Mental Health Assessment Based on Natural Language Processing

Chenyu Liu
College of Computer Science, Chongqing University, Chongqing, 401331, China

Keywords: Natural Language Processing, Mental Health Assessment, BERT.

Abstract: As mental health issues gradually become one of the top global concerns, an effective and scalable assessment tool is needed. Traditional psychological self-assessment scales, such as the Beck depression scale and the self-rating anxiety scale, are widely used in clinical diagnosis. However, they rely on self-report from subjects and are easily influenced by subjective emotions, environmental factors, and comprehension abilities. As a result, they may not always accurately reflect an individual's true psychological state. In addition, the frequency of regular testing is often limited and cannot dynamically track individual emotional fluctuations, especially in the short term. This may lead to missed opportunities for early intervention. The development of Natural Language Processing (NLP) technology has made it possible to analyze potential psychological problems in social media and intervene in advance. This paper proposes an NLP-based framework to detect depression, anxiety, and suicidality from user-generated text. This work uses fine-tuned Bidirectional Encoder Representations from Transformers (BERT) models to classify each mental health state from different dimensions, then employs an ensemble method to detect a person's mental state comprehensively. The system is designed to provide early identification of mental health risks. Experimental results validate the approach's accuracy and its potential impact on mental health interventions.

1 INTRODUCTION

Nowadays, mental health has become a primary health problem for people around the world. According to "Share of population with mental health disorders" as demonstrated in Figure 1, approximately 13.6% of the global population is suffering from mental health disorders (Global Burden of Disease, 2021). The dataset includes different types of mental health disorders such as depression, anxiety, bipolar, eating disorders, and schizophrenia. In high-income countries and low-income countries, this proportion is even higher. Psychological disorders not only cause patients to suffer extraordinary pain, but in severe cases, they can also threaten the patient's life safety. The statistical result of the World Health Organization shows that more than 700 000 people die by suicide every year, which is one person every 40 seconds (World Health Organization, 2014). This is only the number of deaths from suicide, in fact, the number of people attempting suicide each year is several times that number. There are indications that for each adult who died by suicide, there may have been more than 20 others attempting suicide (World Health Organization, 2014). In traditional methods, psychological issues are often assessed through self-report questionnaires.

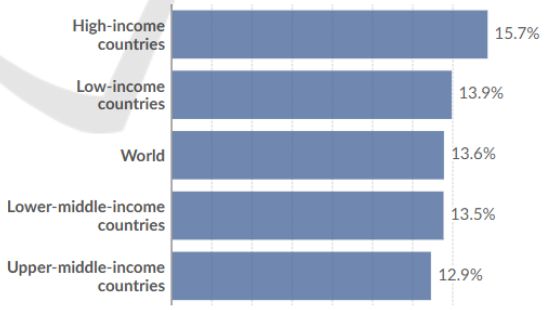

Figure 1: Share of population with mental health disorders (Global Burden of Disease, 2021).

However, this method relies on the individual's perception and ability to express their own psychological state. Many people often struggle to accurately understand or describe their emotions or psychological state when assessing their own psychological state. This subjectivity may prevent the evaluation results from objectively reflecting an individual's true mental health status. Moreover,

Liu, C.
Mental Health Assessment Based on Natural Language Processing.
DOI: 10.5220/0013281600004558

In *Proceedings of the 1st International Conference on Modern Logistics and Supply Chain Management (MLSCM 2024)*, pages 224-228
ISBN: 978-989-758-738-2

many people may intentionally or unintentionally conform to social expectations or their own internal defense mechanisms when filling out questionnaires, which can seriously affect the accuracy of evaluation results. In addition, self-report questionnaires are usually evaluated at specific time points and cannot continuously track changes in an individual's psychological state. In fact, posts on social media often better reflect a person's real-time psychological state. If potential psychological problems can be warned and intervened in a timely manner through social media posts, many mental illnesses will not develop to an irreversible stage. With the development of natural language processing technology, intervening in potential psychological issues through social media is feasible. To further investigate the manifestations of these mental health issues on social media, this work used the Reddit Mental Health Dataset published by Low et al. in 2020 (Low, 2020). This dataset contains posts collected from 28 different subreddits in the Reddit community between 2018 and 2020. This work used a total of 75000 pieces of data from three subreddits, depression, anxiety, and suicide observations in 2020, to train Bidirectional Encoder Representations from Transformers (BERT) models for specific psychological states. The author fine-tuned the BERT model for each situation to learn unique language patterns related to each mental health issue, and employed ensemble learning techniques to combine the outputs of these models for a more powerful and comprehensive assessment.

2 METHOD

2.1 Dataset

The dataset is the Reddit Mental Health Dataset from Zenodo (Low, 2020). Among them, there are 15896 pieces of data on anxiety subreddits, 38033 pieces of data on depression, and 21410 pieces of data on suicidal tendencies. The three subreddits contain a total of approximately 75000 data points, and each subreddit's data includes various kinds of attributes, such as Metadata Attributes which introduce the specific subreddit, author, and date of publication, Text Content Attributes which include the text content of the post, Readability Indices which measure the difficulty and readability of text reading, Sentiment Analysis Attributes which analyze the emotional types of text and Mental Health-related Metrics which provide statistics on the frequency of occurrence of vocabulary related to different types of psychological states.

2.2 Data Preprocessing

In order to adapt the dataset to the training requirements of the BERT model, it is required to process the data (Boukhlif, 2024). Firstly, this work uses the chardet library to detect file encoding and ensure that data can be accurately read and processed. Then, the dataset is loaded into Pandas DataFrame and conducts preliminary missing value evaluation to ensure the integrity of the data. Next, in order to enable the BERT model to process and extract meaningful features more accurately, this work used regular expressions to remove noise from the original text data, such as HTML tags, redundant whitespace, and other irrelevant symbols, completing text cleaning. To facilitate further processing and classification, various sentiment scores and readability indicators stored in string form in the dataset are converted into appropriate numerical types. Then this work performs feature selection. Based on the correlation between features and mental health status detection, the selected features include emotional scores, language usage indicators, and readability indicators. To create classification labels, this work developed a custom function that classifies each record based on predefined thresholds. These label generation functions consider multiple features, allowing for detailed classification of anxious text as "anxious", "potentially anxious" or "non anxious". The label categories for depressive and anxiety texts are similar, divided into "depression", "potential depression", or "non depression". Suicide tendency is divided into "suicidal tendency" and "non suicidal tendency". Finally, this work exported the processed data (including labels) as a CSV file. These annotated datasets can be used for training and fine-tuning BERT models (Chowdhary, 2020).

2.3 Model

This work adopted the BERT architecture and added a classification layer specifically for outputting predictions of corresponding mental health conditions. The basic BERT model, BERT-base-truncated, is initialized with pre-trained weights, and the classification layer is fine-tuned on the training data for each scenario. According to the specific task, the classification layer has been adjusted to match the number of output labels (Devlin, 2018). Each dataset of mental health status is divided into a training set, a validation set, and a testing set, using a hierarchical

partitioning method to ensure consistent label ratios across all subsets. This ensures the model's training and evaluation on balanced data, improving generalization performance. Then this work starts to fine tune the model by using an AdamW optimizer with a learning rate of 1e-5, combined with a linear learning rate scheduler that includes a warm-up phase (Loshchilov, 2017). This prevents overfitting of the model and ensures stable convergence. After fine-tuning, the model is evaluated on the test set and generates a classification report, including metrics such as accuracy, precision (PRE), recall (REC), and F1 score (Novaković, 2017).

2.4 Model Integration

This work integrated three fine-tuned BERT models (used for the classification of depression, anxiety, and suicidal tendencies) together to provide a comprehensive mental health assessment system. The core idea of model integration is to generate a comprehensive judgment of individual mental health by combining the prediction results of various models. Firstly, the input text is pre-processed and converted into a format acceptable to the BERT model. Then, the input text is converted into three independent models for anxiety, depression, and suicidal tendencies, and obtains the predicted probability distribution for each model. For the suicidal tendencies model, since it only has two output labels, this work adds a zero vector to make it consistent with the output format of other models. For each model's output, the author makes separate predictions and obtains their respective classification results. Based on these independent prediction results, the author combines them into a comprehensive assessment that includes anxiety, depression, and suicidal tendencies.

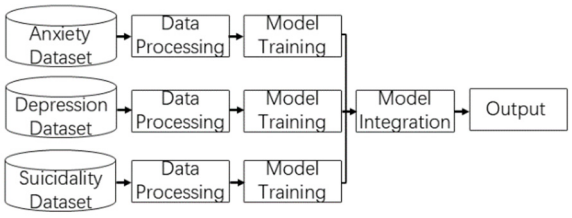

Figure 2: Model architecture (Figure Credit: Original).

After generating independent prediction results, the system will integrate these prediction results into a comprehensive mental health assessment report. This report not only provides classification results for each mental health condition, but also translates these results into actionable recommendations to assist users or medical professionals in making further decisions. Based on the prediction results of each model, the system generates a comprehensive evaluation report. This report includes three main parts: anxiety assessment, depression assessment, and suicidal tendencies assessment. The system provides an overall assessment of the mental health status based on the results of each section and offers corresponding recommendations. The architecture of the entire model is shown in Figure 2.

3 SYSTEM DESIGNN

In the process of anxiety model label generation, this study divides user text into three categories through custom rules: anxiety, potential anxiety, and non-anxiety. The classification criteria combined the text sentiment score and anxiety and negative sentiment traits from the Linguistic Inquiry and Word Count (LIWC) analysis. Here's how it works: Text is marked as anxiety when it has a negative sentiment score greater than 0.5 or a composite sentiment score less than 0 and an anxiety indicator greater than 2 or a negative sentiment indicator greater than 3 in the LIWC analysis. If the text's anxiety or negative sentiment indicators are in the low range (i.e., LIWC anxiety is greater than or equal to 1 or LIWC negative emotion is greater than or equal to 2), the text is marked as potential anxiety. Otherwise, the text will be classified as non-anxiety. For the depression model and the suicidality model, a similar principle was also used for labelling (Ibrahim, 2023).

In order to ensure that the model can effectively handle the text classification task, the input data is loaded and processed in a customized manner. By customizing the TextDataset class, this project uses Huggingface's BertTokenizer to tokenize the input text and convert the results into a tensor format suitable for the model input (Alzahrani, 2021). In addition, this project encodes each piece of text into the format required by the BERT model, including the input ID and attention mask.

4 RESULTS

The training results show that the training loss decreases significantly from 0.861 on the first epoch to 0.162 on the eighth epoch, indicating that the fitting effect of the model on the training set is constantly improving as shown in Table 1. The continuous decline in training loss indicates that the model is continuously optimized during the learning process

and is able to capture patterns in the data over time. The training accuracy increased from 0.566 (56.6%) to 0.942 (94.2%), which indicates a significant improvement in the model's performance on the training set. This result shows that the model is able to accurately distinguish between different classes on the training data. The loss and accuracy trends in the training set are smoothed, as shown in Figure 3.

Table 1: Accuracy in different epochs.

epoch	Train accuracy	Val accuracy
1	0.566576819	0.680100756
2	0.713926325	0.712216625
3	0.787601078	0.743073048
4	0.840880503	0.710957179
5	0.881221923	0.714105793
6	0.919047619	0.715365239
7	0.942138365	0.720403023

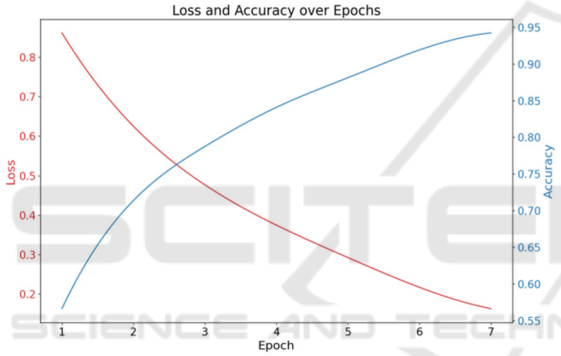

Figure 3: Loss and accuracy over epochs (Figure Credit: Original).

Table 2: Performance on anxiety.

	PRE	REC	F1	Support
Anxiety	0.69	0.86	0.77	657
Potential anxiety	0.68	0.57	0.62	624
Non-anxiety	0.90	0.73	0.80	307
Macro average	0.76	0.72	0.73	1588
Weighted average	0.73	0.72	0.72	1588
Accuracy	0.72			1588

As demonstrated in Table 2, the anxiety detection model had an overall accuracy of 72% on the test set. For anxiety symptoms, the model has a recall rate of 86% and is able to identify most of the anxiety samples well, while the non-anxiety category has an accuracy of 90% but a recall rate of 73%, and there is still room for improvement in identifying potential anxiety.

In the depression model, as shown in Table 3, the recall rate of the model reached 84% for depressive symptoms, which was able to effectively identify most of the samples of depressive symptoms. The recall rate of the potential depression category was 63%, and the performance of the model in this category needs to be further optimized. The accuracy and recall rates for the non-depressive categories were excellent, especially with an accuracy of 93%, indicating that the model was able to identify non-depressive symptoms well.

Table 3: Performance on depression.

	PRE	REC	F1	Support
Depression	0.71	0.84	0.77	1680
Potential depression	0.74	0.63	0.68	1598
Non-depression	0.93	0.81	0.86	524
Macro average	0.79	0.76	0.77	3802
Weighted average	0.75	0.75	0.75	3802
Accuracy	0.75			3802

In the suicidality detection task as displayed in Table 4, the model achieved 89% accuracy. The suicidality category has a 96% recall rate, indicating that the model is able to accurately identify the majority of samples with suicidal tendencies, while having an accuracy of 89%, showing a low false positive rate. The accuracy of the non-suicidal category was also 90%. F1 scores of 0.92 and 0.81 indicate that the model's performance is relatively balanced across these two categories.

Table 4: Performance on suicidality.

	PRE	REC	F1	Support
Suicidality	0.89	0.96	0.92	1466
Non-suicidality	0.90	0.74	0.81	674
Macro average	0.90	0.85	0.87	2140
Weighted average	0.89	0.89	0.89	2140
Accuracy	0.89			2140

Figure 4 illustrates the probability distribution of the anxiety, depression, and suicidality model for one of the input texts in the dataset. The three subgraphs correspond to the predicted probability distributions

of the anxiety model, the depression model, and the suicidality model, respectively. Within each subgraph, the anxiety and depression models have three categories of probability distributions: significant symptoms, latent symptoms, and no symptoms. Due to its dichotomous problem, the suicidality model only includes two categories: significant suicidality and no suicidal tendencies. The input text in the example is an actual user-generated post that expresses the user's emotional distress, including self-identified depression and suicidal tendencies. The models output the probability distributions for different classes for each model, with a negative input from the dataset.

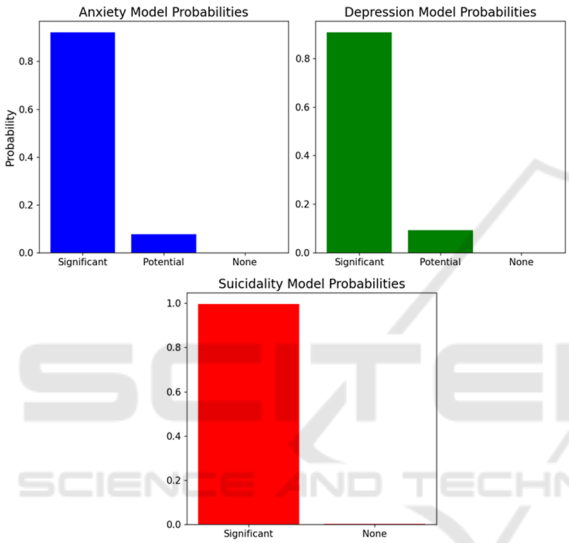

Figure 4: Model probability distribution (Figure Credit: Original).

5 CONCLUSIONS

This project uses natural language processing (NLP) technology and fine-tuned BERT models to detect mental health issues such as depression, anxiety, and suicidal tendencies from social media user-generated text. By using the AdamW optimizer and the linear learning rate scheduler (with warm-up stage), the model achieves good performance. The anxiety, depression, and suicidality models were each trained on a pre-processed custom dataset that included important linguistic features such as sentiment scores, readability index, and generated labels based on specific thresholds, such as "anxiety", "depression" and "suicidality" to ensure that the model was able to accurately classify the text.

During the training process, the performance of each model continued to improve. For example, the training accuracy of the anxiety detection model exceeded 94% in the eighth training round. Still, there is room for improvement in the model in identifying underlying anxiety and depressive symptoms. Overall, the system provides a comprehensive mental health assessment report by integrating the prediction results of multiple models, which provides data support for early intervention and decision-making. This method verifies the effectiveness and potential of NLP-based mental health testing in practical applications.

REFERENCES

Alzahrani, E., & Jololian, L. 2021. How different text-preprocessing techniques using the bert model affect the gender profiling of authors. *arXiv preprint arXiv:2109.13890*.

Boukhlif, M., Hanine, M., Kharmoum, N., Noriega, A. R., Obeso, D. G., & Ashraf, I. 2024. Natural Language Processing-based Software Testing: A Systematic Literature Review. *IEEE Access,* 1-18.

Chowdhary, K., & Chowdhary, K. R. 2020. Natural language processing. *Fundamentals of artificial intelligence*, 603-649.

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

Global Burden of Disease, 2021. Share of population with mental health disorders. URL: https://ourworldindata.org/grapher/share-with-mental-and-substance-disorders. Last Accessed: 2024/09/16

Ibrahim, M. A., Ismail, N. H., Kamarudin, N. S., Nafis, N. S. M., & Nasir, A. F. A. 2023, August. Identifying PTSD Symptoms Using Machine Learning Techniques on Social Media. In *2023 IEEE 8th International Conference on Software Engineering and Computer Systems*. 392-395.

Loshchilov, I. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Low, D. M., Rumker, L., Talkar, T., Torous, J., Cecchi, G., & Ghosh, S. S. 2020. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *Journal of medical Internet research*, *22*(10), e22635.

Novaković, J. D., Veljović, A., Ilić, S. S., Papić, Ž., & Tomović, M. 2017. Evaluation of classification models in machine learning. *Theory and Applications of Mathematics & Computer Science*, *7*(1), 39.

World Health Organization. 2024. Suicide data. URL: https://www.who.int/teams/mental-health-and-substance-use/data-research/suicide-data. Last Accessed: 2024/09/16