Sentiment Analysis for Mental Health Detection

Huu Tien Nguyen
Department of Computer Science
University of Exeter, Exeter, UK

***Abstract*—Mental health issues are becoming a serious public health issue in many nations. To support students, many research studies and efforts have been made to analyze or build depression detectors that are able to filter out which students intend to be depressed or not. In this work, I conducted a sentiment analysis using a BERT pre-trained model as a feature extractor while incorporating additional fully connected layers for classification. The evaluation metrics using the confusion matrix framework indicated the feasibility of the project with positive results. Further experiments are needed to find suitable hyperparameters and ensure that the model effectively captures depressive cases while minimizing misclassification.**

I. Introduction

Student mental health is becoming a serious challenge in higher education in recent years. The 2023 annual Cibyl Mental Health Study [1] found that 39% of students in the UK experienced a decline in mental health after starting university. This not only significantly influences students' lives but also affects financial burdens on society due to more money would need to be allocated to healthcare services. According to an NHS report [2] in 2022, £12 billion was spent on mental health services in the UK, accounting for around 8% of their total budget, and this spending has increased in recent years. In practice, educational organizations have been implementing various solutions to support students such as providing comprehensive mental health services. However, the challenge is that they do not have the ability to monitor students continuously, and mental health issues often lack obvious symptoms. This highlights the importance of developing tools that can detect which students are facing problems and when, at an early stage and on a large scale, so that we can provide timely support.

Nowadays, the rise of social media platforms such as Facebook or Twitter has provided a great opportunity to observe and understand human emotions and behaviors on a large scale. This is because these platforms function as a digital diary for many users and people therefore can share their thoughts, their feelings and theirs experiences every single day. If we can learn writing patterns from this information, we can develop an early depression detection system for mental health problems. The concept of Depression Detector is depicted in figure 1.

To demonstrate the ability of AI to tackle this issue, I selected a dataset from Kaggle named Students Anxiety and Depression Dataset, which includes Facebook comments and posts from undergraduate students. The unstructured data includes 6,982 records with two fields: the first one contains texts and the other contains label field that indicates whether the corresponding text signifies anxiety/depression or not. In this work, I used the Natural Language Processing (NLP) pre-trained model named Bidirectional Encoder Representations from Transformers (BERT) to build a depression detector by training the model on the dataset mentioned above.

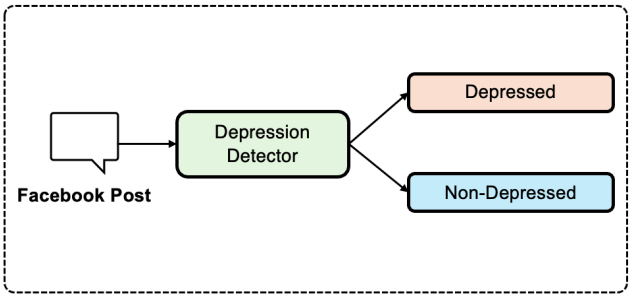

Fig. 1. Concept of Depression Detector.

II. Literature Review

Sentiment analysis of social media data is a promising approach to solving this problem. Many studies and efforts have discussed depression detection using various algorithms. In 2018, Orabi et al. [3] proposed adopting supervised machine learning, utilizing the most effective deep neural architecture from two of the most popular deep learning approaches in the field of NLP: Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), respectively. In 2019, Nafiz Al Asad et al. [4] conducted research to build a depression detection system using Support Vector Machine (SVM) and the Naïve Bayes algorithm to potentially detect depression.

Another study conducted by Obagbuwa et al. [5] in 2023 applied traditional Machine Learning Classification techniques, including four models: Extreme Gradient Boosting (XGB) Classifier, Logistic Regression, Random Forest, and Support Vector Machines (SVM). The comparison between these models provided an interesting insight: SVM and Logistic Regression achieved the most accurate results, with Logistic Regression demonstrating slightly higher accuracy compared to SVM. On the other hand, Logistic Regression models required less execution time.

III. Methodology

This section mentions the model architecture using BERT model, which is a state-of-the-art transformer-based machine learning model in natural language processing (NLP), and then describes how this model works for the Sentiment Analysis

Problem. The BERT versioned in this paper is the BERT base uncased which is suitable for a restricted individual computer resource and the small dataset mentioned above.

A. Data Preprocessing

When observing the dataset, I identified an imbalance where the number of samples labeled as depressed was much smaller than the non-depressed class. Therefore, I applied data augmentation using Augmenting Contextual Word Embedding. This method generates additional depressed samples by inserting words into the original text with contextually similar words based on BERT embeddings. This helps preserve the original sentence structure while introducing diversity into the training data. After this step, I preprocessed the text by applying tokenization using BERT's WordPiece tokenizer. This method converted each text sequence into input token IDs, attention masks, and segment IDs, ensuring compatibility with BERT's input format.

B. Model Architecture

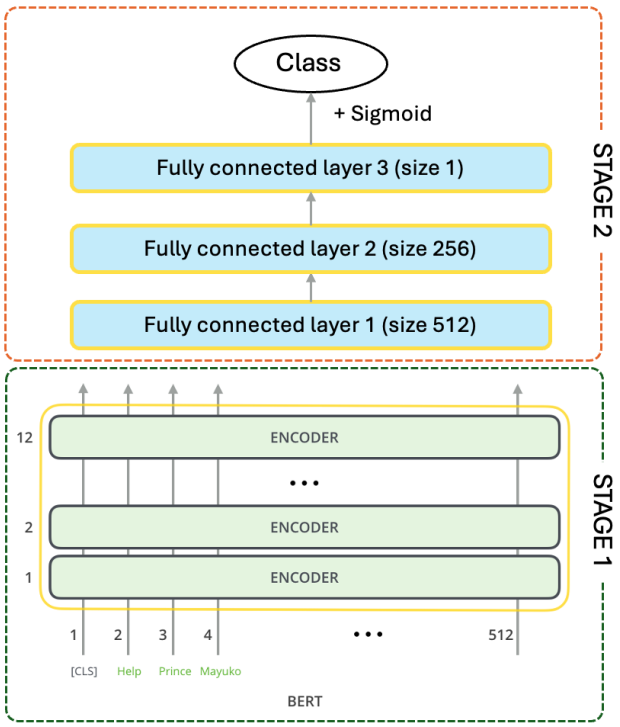

Fig. 2. Model architecture.

Instead of fine-tuning the entire BERT model, I used a transfer learning approach where the pre-trained BERT model serves as a feature extractor (this means that I froze/did not retrain the BERT layers). After that, I added two fully connected layers on top for classification.

Figure 2 illustrates the model architecture, which combines two stages. At the first stage, I kept the BERT encoder to process the input text and extract a contextualized embedding vector, using the [CLS] token as a sentence representation. The BERT base architecture includes a stack of 12 encoder layers from the transformer model and a hidden size of 768 dimensions. At the second stage, I fine-tuned BERT base by adding two fully connected layers. The first dense layer was followed by a dropout layer (30%) to reduce overfitting before applying the ReLU activation function to add nonlinearity. The final layer uses a sigmoid activation function to convert the output into probabilities and determine the class. The embedding vector obtained from the first stage is passed through these layers to perform binary classification.

C. Model Evaluation

Regarding the evaluation process, I fed the model with an evaluation dataset that contained texts and the corresponding labels, then compared the predicted labels with the true labels to compute performance metrics. Although accuracy is a commonly used metric, it is not a reliable measure in the presence of class imbalance. To address this issue, I used the confusion matrix framework, as shown in Figure 3. In this framework:

- True Positives refer to correctly predicted depressed cases.
- False Positives refer to non-depressed cases incorrectly classified as depressed.
- False Negatives refer to depressed cases incorrectly classified as non-depressed.
- True Negatives refer to correctly predicted non-depressed cases.

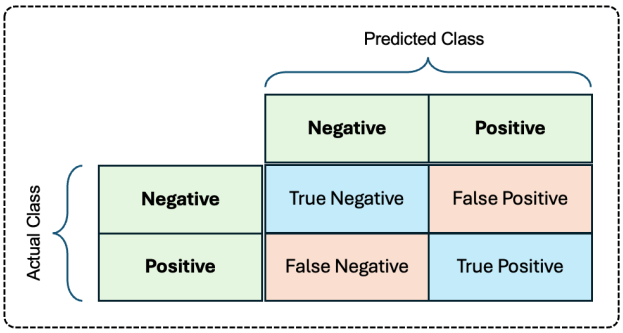

Fig. 3. Confusion matrix framework.

After that, I calculated Precision, Recall, and F1-score. Among these, the F1-score is prioritized because it balances precision and recall, ensuring that both false negatives and false positives are considered carefully.

$$Precision = \frac{true\ positive}{true\ positive + false\ positive} \tag{1}$$

$$Recall = \frac{true\ positive}{true\ positive + false\ negative} \tag{2}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

IV. Result

To conduct sentiment analysis, I split the dataset into training, validation, and test datasets. I trained and evaluated the model using supervised learning. Figure 4 shows the confusion matrix, and from that, we can see that my model correctly detects 577 posts as non-depressed out of 618 total posts and also correctly classifies 250 posts as depressed out of 303 Facebook posts.

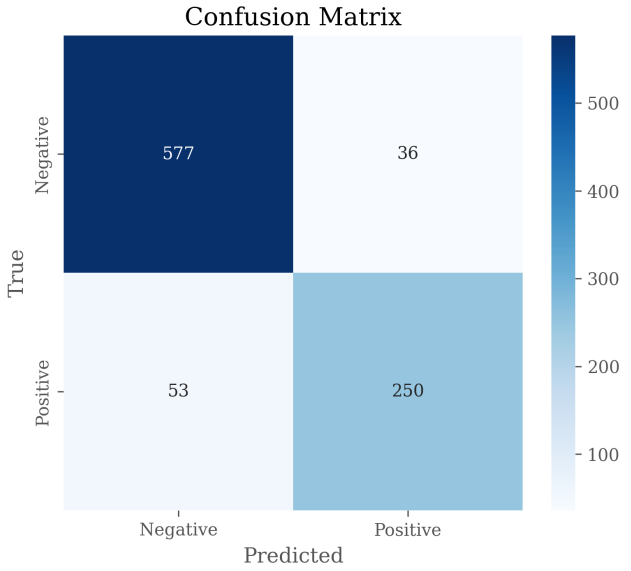

Fig. 4. Example of a figure caption.

Table I shows the "Recall," "Precision," and "F1-scores" for each class and the total model performance. As we can see, the model achieves an overall F1-score of 89%.

TABLE I
Model performance

Sentiment Class	**Evaluation Metrics**			
	Precision	*Recall*	*F1 − score*	*Support*
Non-Depressed	0.92	0.94	0.93	618
Depressed	0.87	0.83	0.85	303
Weight average	0.89	0.88	0.89	916

V. Conclusion

In conclusion, this paper provides a robust foundation for understanding the BERT model and demonstrates the feasibility of using this model to tackle real-world problems. However, further improvements are necessary to enhance its robustness. Future work should focus on exploring alternative data augmentation techniques and hyperparameter tuning to improve classification performance. Beside, expanding the dataset could enhance the model's ability to capture the writing patterns associated with depression.

References

[1] Cibyl, "Mental health launch 2023," 2023, accessed: 17 Mar. 2025. [Online]. Available: https://www.cibyl.com/cibyl-insights/mental-health-launch-2023

[2] H. Treasury, "Public expenditure statistical analyses (pesa) 2022," 2022, accessed: 17 Mar. 2025. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1091951/E02754802_PESA_2022_elay.pdf

[3] A. H. Orabi, P. Buddhitha, M. H. Orabi, and D. Inkpen, "Deep learning for depression detection of twitter users," in *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, 2018, pp. 88–97.

[4] N. A. Asad, M. A. Mahmud Pranto, S. Afreen, and M. M. Islam, "Depression detection by analyzing social media posts of user," in *2019 IEEE International Conference on Signal Processing, Information, Communication Systems (SPICSCON)*, 2019, pp. 13–17.

[5] I. C. Obagbuwa, S. Danster, and O. C. Chibaya, "Supervised machine learning models for depression sentiment analysis," *Frontiers in Artificial Intelligence*, vol. 6, p. 1230649, 2023.