.[5] suggested text data enhancement by syntactic and semantic replacement. In this technique, it generates new samples by changing sentence expressions, like replacing synonyms or changing the order of the words. Producing new samples semantically similar to the original ones but in different forms increases the number and diversity of the training data and also improves the model's generalization ability. However, since semantic substitution algorithms are complex and introduce noise that harms model performance, this method is not preferable.

This is why we chose a simpler and more effective strategy to overcome. New texts are created for positive samples by rearranging the sentence order. These procedures improved the diversity of positive samples with their semantic integrity. The new samples alter only the order of sentences, leaving their meaning unchanged, and thus the model is able to learn more on the features of positive samples. This is very good for the model's ability to recuperate positive instances.

This is the case in most of the cases of depression screening tasks, i.e., there are more negative samples than positive samples; hence, the data is not balanced. This makes the model pay less attention to learning the features of the positive samples, and it is underfitting the majority of the samples. Su et al. [6] proposed a way to modify the model's loss by assigning greater weights to positive samples so that during training the model will be paying more attention to learning the minority class rather than being biased towards negative examples.

Traditional weighting methods, typically rely on fixed weights that we improve over with data augmentation strategies. To address the issue of positive samples, we also achieve data augmentation with the help of dynamically adjusting the weights of the positive samples according to the proportion of samples in distinct emotion categories. This combination helps the performance of the model, specifically in the case of key evaluation metrics, i.e., precision, recall, and F1 score. Positively weighted samples imply that the model pays more attention to the minority samples and balances the amount of data to enhance the model's capability to identify depressive tendency behavior emotionally.

2.3 Sentiment analysis in medical text analysis and limitations of pre-trained models

In the context of medical text sentiment analysis, specifically in the area of depression screening, pre-trained models like Bio-BERT and Med-BERT have been extraordinarily useful. Nevertheless, there are a number of obstacles to overcome in order for depressive emotions to be accurately detected. While these models achieve strong generalization in the medical domain, being able to understand medical terminology and common emotional expressions, they struggle with particular downstream tasks such as depression-related sentiment analysis. This is especially true for small or imbalanced datasets since the models are trained on a large amount of general medical corpora, and it's hard for them to learn nuanced emotional features from depression screening. [12]

Thus, we have optimized BERT for depression-related sentiment analysis to overcome these challenges. To this end, we improve the model's ability to detect these emotions by introducing a specialized classification layer and fine-tuning it to self-label depressive emotions. [4]Moreover, we also apply data augmentation methods like weighted negative and positive samples and sampling diverse positive samples to alleviate data imbalance and strengthen the generalization ability of the model. Assisting the model in identifying the depressive emotional features even in reduced resource scenarios empowers these strategies to bolster the performance and adaptability of the model in its deployment in real-world applications.

3 Data collection and system architecture

In this section, we provide data collection and analysis training and the comprehensive architecture of the text sentiment analysis model for depression screening.

3.1 Dataset

Although there are depression-related datasets proposed for depression, very few authors in the field of depression have depression data in Chinese. However, in recent years, Ying Shen et al. from [11]introduced the Chinese depression dataset they called EATD-Corpus. This dataset is a multidimensional evaluation content with a clear basis for grouping, so it suffices our needs.

The results described here were built using 62 volunteers located at Tongji University on audio and text from interviews, which comprised the EATD-Corpus. Their information was all confirmed through informed consent forms signed by all volunteers. Three random questions were selected to be asked to each volunteer, and a 20-item SDS questionnaire was used to assess four aspects of depression: general affective, physical, other disorders, and psychomotor activity.[7] According to the Chinese standard (SDS score ×1.25 ≥ 53 as depression), the 162 volunteers were divided into a depression group (30 participants) and a non-depression group (132 participants). The total audio duration in the dataset was approximately 2.26 hours.

The dataset construction involved two steps: data collection and data preprocessing. Data collection was performed using a specially developed app. Mute or short audio segments were removed, background noise was eliminated using RNNoise, and audio transcripts were extracted using Kaldi. Finally, the transcripts were manually checked and corrected to ensure high data quality.

3.2 System architecture

Figure 1 illustrates the complete architecture of our model. At the core of the system is a customized BERT model that extracts and classifies features from depression-related texts.

First, the system receives input data from the EATD-Corpus dataset, including participants' text records and their corresponding labels. The text is tokenized into word slices, and data augmentation is performed on positive samples as needed. The augmented text data is then encoded into fixed-length input vectors, preparing them for model input.

During the model training phase, the training loss is recorded for each epoch, and the model's performance is evaluated using test data. Evaluation metrics include Precision, Recall, and F1 Score. Finally, the model generates predictive outcomes (depressed/non-depressed) along with their corresponding probabilities. These results are used for subsequent system deployment or model weight updates. When sufficient data accumulates, the system retrains the model to ensure continuous performance optimization.

4 Classification model

4.1 Word vector representation

In text classification tasks, word vector representation plays a crucial role in the model's ability to process natural language input. In this study, we use the pre-trained BERT model as a tool for feature extraction. Based on the Transformer architecture[8], BERT effectively captures semantic information in the context of text through bidirectional language modeling. In the implementation, we use the BERT-base-Chinese model for training. First, the sentences in the text are tokenized and mapped into fixed-length word vectors. The dimension of each word vector corresponds to the hidden layer size of the BERT model. To fully utilize the context information of the text, the model outputs a pooled vector during feature extraction[9]. Specifically, the hidden state corresponding to the [CLS] tag of the BERT model is used as the global semantic representation of the entire text. To further enhance the robustness and generalization ability of the model, we apply a Dropout operation after feature extraction. This randomly deactivates certain features to reduce the risk of overfitting. Finally, the features are non-linearly transformed through the fully connected classification layer to produce a single

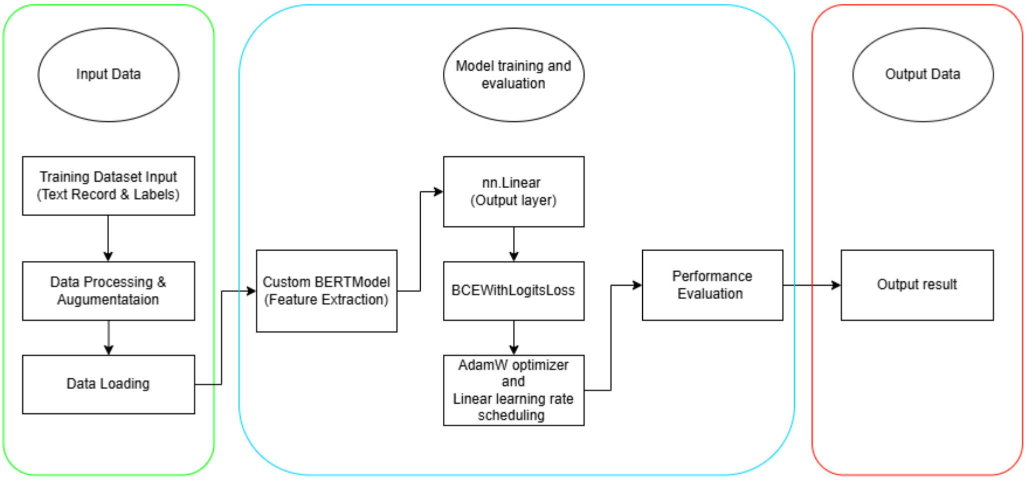
Input Data
Training Dataset Input (Text Record & Labels)
Data Processing & Augumentataion
Data Loading
Custom BERTModel (Feature Extraction)
Model training and evaluation
nn.Linear (Output layer)
BCEWithLogitsLoss
AdamW optimizer and Linear learning rate scheduling
Performance Evaluation
Output Data
Output result

Figure 1: Process Architecture Diagram of Depression Text Sentiment Analysis Model Based on BERT

classification probability. The entire word vector representation process captures the semantic relationship within the sentence and expands the expression of positive samples through data augmentation[10], thereby providing more comprehensive and high-quality feature representations for the subsequent classification tasks.

4.2 BERT pre-trained model

BERT (Bidirectional Encoder Representations from Transformers) is a landmark model in the field of natural language processing. Its core idea is to capture semantic information from text context through deep bidirectional encoders. Based on the Transformer encoder architecture, BERT combines a bidirectional attention mechanism with pre-training tasks, which significantly improves its performance in various downstream tasks.

The basic structure of the BERT model consists of stacked Transformer encoders. Each encoder layer comprises two key components: the Multi-Head Self-Attention mechanism and the Feedforward Neural Network. The input text is first processed into a sequence of words, then mapped into a word embedding space. Positional Encoding is added to incorporate sequence information. The self-attention mechanism in each Transformer layer captures the relationship between all words in the sequence, and the calculation formula for self-attention is as follows 1:

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

where Q,K,V represent the query, key and value vectors, respectively, and the dimension of the key vector is used as a scaling factor to prevent the inner product value from being too large. $\mathrm{d_k}$The multi-head attention mechanism captures the features of different subspaces by introducing multiple attention heads, which are calculated as follows 2 and 3:

$$\begin{gathered}\mathrm{MultiHead}\left(\mathrm{Q},\mathrm{K},\mathrm{V}\right) = \\ Concat\left(\mathrm{head_1},\mathrm{head_2},\ldots,\mathrm{head_h}\right)\mathrm{W^O}\end{gathered} \tag{2}$$

$$\mathrm{head_i} = Attention\left(\mathrm{QW_i^Q},\mathrm{KW_i^K},\mathrm{VW_i^V}\right) \tag{3}$$

Each head independently learns different subspace representations and then concatenates its outputs to form the final result of multi-head attention. After passing through the self-attention module, BERT performs a nonlinear transformation of the representation of each marker with the formula 4 and 5:

$$\mathrm{H^{l+1}} = LayerNorm\left(\mathrm{H^l} + \mathrm{FFN}\left(\mathrm{H^l}\right)\right) \tag{4}$$

$$\mathrm{FFN}\left(\mathrm{x}\right) = ReLU\left(\mathrm{xW_1} + \mathrm{b_1}\right)\mathrm{W_2} + \mathrm{b_2} \tag{5}$$

where is the output $\mathrm{H^l}$of layer l, FFN is a feedforward neural network, and a two-layer fully connected network is used to enhance the expression ability of the model. The Residual Connection and Layer Normalization are used to optimize the training stability of each layer.

In order to improve the applicability of the model in the downstream tasks, BERT designed two kinds of early training missions: mask Language Model (Masked Language Model, MLM) and Next Prediction (Next Sentence Prediction, NSP). MLM, through the random cover part of the input sequence tags, asks models to predict the concealed tag. The loss function is defined as 6:

$$L_{MLM} = -\sum_{i\in M\backslash} logP\left(x_i|X_{\backslash textmasked}\right) \tag{6}$$

where i is the index set of masked markers, representing the input sequence after masking. $\mathcal{M}\mathrm{X_{masked}}$NSP (Next Sentence Prediction) is designed to help the model understand semantic relationships between sentences. It works by determining whether two pieces of text are continuous. The binary loss function for NSP is defined as follows 7:

$$\mathcal{L}_{\mathcal{NSP}} = -\left(\mathrm{y}\log\mathrm{P}\left(\mathrm{isNext}\right) + \left(1-\mathrm{y}\right)\log\mathrm{P}\left(\mathrm{notNext}\right)\right) \tag{7}$$

The input to BERT takes a fixed format and is defined by special tokens [CLS] and [SEP]. Specifically, each input sequence begins with [CLS] as a global representation of the whole sequence, while

[SEP] is used to separate sentences. In the output layer of the model, the hidden state vector of [CLS] is designed to be a semantic representation of the entire text with the following formula 8:

$$P\left(\mathrm{y}|\mathrm{X}\right) = \sigma\left(\mathrm{Wh}_{[\mathrm{CS}} + \mathrm{b}\right) \tag{8}$$

where W and b are the weights and biases of the fully connected layer, are the hidden states corresponding to the $\mathrm{h}_{[\mathrm{CLS}]}$[CLS] tags, and are the activation functions (usually $\sigma(\cdot)$sigmoid or softmax). With this design, BERT is able to compress the entire input sequence into a fixed-length vector, which facilitates tasks such as classification, regression, and so on. BERT is trained using large-scale corpora (such as Wikipedia and BooksCorpus) for unsupervised learning, and Adam optimizer is used to adjust the parameters. The weights generated in the pre-training stage can be transferred to various downstream tasks to achieve efficient application under supervised training with a small amount of data.

The powerful performance of BERT in natural language processing is due to its bidirectional coding mechanism, innovative pre-training tasks, and powerful model capacity. However, BERT faces some challenges in computational resources and inference speed due to its highly complex architecture, which have been partially solved by techniques such as Distillation.

4.3 Depression classification model design

The sentiment classification model designed in this paper is based on the pre-trained BERT model, combined with a custom classification head and optimization strategy, aiming to achieve accurate classification of text sentiment. The BERT model is built through a multi-layer bidirectional Transformer encoder. The core of the Bert model is a self-attention mechanism, which generates a deep semantic representation by capturing the context dependence of the input text. In this study, the model architecture is composed of a BERT encoder, custom classification head, and Dropout layer. The data augmentation strategy and weight initialization method are combined to improve the performance of the model further in the sentiment classification task. The input of the model consists of text sequences, which are transformed into Input IDs, Attention masks, and Token Type IDs after processing by BERT word segmentation. These features are fed to the BERT encoder for context modeling, and the last layer output of the encoder contains the context representation of each token. In particular, the hidden states of BERT's [CLS] tags are designed to be global semantic representations of whole pieces of text for downstream classification tasks. On this basis, the model adds a layer of fully connected classification heads to map the output of [CLS] into the probability distribution of classification by linear transformation, which is calculated as follows 9:

$$\mathrm{logits} = \mathrm{W} \cdot \mathrm{h}_{[\mathrm{CLS}]} + b \tag{9}$$

Where logits is the hidden state corresponding to $\mathrm{h}_{[\mathrm{CLS}]}$BERT's [CLS] tag W and b are the weights and biases of the classification head. In order to reduce the risk of model overfitting, a Dropout layer is added before the fully connected layer, and its retention probability is set to 0.3 to enhance the model's robustness by randomly masking some neurons' output. In addition, to adapt to the emotion classification task, the weights of encoders in the last six layers of BERT are randomly initialized, and the pre-trained weights of the other layers are retained. This initialization process uses a Gaussian distribution with a mean of 0 and a standard deviation of 0.02 to improve the model's ability to adapt to specific tasks.

In terms of data processing, according to the characteristics of the sentiment classification task, this paper conducts fine labeling and enhancement of the data. Samples with sentiment scores greater than 53 were labeled as positive samples; the rest were negative. In order to solve the problem of an unbalanced proportion of positive and negative samples in the dataset, this paper adopts a data enhancement strategy based on sentence rearrangement, especially for positive samples. Specifically, each positive sample is composed of three clauses, and all possible sentence orders are generated by permutation and combination to form new training samples, thus greatly increasing the diversity and number of positive samples. In addition, after being encoded by the BERT word splitter, the text is transformed into a fixed-length tensor representation. The samples with insufficient length are filled up, and the samples beyond the limit are truncated. This method not only ensures the consistency of the input data format but also retains the key information of the text to the maximum extent.

The Loss function of the proposed model uses Weighted Binary Cross-Entropy loss to alleviate the effect of class imbalance on the performance of the model. The positive and negative sample weights are set to 4 and 1, respectively. The loss function is calculated using the following formula 10:

$$L = -\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \begin{bmatrix} \mathrm{w}^{+}\mathrm{y}_{\mathrm{i}} \log\left(\sigma\left(\mathrm{z}_{\mathrm{i}}\right)\right) \\ +\mathrm{w}^{-}\ \left(1 - \mathrm{y}_{\mathrm{i}}\right) \log\left(1 - \sigma\left(\mathrm{z}_{\mathrm{i}}\right)\right) \end{bmatrix} \tag{10}$$

where w^{+} and w^{-} are the positive and negative sample weights, y is the true label, z is the model predicted value, and σ is the sigmoid activation function. In the experiment of the optimization strategy, the AdamW optimizer for the experiments and different learning rates assigned to the BERT encoder and the classification head are set. The former is 1×10^{-5}, and the latter is 5×10^{-4}. In addition, the model introduces a linear learning rate scheduler, which gradually increases the learning rate in the early stage of training and then gradually decreases it, ensuring the stability and efficiency of training. The training process of the model includes two stages: forward propagation and backpropagation. In the forward propagation stage, the model generates the predicted value of sentiment classification through the BERT encoder and the classification head, and calculates the loss between the predicted value and the true label. In the backpropagation phase, the model uses the gradient descent algorithm to optimize the parameters to minimize the loss function. The whole training process is carried out in a batch size of 2, with a total of 158 iterations. At the end of each cycle, the model is tested on the validation set to evaluate its performance. In this paper, Precision, Recall, and F1 value are used as the main indicators of model performance evaluation. The formulas for calculating these metrics are as follows 11:

$$\begin{gathered} \mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}, \quad Recall = \frac{\mathrm{TP}}{\mathrm{TP+FN}}, \\ F1 = 2 \cdot \frac{\mathrm{Precision \cdot Recall}}{\mathrm{Precision+Recall}} \end{gathered} \tag{11}$$

here, TP, FP, and FN denote true positive, false positive, and false negative, respectively. The experimental results show that the proposed model shows good robustness under the condition of unbalanced proportion of positive and negative samples, especially achieving significant improvement in the recall rate and F1 value of positive samples.

5 Experimental Results and Analysis

In this section, we introduce the experimental environment and conduct comparative experiments and ablation studies.

5.1 Experimental environment

5.1.1 Dataset. The dataset used in this experiment is the publicly available EATD-Corpus. Each subfolder within the dataset contains text files named negative.TXT, neutral.TXT, positive.TXT, and a corresponding new_label.TXT file representing label information. Samples are labeled as positive (1) or negative (0) based on whether the score in new_label.txt is greater than 53. For positive samples, data augmentation techniques were used to generate new samples through various combinations of text orders to enrich sample diversity. Negative and neutral samples do not undergo data augmentation.

5.1.2 Model setup. We propose an improved depression screening framework based on the BERT pre-trained model, referred to as **Mood-BERT**. The model consists of a pre-trained BERT component and a custom classifier. Leveraging BERT's bidirectional pre-training on a large-scale corpus, the model can capture both global and local semantic information, making it suitable for handling complex emotional patterns in medical texts. We have optimized the BERT model to address challenges such as data scarcity, imbalance, and dynamic emotional changes in medical sentiment analysis. Specifically, we randomly initialized the weights of the last few BERT layers to enhance task adaptability and used a custom linear classifier with a Sigmoid activation function for binary classification.

The data loading module dynamically loads preprocessed data in batches for both training and testing sets. The training module optimizes model parameters using the BCEWithLogitsLoss function, assigning higher weights to positive samples to balance class distribution. Additionally, the AdamW optimizer and a dynamic learning rate scheduler are employed to improve training efficiency and stability.

Given the issue of data imbalance in depression screening—particularly the scarcity of positive (depressive) emotional samples—we adopted positive sample weighting techniques. By increasing the weight of minority class samples, the model focuses more on rare emotional features during training. A dynamic weight adjustment mechanism further adjusts weights based on the distribution of each batch, enhancing model adaptability at different training stages.

For data augmentation, we generate diverse training samples by rearranging and recombining emotional units. This strategy enhances the model's sensitivity to emotional changes and produces more training samples while maintaining semantic integrity.

5.1.3 Benchmarking and comparison. In this study, we compare Mood-BERT with several benchmark models using only text modality data scores for contrast:

- **LSTM**: Based on a Bi-LSTM structure for processing long sequences.
- **SVM**: Uses an optimal hyperplane for classification, employing kernel functions for nonlinear data.
- **RF**: Integrates predictions from multiple decision trees trained by random sampling.
- **Decision Tree**: Gradually divides data based on features, making decisions in a tree structure.
- **BiLSTM**: Processes sequences bidirectionally to capture more information suitable for NLP tasks.

5.2 Overall performance and comparison

We analyze the performance of each model in comparison to Mood-BERT. The experimental results are presented in the accompanying table, where the best results are in bold and the second-best are underlined.

The table shows that Mood-BERT achieves the best performance with an F1 score of 0.667 and the highest accuracy of 0.667. Theoretically, random results for these three metrics would usually fluctuate around 0.5. Among the other models, SVM achieved a recall of 1.00 but a precision of 0.48. LSTM, RF, and Decision Tree models showed moderate performance across metrics. Table 1 summarizes the results.

Notably, the F1 score of Mood-BERT is 0.667, which is 2.62% higher than the second-best BiLSTM model (0.660), highlighting its superior balance of precision and recall. The recall values of Mood-BERT and BiLSTM are very similar (0.667 vs. 0.660), indicating comparable effectiveness in retrieving a sufficient number of positive samples. However, Mood-BERT's precision is 2.62% higher than that of BiLSTM, demonstrating that its predictions are more accurate. Overall, Mood-BERT performs largely better in terms of precision and F1 score performance in the depression screening task. Though other models might be stellar at a particular metric while Mood-BERT is not, it generally outperforms them overall. However, the SVM model has a high error rate of prediction, and this is not practical for use in real situations since there are better performance models available, such as LSTM, Random Forest (RF), and Decision Tree, which are also slightly flawed in screening depression scenarios.

In sum, Mood-BERT outperforms baselines in the screening of depression tasks and demonstrates high specificity and F1 score. Other models have their own things to offer in terms of different metrics, though the overall performance comparison remains quite challenging with Mood-BERT. With the given dataset, the accuracy of the SVM model is too low and should be used with careful consideration in the practical application. Regarding models like LSTM, RF, and Decision Tree, their performance on a number of metrics is not anything extraordinary, while their effectiveness in screening for depression scenarios is not particularly significant.

5.3 Ablation experiments

For an ablation study to evaluate the effect of each of the four main components of Mood-BERT, namely data augmentation, dropout layers, and freezing the part of the BERT model layers, we independently deactivated each one. Table 2 summarizes the results.

As observed, disabling data augmentation and the dropout layer results in a great drop in precision and F1 score. Furthermore, when the Dropout layer is turned off, the recall plateaus at 1.0, and with a

Table 1: Comparison Table of Performance Indicators of Different Models in Depression Text Sentiment Analysis

Models	**F1**	**Recall**	**Precision**
Decision Tree	0.490	0.530	0.590
LSTM	0.570	0.630	0.530
RF	0.570	0.530	0.610
SVM	0.640	1.000	0.480
BiLSTM model	0.6 50	0. 660	0.650
Mood-BERT	**0.667**	**0.667**	**0.667**

Table 2: Summary Table of Ablation Experiment Results for Depression Text Sentiment Analysis Model

Models	Precision	Recall	F1
WITHOUT-data augmentation	0.250	0.667	0.363
WITHOUT-Dropout layer	0.037	1.000	0.073
Frozen layer 1	1.000	0.333	0.500
Frozen layer 2	0.250	0.333	0.285
Frozen layer 3	0.055	0.667	0.102
Full Model	0.667	0.667	0.667

higher loss of accuracy in detecting depressive cues. These metrics were equally led astray by the first three layers of the BERT model being frozen.

This shows that BERT model layers should be maintained, and dropout layers and data augmentation are necessary for Mood-BERT's performance.

6 DISCUSSION

Data Enhancement Strategy: This study proposes a BERT-based sentiment analysis model for depression-related texts, achieving significant improvements in precision, recall, and F1 score. However, the current data augmentation strategy, based on sentence rearrangement, though effective in enhancing sample diversity, is limited in capturing complex emotional variations such as shifts in intensity or implicit expressions. Advanced techniques like synonym substitution, contextual rewriting, or semantic similarity sampling could address these limitations by introducing greater linguistic variability, though they may increase computational complexity. Despite these challenges, the sentence rearrangement strategy demonstrated clear benefits, improving sample diversity and boosting performance, with an F1 score increase of 2.62% compared to the next-best model. Its simplicity and efficiency make it suitable for resource-limited settings. Future work could explore more advanced augmentation techniques, refine model architectures, and validate the system in clinical applications to enhance its robustness and practicality.

Dynamic Weighting Mechanism: The dynamic weighting mechanism in this study adjusts weights based on sample category proportions, effectively addressing class imbalance and improving model performance. However, it does not account for semantic difficulty or complexity, which may lead to the neglect of key features in nuanced or challenging texts. Research indicates that factors like perplexity or linguistic intricacy can significantly impact a model's learning effectiveness. Incorporating these factors into the weighting mechanism could enhance the model's focus on critical features in depression-related texts. Future improvements could integrate semantic complexity measures, such as perplexity scores or attention-based importance metrics, to assign higher weights to semantically challenging samples. While this may increase computational demands, it could improve the model's robustness and ability to capture subtle patterns. Despite these limitations, the current mechanism has achieved strong performance, balancing class distributions effectively and boosting precision, recall, and F1 score. Its simplicity and efficiency make it suitable for imbalanced datasets. Future work should explore incorporating semantic complexity into the weighting strategy to further enhance its adaptability and performance.

Initialization Strategy: This study employs a random initialization strategy for some BERT layers, which has shown initial improvements in model adaptability and performance. However, this approach does not fully explore optimal layer configurations, potentially limiting the model's ability to leverage BERT's pre-trained knowledge. Randomly initializing certain layers may result in suboptimal feature learning, particularly for tasks requiring deeper semantic understanding. Future improvements could explore layer-wise initialization, where shallow layers retain pre-trained weights for general semantics, while deeper layers are randomly initialized for task-specific adaptation. Tools like Optuna could also systematically optimize initialization configurations and learning rates, providing a more structured approach. Despite these limitations, the current strategy has achieved notable experimental results, improving metrics such as F1 score and precision. Its simplicity and adaptability make it practical, but future work should aim to refine initialization methods to further enhance model performance.

7 CONCLUSION

In this work, we develop a new sentiment analysis model based on BERT for screening depression-related text. With the help of fine-tuned BERT and designing some specialized classification layers, we make the model more adaptable for depressive emotions. In addition, we present a permutation and combination method to create positive sample text in different orders. Furthermore, the method not only enhances the sample diversity but also has semantic consistency, which contributes a lot to the model's ability to identify positive emotional signals.

Additionally, we suggest a dynamic weight strategy in the BCEWithLogitsLoss function to dynamically adjust the weight of the positive samples in real time. This alleviates the need for manual selections for the imbalanced data. In addition, the model can also capture complex emotional changes across the text using dynamic representations of global as well as local semantic features.

In comparison to previous models, the experimental results reported in this paper indicate that our method has the highest performance, outperforming previously set state-of-the-art. We demonstrate how the performance is impacted on all of the ablation experiments: the use of data augmentation, dropout layers, and BERT model layers. In particular, the specialized classification layer for depression screening is important for enhancing detection accuracy.

In future work, we will try to optimize the model architecture even further and range out superior BERT fine-tuning strategies. Another thing we will also work on is innovating classification layer designs as well as setting up data augmentation techniques. We will also handle data imbalance, improve the dynamic weight mechanism, and find more semantic features using new sampling strategies. We wish to investigate cross-modal fusion, multi-variate evaluation indicators, optimization in training, and interpretability of a model. We will also validate the system in the clinical support settings to test its practical utilization.

References

[1] Poria S, Majumder N, Hazarika D, *et al.* Recognizing emotion cause in conversations[J]. Cognitive Computation, 2021, 13: 1317-1332.

[2] Zhang Y, Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1510.03820, 2015.

[3] Rasmy L, Xiang Y, Xie Z, *et al.* Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction[J]. NPJ digital medicine, 2021, 4(1): 86.

[4] Lee J, Yoon W, Kim S, *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36(4): 1234-1240.

[5] Feng S Y, Gangal V, Wei J, *et al.* A survey of data augmentation approaches for NLP [J]. Journal of arXiv preprint arXiv: 2105.03075, 2021.

[6] Su Y, Zhang R, Erfani S, *et al.* Detecting beneficial feature interactions for recommender systems[C]//Proceedings of the AAAI conference on artificial intelligence. 2021, 35(5): 4357-4365.

[7] Devlin J. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[8] Vaswani A. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017.

[9] Ghosh S, Anwar T. Depression intensity estimation via social media: A deep learning approach[J]. IEEE Transactions on Computational Social Systems, 2021, 8(6): 1465-1474.

[10] Y. Shen, H. Yang and L. Lin, "Automatic Depression Detection: an Emotional Audio-Textual Corpus and A Gru/Bilstm-Based Model," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 6247-6251, doi: 10.1109 / ICASSP43922.2022.9746569.

[11] Y. Shen, H. Yang and L. Lin, "Automatic Depression Detection: an Emotional Audio-Textual Corpus and A Gru/Bilstm-Based Model," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 6247-6251, doi: 10.1109 / ICASSP43922.2022.9746569.

[12] Y. Zhang, J. Chen, Z. Zou, M. Yao, S. Zhang and L. Xiao, "CSIA-GCN: A Doctor Recommendation Model Based on Interactive Graph Convolutional Networks," 2024 International Joint Conference on Neural Networks (IJCNN), Yokohama, Japan, 2024, pp. 1-8, doi: 10.1109/IJCNN60899.2024.10650337.