Fine-Tuning Large Language Models For Depression And Anxiety Detection On Twitter

Submitted By

MD Siyam Bhuiyan

Section: A

ID: 211-35-702

Department of Software Engineering

Daffodil International University

Supervised By

Nadira Islam

Senior Lecture

Department of Software Engineering

Daffodil International University

Thesis submitted in fulfillment of the requirements for the award of the degree of

Bachelor of Science

Fall - 2024

APPROVAL

This thesis titled on "**Fine-Tuning Large Language Models for Depression and Anxiety Detection on Twitter**", submitted by **MD Siyam Bhuiyan (ID: 211-35-702)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Chairman

Dr. Imran Mahmud
Associate Professor & Head
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

Nuruzzaman Faruqi
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

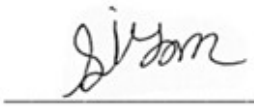
Md. Rajib Mia
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Md. Fazle Munim
Associate Director & Vice President
Government & Public Sector
Ernst & young (EY)

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.



(Student's Signature)

Full Name : MD Siyam Bhuiyan

ID Number : 211-35-702

Date : Januray 2025

SUPERVISOR's DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science.

(Supervisor's Signature)

Full Name : Nadira Islam

Position : Senior Lecturer

Date : January 2025

ACKNOWLEDGEMENT

First and foremost, I am deeply grateful to Almighty Allah, who has given me the strength, wisdom, and perseverance to complete this research. All through my academic journey, I am grateful for the unconditional love, support, and encouragement of my parents. It is always their belief in me that has motivated and inspired me the most.

I would like to thank my supervisor, Senior Lecturer Ms. Nadira Islam for his valuable advice, support, and guidance throughout the research. Her knowledge and insight have vastly affected this work. The departmental head, Dr. Imran Mahmud, is also highly appreciated for his support, guidance, and valuable comments which helped me to successfully complete my journey. Finally, I would like to thank all my friends, colleagues, and all those who helped and encouraged me during this process.

Abstract

Social media platforms, such as Twitter, have become influential channels for capturing expressions of emotion and thought, serving as proxies for users' mental states. With the growing prevalence of anxiety and depression in today's society, there is an increasing need for scalable and effective methods to detect mental health issues. Traditional approaches, such as surveys and clinical assessments, are limited in their capability to provide real-time, large-scale population coverage. Recent advancements in large language models (LLMs) offer the opportunity to analyze textual data and automate mental health detection on a broader scale.

This study leverages pre-trained transformer models, including DistilBERT, MentalBERT, and BERTweet, to detect anxiety and depression from tweets. The primary objective is to develop an automated system capable of identifying signs of mental health concerns by analyzing linguistic patterns, sentiment markers, and behavioral cues in social media posts. To achieve this, a labeled dataset of tweets was processed through cleaning, tokenization, and splitting into training, validation, and test sets. Fine-tuning techniques were applied to optimize model performance, addressing challenges such as class imbalance and scalability.

Among the models, BERTweet demonstrated superior performance overall, excelling in precision, accuracy, and a balanced detection of mental health indicators. MentalBERT performed notably in recall, highlighting its strength in identifying instances of anxiety and depression with minimal false negatives. The lightweight DistilBERT model offered a computationally efficient solution while maintaining strong overall performance, making it suitable for real-world applications.

This research has therefore shown that fine-tuned LLMs are effective in the detection of anxiety and depression using Twitter data. This work describes a scalable, efficient, on-time methodology for the identification of at-risk persons through automation of mental health detection. The contribution of this work contributes to the bigger effort of integrating artificial intelligence into mental health care subsequent to enable early intervention and support. Further research can expand this framework to include multilingual datasets and integrate multimodal features such as images and user behavior to develop a more holistic mental health monitoring system.

Contents

List of Figures

Chapter 1

Introduction

1.1 Background

User mental states are represented via social networking sites like Twitter, which also record the users thoughts and emotions. This transformation extends to fields like public health epidemiology, where analyzing social media has become a valuable tool for understanding mental health trends. Anxiety and depression remain the most widespread mental health issues globally, with their prevalence growing over the last decade.(Organization,) Social media,especially Twitter,offers a unique glimpse into individuals mental states,and researchers have found that analyzing tweets can provide real-time insights into shifts in mental well-being.(Choudhury, Counts, & Horvitz,)The integration of artificial intelligence, especially natural language processing (NLP),has opened doors to processing massive amounts of unstructured text. These advancements make it feasible to create systems that monitor mental health at scale, offering a practical approach to identifying and addressing mental health concerns through digital footprints.

1.2 Research Motivation

Clinical interviews and self-report surveys, which are considered to be the traditional methods of assessing mental health, have been seen as being practical, scalable, and up to date (Kessler & Bromet,). However, Automated systems that are based on large language models (LLMs) are able to swiftly, scalable, and non-invasively identify mental health concerns. This is possible because of the millions of tweets that are produced every day.This effort aims to construct a system that can quickly detect anxiety and sadness using textual information that is readily

available to the public. This will be accomplished by overcoming the approaches that are now in use.

1.3 Problem Statement

It is still challenging to use data from social media platforms to diagnose mental health concerns such as anxiety and depression, despite the fact that large-scale language models (LLMs) have improved. Text data from social media platforms is typically unorganized, which makes it significantly more challenging to evaluate. In addition, these activities make use of datasets that have unequal class distributions, which results in predictions and models that are less reliable. TThere is also a trade-off between the precision of the model and the speed of the computation. Although it is challenging, striking a balance between these two qualities is essential. The absence of mental health databases that are tied to social media is a key cause for concern. Because of this, the process of testing and producing models is made more complicated. As a result of these problems, there is a pressing requirement for models that are capable of properly and rapidly identifying feelings of anxiety and depression by utilizing this data.

1.4 Research Questions

This research seeks important answers.:

1. What are the most effective ways to fine-tune large language models for detecting anxiety and depression using Twitter data?
2. To get an accurate binary categorization of mental health conditions, what kinds of pre-processing and hyper-parameter optimization are required?
3. How does incorporating class weights affect a model's ability to accurately identify minority classes in datasets with imbalanced distributions?

1.5 Research Objectives

Research objectives are as follows:

- To fine-tune large language models (LLMs) for identifying signs of anxiety and depression from Twitter data.
- To evaluate the selected model's performance using metrics such as accuracy, precision, recall, and F1-score.
- Eliminating datasets(Surin,) imbalances is accomplished by the utilization of class weights inside the loss function.
- To help raise awareness of the use of LLM to detect mental health issues on social media platforms.

1.6 Research Scope

This study emphasizes classifying mental health awareness based on English-language tweets. The model will be pre-trained and then evaluated on an annotated dataset composed of tweets indicative of anxious, depressed, and neutral states of a person's mind. It does not develop research on either multilingual datasets or multimodal approaches with images and user behavior included. Instead, it establishes a benchmark for applications that need only textual input to detect signs of mental health conditions.

1.7 Thesis Organization

The thesis is structured into the following chapters:

- **Chapter 1**: Introduces the study's background, motivation, problem statement, and objectives.
- **Chapter 2**: The purpose of this article is to provide a summary of studies on the utilization of models to identify and explain mental health disorders by providing a literature review.
- **Chapter 3**: Detailed explanations of our study's preprocessing, model adjustment, and evaluation procedures are provided here.
- **Chapter 4**: Regarding the success of our model, we discuss the results.
- **Chapter 5**: Concludes with a discussion on findings, limitations, and recommendations for future research.

Chapter 2

Literature Review

The detection of mental health issues on Twitter and other websites has been improved because to the usage of large language models (LLMs), which have enabled user-generated text analysis. Twitter's short length, user-friendliness, and ability to support multiple languages make it an effective tool for identifying patterns related to mental health. Recently efforts have been emphasized developing datasets, applying LLMs, and systematically reviewing their impacts, highlighting both progress and ongoing challenges in this field.

One major focus has been the creation of high-quality datasets to support effective mental health analysis. Skianis et al.((Skianis, Pavlopoulos, & Dogruöz,)) introduced a multilingual dataset that assesses mental health conditions and their severity across platforms like Twitter. This dataset facilitates cross-platform comparisons while promoting inclusivity by capturing linguistic diversity and fostering cross-linguistic insights. Similarly, Bucur((Bucur,)) explored LLM-generated synthetic data as a means to supplement existing datasets, particularly for detecting major depression, demonstrating the value of synthetic methods in dataset enhancement.

LLMs have shown immense potential for mental health detection, with research increasingly exploring linguistic markers indicative of psychological states. For instance, Xu et al.(Xu, Yao, Dong, Gabriel, & Yu,) fine-tuned an LLM to predict mental health conditions, showing how such models can enhance accuracy using annotated Twitter data. Shah et al.(Shah, Gillani, Baig, & Saleem,) adopted a similar fine-tuning approach to detect depression in tweets, focusing on nuanced text features specific to mental health themes.

Systematic reviews further illuminate the broader field. Guo et al.(Guo, Lai, Thygesen, Farrington, & Keen,) conducted an in-depth review of LLM applications in mental health, highlighting both accomplishments and challenges. While they emphasized the promise of LLMs in identifying disorders like depression and suicidality, their findings also pointed to

ethical and methodological limitations that need addressing.

There are some areas of mental health that are the subject of certain study. For example, Thamrin et al.(Thamrin & Chen,) developed model For the purpose of identifying bipolar disorder in Twitter data, word embeddings were improved for many domains. The application of clinical and biological knowledge resulted in an improvement of these. Similarly, Alhamed et al.(Alhamed, Ive, & Specia,) examined the effects of a sad diagnosis on the language used before and after the diagnosis. The alterations in mental health were better understood as a result of this.

Despite the fact that there have been improvements made, there are still many problems. There are significant ethical problems around the misuse of diagnostic tools and the safeguarding of data. There are many different cultures and languages represented in the data that Twitter collects, which makes models less universal. It is necessary to take a holistic approach that places an emphasis on the ethical and appropriate utilization of LLM in order to address these concerns.

The development of Large Language Models (LLMs) has revolutionized the analysis of social media data for mental health applications, particularly in detecting depression. By capitalizing on the unique characteristics of platforms like Twitter, real-time accessibility, and diverse linguistic content—researchers have achieved significant advancements in understanding and identifying mental health patterns. (Shah, Gillani, Baig, Saleem, & Siddiqui,) explores the application of fine-tuned GPT-3.5 Turbo and LLaMA2-7B models for detecting depression on social media. Using a robust dataset from prior research, the models were trained to identify nuanced linguistic patterns indicative of depression, achieving a high accuracy of 96%. This work highlights the effectiveness of adapting large-scale language models to mental health diagnostics, outperforming traditional methods. It emphasizes the potential for early depression detection through real-world applications, though ethical considerations such as privacy and bias remain challenges. Future directions include expanding to multilingual datasets and improving model generalizability across demographics.

With the increasing reliance on social media as a platform for self-expression, opportunities have emerged to analyze user-generated content for early signs of mental health conditions like depression. Social platforms such as Reddit and Twitter, characterized by their openness and frequent updates, provide valuable data for building automated detection systems. Recognizing this potential, (Tavchioski, Robnik-Šikonja, & Pollak,) explored the use of transformer-based

models and ensemble methods for detecting depressive patterns in social media posts. The authors fine-tuned widely-used models, including BERT, RoBERTa, BERTweet, and mental-BERT, while also constructing ensemble configurations to enhance model performance. Tested on datasets from Reddit and Twitter, the results demonstrated that ensemble approaches outperformed individual transformer models, highlighting the benefits of combining classifiers for improved accuracy. Transfer learning was also employed across datasets, showcasing the adaptability of these models in diverse contexts. This work underscores the promise of transformer-based ensembles in advancing mental health diagnostics through social media analysis. Future directions may include addressing ethical considerations and expanding research to multilingual datasets and other social platforms to ensure inclusivity and scalability.

In summary, LLMs are transformative tools for detecting mental health disorders through social media analysis. While considerable progress has been made in dataset development, model refinement, and systematic reviews, future work must tackle the ethical, cultural, and technical challenges that persist. With sustained research and responsible practices, Mental health monitoring can be revolutionized by LLM, which can also assist individuals in receiving assistance at an earlier stage.

Chapter 3

Methodology

3.1 Description of Dataset

The methodology for this research follows a systematic approach that integrates experimental, descriptive, and analytical techniques to develop and assess fine-tuned LLM models for detecting anxiety and depression using Twitter data. The process is carefully structured to ensure reproducibility, guiding readers through the following key steps:

- **Data Preparation:** Collecting and organizing relevant Twitter data to serve as the foundation for analysis.
- **Preprocessing:** Cleaning and refining the data to remove noise and ensure suitability for model input.
- **Model Fine-Tuning:** Adjusting LLMs to enhance their performance in identifying mental health conditions based on the specific dataset.
- **Evaluation:** Measuring the model's performance using metrics such as accuracy, precision, recall, and F1-score.
- **Validation:** Ensuring the robustness and reliability of the model through additional tests and cross-validation techniques. This structured methodology ensures transparency and provides a replicable framework for future studies in this area.

3.2 Preprocessing Data

The datasets used for this research was sourced from Kaggle, a widely recognized platform for datasets and data science competitions. It features tweets labeled with binary indicators: a label of 1 signifies signs of anxiety or depression, while 0 indicates otherwise. This publicly available dataset is titled ”combined Depression Twitter Dataset Feature Extraction” on Kaggle.(Surin,) Among the three source files in the dataset, only mental-health.csv was utilized for this study. The data was then split to ensure unbiased performance evaluation: 80% was allocated for training, 10% for validation, and 10% for testing. This division helps maintain a balanced approach to training and evaluating the model's accuracy and reliability.

3.2.1 Preprocessing Training Data

The process of preprocessing is an essential step in the transformation of raw tweets into code that can be read by machines. Taking the following steps:

- **Cleaning Text:** In an effort to make the process of data analysis more straightforward, URLs, references, hashtags, and odd characters were eliminated.
- **Data Tokenization:** For the purpose of tokenizing the dataset, DistilBERT was utilized. Through the use of this model, text is converted into numeric input data of the Transformer model. Every single token entry is stuffed with 128 tokens in order to guarantee that the dataset is consistent.
- **Handling Class Imbalance:** During the process of training the model, class weights were incorporated into the loss function in order to correct the mismatch between the datasets. Other studies suggest that the major class and the minority class should be given equal weight, which is something that this modification made it possible for the model to do. (Garcia & Wong,).

3.2.2 Fine-Tuning Model

MODEL	DistilBERT
DESCRIPTION	A lightweight, faster version of BERT with 97% of its performance.
CONTEXT WINDOW	512 tokens
TRAINING DATA	English Wikipedia and BookCorpus.

This study employed large language models (LLMs) such as DistilBERT(Sanh, Debut, Chaumond, & Wolf,), which is a lightweight version of BERT transformer model optimized for efficient text classification. The fine-tuning process adapted the pre-trained DistilBERT model to perform a binary classification task, targeting the detection of anxiety and depression. Key steps in the process included:

- **Optimizations of Hyperparameter:**
- For optimal Gradient updates Learning rate was set to 2e-5
- For efficient Gpu Utilization Batch size was set to 16
- Epochs were carefully increased and set to 5
- **Loss Function:** The **cross-entropy loss function** was used for binary classification, defined as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} w_{y_i} \log \hat{p}_{y_i}$$

Where:

- y_i: sample i with true class label .
- N: Total samples number i.
- $\hat{p}_{y_i}$: True class of predicted probability y_i.
- w_{y_i}: weight added to the classes y_i that solves imbalance.
- **Training:** The Hugging Face Trainer API was used to streamline the fine-tuning process by integrating preprocessing, training, and evaluation into a seamless workflow.(Wolf, Debut, Sanh, & Chaumond,)

MODEL	MentalBERT
DESCRIPTION	A BERT-based model fine-tuned for mental health-related text analysis.
CONTEXT WINDOW	512 tokens
TRAINING DATA	Mental health forums and clinical datasets.

The study utilized LLM Models such as MentalBERT(Ji et al.,), a transformer model fine-tuned for analyzing mental health-related text. Fine-tuning involved training the pre-trained MentalBERT(Ji et al.,) is a pre-trained model which was trained for fine-tuning for a binary classification task to distinguish between ”Signs of Anxiety/Depression” and ”No Signs.” Key steps included:

- **Optimizations of Hyperparameter:** The Same DistilBERT Optimization were used for this model also
- **Loss Function:** The **cross-entropy loss function** was used for binary classification, defined as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(\hat{p}_{y_i}) + (1 - y_i) \cdot \log(1 - \hat{p}_{y_i})]$$

- **Training:** TThe Hugging Face `Trainer` API was used to fine-tune MentalBERT by preprocessing, training, evaluation, logging to be flawless. The training pipeline included:
  - Tokenization: Text was tokenized using the MentalBERT tokenizer with a maximum sequence length of 128 tokens.
  - Dynamic Padding: The ‘DataCollatorWithPadding‘ ensured that batches were padded dynamically.
  - Evaluation: The model was evaluated at the end of every epoch using the validation dataset.
  - Logging: Training and evaluation logs were saved for further analysis, and the best model was automatically saved at the end of training.

MODEL	BERTweet
DESCRIPTION	A transformer model optimized for processing tweets and social media text.
CONTEXT WINDOW	512 tokens
TRAINING DATA	850M English tweets from 2012 to early 2021.

The study utilized LLM Models such as **BERTweet**(D. Q. Nguyen, Vu, & Nguyen,), a transformer model fine-tuned for analyzing social media text, particularly tweets. Fine-tuning involved training the pre-trained BERTweet(D. Q. Nguyen et al.,) model on a binary classification task to distinguish between *"Signs of Anxiety/Depression"* and *"No Signs."* The process included the following key steps:

- **Optimizations of Hyperparameter:** The Same DistilBERT Optimization were used for this model also. additionally Weight decay was added.
  - **Weight decay:** for regularization to prevent overfitting it was Set to 0.01.
- **Loss Function:** The **cross-entropy loss function** was used for binary classification, defined as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot \log(\hat{p}_{y_i}) + (1 - y_i) \cdot \log(1 - \hat{p}_{y_i})\right]$$

  Where:
- **Training:** The Hugging Face `Trainer` API was used to fine-tune BERTweet, which streamlined the training process by:
  - **Tokenization:** Text data was tokenized using the `BERTweet tokenizer` (`vinai/bertweet-base`) with a maximum sequence length of 128 tokens, ensuring consistency across inputs.
  - **Dynamic Padding:** The `DataCollatorWithPadding` dynamically padded inputs during training and evaluation for optimized memory usage.
  - **Evaluation:** The model was evaluated at the end of every epoch using the validation dataset, tracking metrics such as accuracy, precision, recall, and F1-score.
  - **Logging and Checkpointing:** Training logs were saved, and the best model was automatically stored at the end of training.

- **Evaluation and Visualization:**
  - **Performance Metrics:** The model's performance on the test dataset was evaluated, calculating metrics such as accuracy, precision, recall, and F1-score.
  - **Confusion Matrix:** A confusion matrix was generated to visualize classification results, distinguishing between *"No Signs"* and *"Signs of Anxiety/Depression."*
  - **Loss Curves:** Training and validation loss curves were plotted to assess model convergence and identify overfitting or underfitting.

3.3 Evaluation Metrics

Model performance was evaluated using standard metrics:

Accuracy

The proportion of correctly classified instances out of the total instances. The accuracy metric is given by:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Where:

- TP: True Positives.
- TN: True Negatives.
- FP: False Positives.
- FN: False Negatives.

Precision

The proportion of true positive predictions among all positive predictions. The precision metric is given by:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall

The ability of the model to detect all actual positive instances. The recall metric, also known as sensitivity, is given by:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1-Score

The harmonic mean of precision and recall, providing a balanced measure of the model's performance. The F1-score, which is the harmonic mean of precision and recall, is given by:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Confusion Matrix:

Visualized the distribution of true positive, true negative, false positive, and false negative predictions. The confusion matrix for the classification model is presented below:

Actual \Predicted	Positive	Negative	Total
Positive	80	20	100
Negative	10	90	100
Total	90	110	200

Metrics Derived from the Confusion Matrix

Using the confusion matrix values, the following metrics are calculated:

- **Accuracy:**

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total}} = \frac{80 + 90}{200} = 85\%$$

- **Precision:**

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{80}{80 + 10} = 88.9\%$$

- **Recall:**

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{80}{80 + 20} = 80\%$$

- **F1-Score:**

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \cdot \frac{88.9 \cdot 80}{88.9 + 80} \approx 84.2\%$$

3.4 Validation

The trained model underwent evaluation on a validation set during training and was later tested on a separate hold-out test set. Several measures were implemented to ensure robustness and reliable performance:

- **Monitoring Validation Loss:** During training, validation loss was tracked to identify potential overfitting, ensuring the model maintained generalization capability.
- **Confusion Matrix Analysis** A confusion matrix was utilized to assess the model's performance in detecting mental health indicators. This tool provided insights into the model's strengths and weaknesses by breaking down true positives, false positives, true negatives, and false negatives.

3.5 Reproducibility

In order to do the research, open-source tools such as Hugging Face Transformer and Python are utilized.(Wolf et al.,) e.g. Scikit-learn and Pytorch. The methodology was carefully designed with reproducibility in mind. To ensure other researchers can replicate the results, all preprocessing scripts, training configurations, and evaluation metrics have been shared. This transparency not only promotes collaboration but also strengthens the reliability and accessibility of the research findings.

Chapter 4

Results

4.0.1 Training Performance

The DistilBERT(Sanh et al.,) model had very good training with great efficiency in processing and accuracy in its performances across the different iterations. The performance improves further because of the MentalBERT(Ji et al.,) model specialized for mental health-related texts. Five-epoch training over both models allowed for good convergence, but it is visible that MentalBERT (Ji et al.,)achieved slightly better stability in the training loss.

4.0.2 Validation Performance

During validation, DistilBERT(Sanh et al.,) demonstrated strong generalization capabilities with an accuracy of **96.21%**, reflecting its reliability in the task. MentalBERT(Ji et al.,), however, slightly outperformed it, MentalBERT had accuracy of **96.33%**, An emphasis was placed on the fact that it contributes to the construction of complex narratives around mental health. Both of these models have some degree of validity loss, but MentalBERT(Ji et al.,) showed a marginal advantage in recall, indicating better sensitivity in identifying relevant mental health indicators.

4.0.3 Quantitative Analysis

Metric	DistilBERT (%)	MentalBERT (%)	BERTweet (%)	Comparison
Accuracy	96.21	96.32	**96.93**	BERTweet highest
Precision	96.46	95.47	**96.97**	BERTweet highest
Recall	95.89	**97.19**	96.83	MentalBERT highest
F1-Score	96.17	96.32	**96.90**	BERTweet highest
Evaluation Loss	0.1218	0.1279	**0.1047**	BERTweet lowest

Table 4.1: DistilBERT, MentalBERT and BERTweet Models Performance for Comparison

4.0.4 Tracking Results

- **Accuracy**: It is estimated that these models are accurate 96% of the time, with BERTweet coming out on top.
- **Recall & Precision**:The BERTweet Model was distinguished by its high Precison and minimal number of false positives. In addition to having a strong recall, MentalBERT discovered a greater number of true positives than false negatives.
- **Evaluation Losses**: As a result of BERTweet's significantly lower rating loss, it is more effective at disseminating new unseen information.
- **F1-Scores**: The performance of the three models was comparable, with BERTweet doing marginally better.

4.0.5 Training and Validation Loss

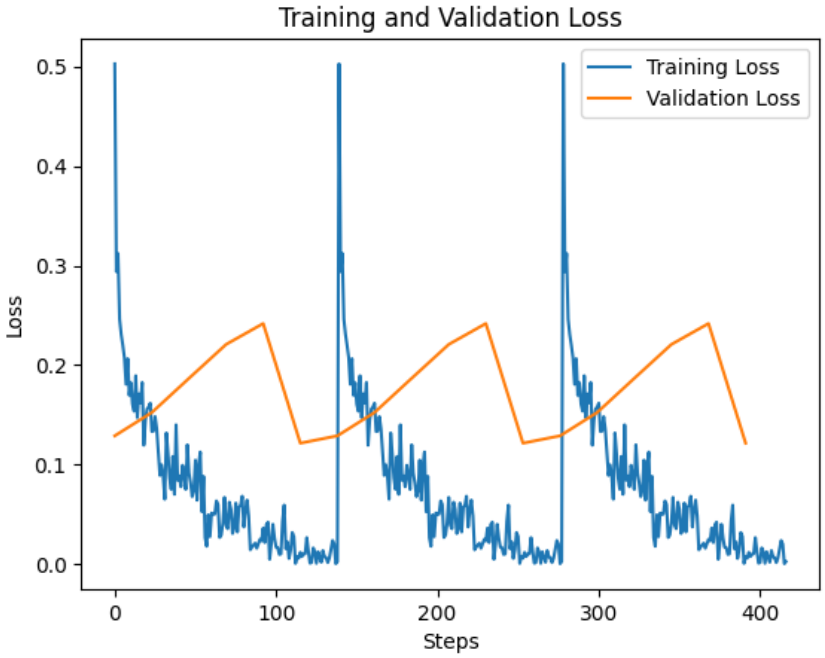


Figure 4.1: DistilBERT Loss Curves

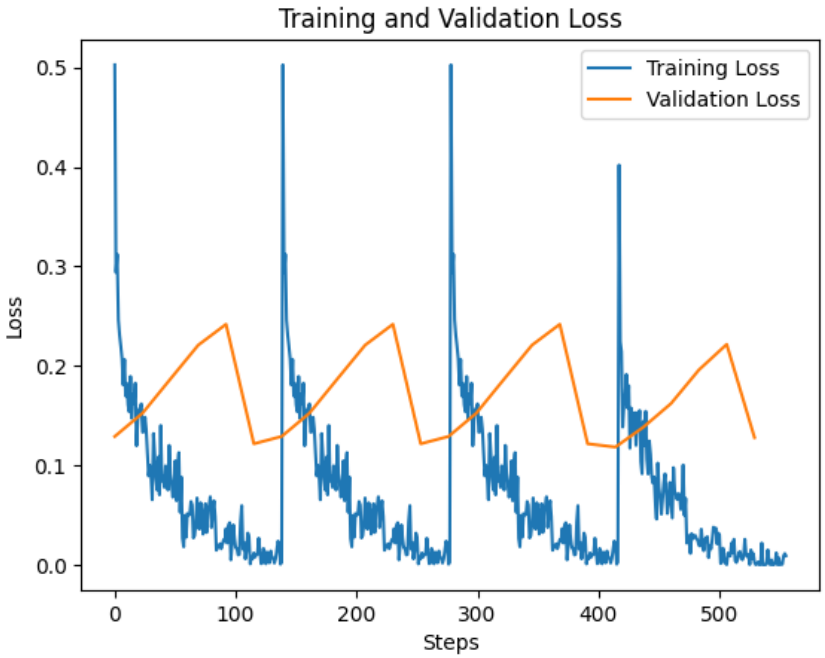


Figure 4.2: MentalBERT Loss Curves

Figure 4.3: BERTweet Loss Curves

4.0.6 Visual Analysis

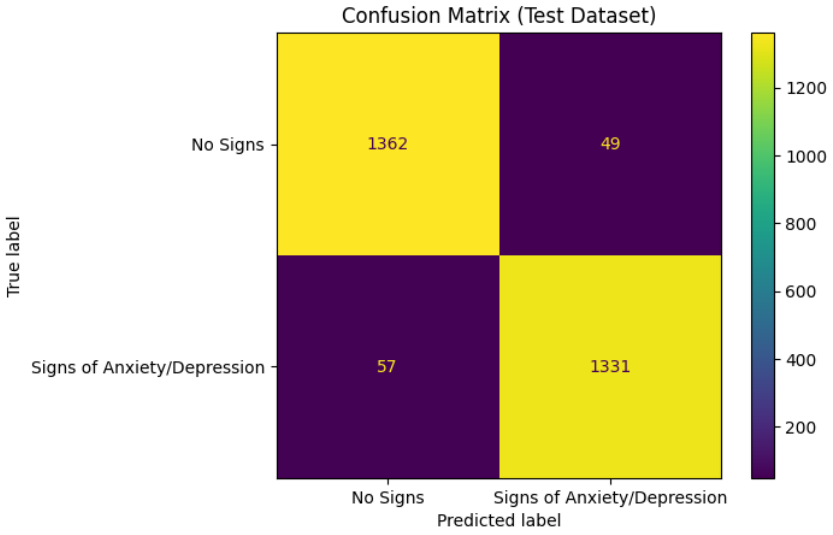


Figure 4.4: DistilBERT Confusion Matrix

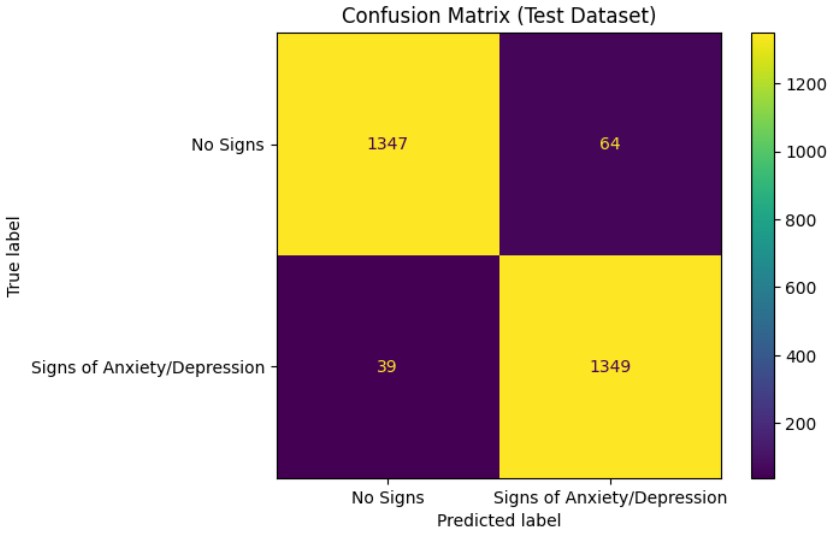


Figure 4.5: MentalBERT Confusion Matrix

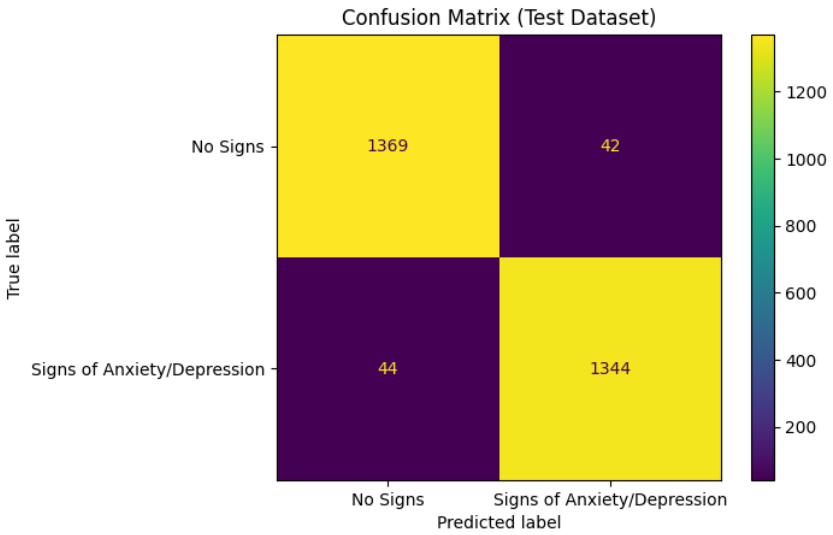


Figure 4.6: BERTweet Confusion Matrix

4.0.7 Results Summary

The fine-tuned DistilBERT,MentalBERT and BERTweet models were evaluated on the task of detecting anxiety and depression from Twitter data. Below is a concise summary of their performance and comparison between them:

Metric	DistilBERT (%)	MentalBERT (%)	BERTweet (%)	Comparison
Accuracy	96.21	96.32	**96.93**	BERTweet highest
Precision	96.46	95.47	**96.97**	BERTweet highest
Recall	95.89	**97.19**	96.83	MentalBERT highest
F1-Score	96.17	96.32	**96.90**	BERTweet highest
Evaluation Loss	0.1218	0.1279	**0.1047**	BERTweet lowest

Figure 4.7: DistilBERT, MentalBERT and BERTweet's Performance Metrics for Comparison

4.1 Discussion

- **MentalBERT:** This test has a high level of accuracy (97.19%), which helps identify feelings of worry and depression while also reducing the number of false negatives. While it is more accurate than DistilBERT and BertWrite, it is (95.47%) less accurate than other two.

- **DistilBERT:** The model was able to function successfully on fresh data and had fewer false positives, as seen by its low evaluation loss (0.1218) and high accuracy (96.46%). When compared to other models, the memory rate and F1 score are both significantly lower.

- **BERTweet:** Outperformed both DistilBERT(Sanh et al.,) and MentalBERT in accuracy (96.93%), precision (96.97%), and F1-score (96.90%). Additionally, it achieved the lowest evaluation loss (0.1047), demonstrating excellent generalization and overall balanced performance.

- **Trade-offs:** MentalBERT's higher recall makes it ideal for applications prioritizing the detection of true positives, such as early intervention for mental health. Meanwhile, BERTweet offers the best overall balance of metrics, making it a strong candidate for robust and accurate mental health monitoring.

Chapter 5

Conclusion

Findings and Contributions

This research focuses on detecting anxiety and depression on Twitter by fine-tuning large language models (LLMs), specifically **DistilBERT** and **MentalBERT**. Both models performed exceptionally well, achieving over 96% accuracy on the test dataset. However, BERTweet emerged as the top-performing model, achieving the highest accuracy (96.93%), precision (96.97%), and F1-score (96.90%), making it the most balanced model overall.

DistilBERT demonstrated robust performance with a precision of 96.45% and an evaluation loss of 0.1217, suggesting fewer false positives and strong generalization. In contrast,**MentalBERT** excelled in recall with a score of 97.19%, reflecting its ability to accurately identify true positive cases. This makes **MentalBERT** particularly effective for tasks where capturing mental health concerns is critical.

The models utilized a standard cross-entropy loss function, and dynamic token padding was incorporated to optimize computational efficiency. The results underscore that fine-tuning transformer-based LLMs is a reliable approach for detecting mental health conditions on social media. The superior performance of **BERTweet** highlights the significant advantage of domain-specific pretraining for real-world applications, including large-scale mental health monitoring.

Despite these achievements, the study acknowledges some limitations. The reliance on English-language textual data restricts the application of these models to multilingual and multimodal scenarios. Additionally, the use of annotated data introduces an element of subjectivity, as expressions of mental health vary widely across cultures and individuals.

5.0.1 Recommendations for Future Work

Future research directions to improve the applicability and performance of models for detecting mental health conditions include the following:

1. **Multilingual Datasets:** FTraining models on multilingual datasets could enhance global relevance and inclusivity, addressing the diversity of linguistic expressions across cultures.

2. **Features Multimodal:** Understanding mental health difficulties and improving the model's projections could be facilitated by the addition of images, videos, and user behavior.(P. Nguyen & Lin,).

3. **Dynamic Thresholding:** IUsing dynamic thresholds to balance precision and recall based on the context, such as public health monitoring versus individual-level detection, could improve practical utility.

4. **Explainability(XAI):** (XAI) explainable ai can improve model predictions, especially for sensitive issues like mental health. (Doshi-Velez & Kim,).

5. **Real-Time Systems:** DDeveloping systems capable of real-time monitoring would enable early identification and intervention for mental health crises, making these tools more impactful.

6. **Cross-Domain Applications:** Applying these methodologies to other social media platforms, like Facebook or Reddit, could broaden the models' scope and generalizability.

5.0.2 Concluding Remarks

This study demonstrates that fine-tuned large language models (LLMs) such as DistilBERT(Sanh et al.,), MentalBERT(Ji et al.,), and BERTweet(D. Q. Nguyen et al.,) are capable of detecting potential mental health disorders using Twitter data. The models showed exceptional performance in terms of accuracy, recall, and F1-scores, making them highly reliable for real-world applications. The research contributes valuable insights to the growing field of AI-driven mental health monitoring by offering a cost-effective and scalable approach. This enables early detection and timely interventions, which are critical for improving mental health outcomes. While limitations persist—particularly in the language and modality constraints of the models—this work lays a solid foundation for addressing these challenges in future studies. Ultimately, this research highlights the transformative impact of LLMs on the evolution of mental health care. By bridging gaps in current technologies, it represents a significant step toward fostering better mental well-being through innovative AI solutions.

References

Alhamed, F., Ive, J., Specia, L. (2024). Classifying social media users before and after depression diagnosis via their language usage: A dataset and study. *ACL Anthology*.

Bucur, A. M. (2024). Leveraging llm-generated data for detecting depression symptoms on social media. *Springer*.

Choudhury, M., Counts, S., Horvitz, E. (2013). Social media as a measurement tool of depression in populations. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3267-3276.

Doshi-Velez, F., Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*. Retrieved from https://arxiv.org/abs/1702.08608

Garcia, M., Wong, E. (2020). Improving model performance with class weights. *AI Research*, *30*, 300-310.

Guo, Z., Lai, A., Thygesen, J. H., Farrington, J., Keen, T. (2024). Large language models for mental health applications: Systematic review. *JMIR Mental Health*.

Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., Cambria, E. (2022). MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of lrec*.

Kessler, R., Bromet, E. (2013). The epidemiology of depression across cultures. *Annual Review of Public Health*, *34*, 119-138.

Nguyen, D. Q., Vu, T., Nguyen, A. T. (2020). BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 9–14).

Nguyen, P., Lin, A. (2021). Advances in multimodal ai for mental health applications. *AI and Society*, *36*, 123-135.

Organization, W. H. (2022). Mental health and covid-19: Early evidence of the pandemic's impact. *WHO Publications*. Retrieved from https://www.who.int/publications/i/item/WHO-2019-nCoV-Sci-Brief-Mental-health-2022

Sanh, V., Debut, L., Chaumond, J., Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, *abs/1910.01108*.

Shah, S. M., Gillani, S. A., Baig, M. S. A., Saleem, M. A. (2024). Advancing depression detection on social media platforms through fine-tuned large language models. *arXiv preprint*. doi: 10.48550/arXiv.2409.14794

Shah, S. M., Gillani, S. A., Baig, M. S. A., Saleem, M. A., Siddiqui, M. H. (2024). Advancing depression detection on social media platforms through fine-tuned large language models. *arXiv preprint arXiv:2409.14794*.

Skianis, K., Pavlopoulos, J., Dogruöz, A. S. (2024). Severity prediction in mental health: Llm-based creation, analysis, evaluation of a novel multilingual dataset. *arXiv preprint*. doi: 10.48550/arXiv.2409.17397

Surin, S. (2024). *Depression twitter dataset feature extraction.* Retrieved from https://www.kaggle.com/datasets/sumitsurin/combined (Accessed: January 2025)

Tavchioski, I., Robnik-Šikonja, M., Pollak, S. (2023). Detection of depression on social networks using transformers and ensembles. *arXiv preprint arXiv:2305.05325*.

Thamrin, S. A., Chen, A. L. P. (2024). Detection of bipolar disorder on social media data utilizing biomedical, clinical and mental health domain fine-tuned word embeddings. *IEEE Explore*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J. (2021). Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Xu, X., Yao, B., Dong, Y., Gabriel, S., Yu, H. (2024). Mental-llm: Leveraging large language models for mental health prediction via online text data. *ACM Digital Library*. doi: 10.1145/3643540

Fine-Tuning Large Language Models For Depression And Anxiety Detection On Twitter

ORIGINALITY REPORT

9%	7%	3%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
2	www.coursehero.com Internet Source	1%
3	neptune.ai Internet Source	1%
4	Submitted to Midlands State University Student Paper	1%
5	ijsrst.com Internet Source	1%
6	danieljude1992.medium.com Internet Source	<1%
7	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	<1%
8	Ali, Muhaddisa Barat. "Deep Learning Methods for Classification of Gliomas and their Molecular Subtypes: From Central	<1%

	Learning to Federated Learning", Chalmers Tekniska Hogskola (Sweden), 2023 Publication	
9	Yang Xiao, Yueshan Huang, Yu Zhao, Fan Xu, Qin Ren, Bing He, Jianhua Yao, Xiao Liu. "Multimodal-AIR-BERT: A Multimodal Pre-trained Model for Antigen Specificity Prediction in Adaptive Immune Receptors", 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2023 Publication	<1 %
10	ebin.pub Internet Source	<1 %
11	researchspace.ukzn.ac.za Internet Source	<1 %
12	arxiv.org Internet Source	<1 %
13	www.mdpi.com Internet Source	<1 %
14	Submitted to UCL Student Paper	<1 %
15	eprints.utem.edu.my Internet Source	<1 %
16	Anshuman Tripathi, Shilpi Birla, Mamta Soni, Jagrati Sahariya, Monica Sharma.	<1 %

	"Multidisciplinary Approaches for Sustainable Development", CRC Press, 2024 Publication	
17	Guo, Wenbin. "Modeling Site-Site Dependency in DNA Methylation Sequencing Data.", University of California, Los Angeles Publication	<1 %
18	vtiya.medium.com Internet Source	<1 %
19	www.speech.kth.se Internet Source	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		