Decoding Mental Health: Assessing Depression Severity Levels via Binary and Multi-Class Classification

Daisy Bui **Yikai Sun**

1 Introduction

Since the start of the 20th century, psychiatry has been considered a field that heavily relied on human expertise, which necessitates mental health professionals with extensive practical experience. Diagnosis in mental health is a complex, ongoing process that involves assessing familial history, situational factors, and environmental influences. Given the ongoing efforts of thousands of mental health and medical experts, the Diagnostic and Statistical Manual of Mental Disorders (DMS) has undergone five significant revisions and will continue evolving to reflect advancements in our comprehension of mental health issues (Association, 2022).

Notably, there has been a significant emphasis on improving context-sensitive analysis and cultivating authentic human-like interactions, characterized by empathy and compassion. Additionally, efforts have been directed towards training large language models (LLMs) to generate empathetic, purposeful text (Hua et al., 2024), with the goal of creating a more therapeutic virtual environment and fostering interactions that mimic human mentalizing processes for users.

In this project, we apply concepts learned in class about LLM models, transformers, and text classification to compare MentalBert, a specialized variant tailored for mental health-related tasks, with the DistilBERT model in the context of depression detection. Given that MentalBert is the pretrained language model for mental healthcare (Ji et al., 2021), we speculate that the accuracy level for this model is thus higher than DistilBERT when performing clinical diagnosis (i.e., depression detection).

Leveraging our understanding of transformer architectures and fine-tuning techniques, we frame depression detection as a text classification task similar to those covered previously in class (i.e., hate speech classification). By conducting error analysis on the predictions of both models, we aim to identify patterns of misclassification and understand their underlying causes, as well as identifying gaps in clinical applicability and ethical considerations to aim at a more responsible development of LLMs in psychotherapy. Thus, analyzing the performance of MentalBert would be the foundational basis to figure the actual applicability of a large language model in assisting or even replacing the work of mental health professionals in the future.

2 Methods & Setup

2.1 Data

We obtained our binary dataset from the HuggingFace Depression Depression dataset, sourced from Reddit and Twitter (Li, 2024). Given the limited background information available on how the data was gathered, it is crucial to acknowledge potential ethical concerns regarding user privacy. To address this, we want to emphasize that we solely collected textual content and refrained from accessing any user profiles or personal information. This dataset comprises 1,546 lines of text, categorized into 775 lines classified as 'Normal' and 770 lines as 'Depression.' (Binary dataset is shown in Table 1)

Depression Detection	Count
Normal	775
Depression	770

Table 1: Binary Classification Dataset Label Distribution

The second dataset for our multi-class classification task comes from a relabeled Reddit dataset containing 3,553 posts, categorized into four depression severity levels: 'minimum', 'mild', 'moderate', and 'severe' (Naseem et al., 2022). A significant challenge with this dataset is its class im-

balance, with the 'Minimum' label being the most prevalent at 2,587 lines, followed by 'Moderate' with 394 lines, 'Mild' with 290 lines, and 'Severe' with 282 lines. (Multi-class dataset is shown in Table 2)

It is important to note that despite the issue with imbalance classification, we selected this dataset because it is one of the few publicly available datasets that adheres to the Depressive Disorder Annotation Scheme (DDAS), which closely aligns with the APA's Diagnostic and Statistical Manual for Mental Disorders (DSM-5) (American Psychological Association, 2013). The DDAS uses a hierarchical structure divided into two major categories: 'No evidence of clinical depression' and 'With evidence of clinical depression,' further branching into various depression symptoms and psychosocial stressor subclasses. These include low mood, disturbed sleep, recurrent thoughts of death, suicidal ideation, educational and occupational problems, feelings of worthlessness or excessive guilt, difficulty concentrating, and indecisiveness, among others (Mowery et al., 2015). Given that our primary goal for this project is to evaluate MentalBERT's performance in detecting the symptoms and severity of mental disorders such as depression, using a dataset that follows the strict guidelines of such manuals would ensure the validity and effectiveness of our model in performing these critical tasks.

Depression Severity	Count
Minimum	2587
Mild	290
Moderate	394
Severe	282

Table 2: Multi-class Classification Dataset Label Distribution

2.2 Methods

- *Hyper-parameters & Test Split*

Our project involves two models, each utilizing three key Python files: model.py, train.py, and util.py. model.py is responsible for defining the models' architecture. train.py contains functions for training the models and computing their final accuracy. util.py offers support functions for model evaluation and data management.

For reproducibility, the README.md file explicitly details the configuration settings that achieve the best accuracies for the Bert and MentalBert models. These settings, specified for both binary and multi-class classification scenarios, include critical hyperparameters such as batch size, learning rate, and number of epochs. The configurations are as follows:

- Binary classification Settings

  Bert:
  - Batch Size: 4
  - Learning Rate: 0.005
  - Epochs: 7

  MentalBert:
  - Batch Size: 4
  - Learning Rate: 0.001
  - Epochs: 5

- Multi-class Classification Settings

  Bert:
  - Batch Size: 12
  - Learning Rate: 0.0003
  - Epochs: 35

  MentalBert:
  - Batch Size: 12
  - Learning Rate: 0.0003
  - Epochs: 35

To comprehensively evaluate the performance of MentalBert, we devised two distinct tasks, each targeting different aspects of depression detection in Reddit posts. The first task focused on binary classification, with a goal to discern whether a given post exhibited signs of depression or not. This task serves as a foundational step in identifying posts that may warrant further intervention or support. In contrast, the second task delved into multi-class classification, aiming to categorize posts based on the severity of depression expressed. This approach enables a better understanding of MentalBert's ability to differentiate between different levels of depression symptoms.

When assessing the performance of MentalBert and Bert models, we utilized a set of metrics, which included accuracy, precision, and recall. Accuracy serves as a fundamental measure, offering an overall assessment of the model's correctness in its predictions. Further, precision provides insights into the model's ability to accurately identify positive cases, such as depression posts, by measuring the ratio of true positive predictions to the total number of positive predictions made by the model. On the other hand, recall assesses the model's capacity to avoid false negatives by measuring the ratio of true positive predictions to the total number of actual positive instances in the dataset. Together,

these metrics allow us to analyze the model's effectiveness in identifying depression-related content while minimizing both false positives and false negatives.

We conducted a binary dataset split with 70% allocated for training, 15% for testing, and 15% for development. The development set is used for fine-tuning hyperparameters to achieve the best results. In contrast, for our multi-class dataset, which is significantly larger (3,553 lines compared to 1,546 lines in the binary dataset), we opted for a different split: 80% for training, 10% for testing, and 10% for development.

The reason for these differing splits lies in the size of the datasets. With the binary dataset being smaller, we allocate a larger proportion to testing and development to ensure effective evaluation and fine-tuning without compromising the training set's size. For the larger multi-class dataset, a higher percentage for training (80%) allows the model to learn from a more extensive set of examples, which is beneficial given the increased complexity of the multi-class classification task. The smaller proportions for testing and development (10% each) are sufficient to evaluate the model and fine-tune hyperparameters, given the overall larger dataset size. This approach ensures that both models are trained optimally according to the specific requirements and characteristics of each task.

- *Strategies for Learning in Class Imbalance*

Incorporating class weights is particularly important in applications such as mental health classification, where the accurate identification of all classes is crucial. In particular, for our multi-class dataset with varying levels of depression severity, it is essential for the model to accurately distinguish between different severity levels, not just the most common one. In imbalanced datasets, models tend to be biased towards the majority class, as it dominates the learning process. By assigning higher weights to the minority classes, we can make the model more sensitive to these underrepresented classes. This adjustment helps ensure that the model pays adequate attention to each class, thereby improving its overall performance and fairness. As a result, the model can be improved in terms of reliability and effectiveness in real-world scenarios, ultimately contributing to better diagnostic and therapeutic outcomes.

Undersampling involves reducing the number of instances in the majority class(es) of a dataset to balance it with the minority class(es). In our case, the original model predominantly predicted one major class, reflecting an imbalanced dataset where one class overshadowed the others.

When we perform undersampling to balance the classes, the model must learn to predict all four classes equally. This shift in focus from a single dominant class to all classes means the model's accuracy is now influenced by its ability to predict each class accurately, rather than just excelling in the dominant class. Consequently, the model's overall accuracy can decrease because it is now evaluated on its performance across all classes, including the previously underrepresented ones (i.e., Mild, Moderate, Severe).

This change is significant because it makes the model's performance more reflective of its ability to make accurate predictions across the entire dataset, rather than being biased towards the dominant class (i.e., Minimum). As a result, the accuracy might drop when the model is assessed on balanced data compared to when it was only focusing on the majority class.

3 Results

As showed in Table 3, in binary classification tasks, MentalBert slightly outperforms Bert, achieving 98.1% across accuracy, precision, and recall, compared to Bert's 97.0%. This performance underscores MentalBert's enhanced capability in straightforward task settings, suggesting it has better mechanisms for generalizing from training data without capturing noise and outliers in the context of mental-health-related tasks.

In multiclass classification tasks, the differences become more nuanced. While MentalBert edges out Bert in accuracy with 65.9% versus 65.3%, it slightly lags behind in precision and recall. Bert's precision of 44.8% and recall of 52.5% surpass MentalBert's 43.0% and 51.6%, respectively, indicating that Bert might be more effective at identifying relevant cases within a complex label space. We hypothesize that MentalBert's algorithm might prioritize overall accuracy over precision and recall, making it slightly better at classifying more cases overall but at a minor cost to its ability to precisely identify and recall all relevant instances.

To further understand the performance differences between these two models, accuracies at different epochs and learning rates were analyzed[1].

[1]All results in the figures are obtained from the test data

Models	Binary			Multi-Class		
	Acc.	Pre.	Rec.	Acc.	Pre.	Rec.
Bert	97.0%	97.0%	97.0%	65.3%	44.8%	52.5%
MentalBert	98.1%	98.1%	98.1%	65.9%	43.0%	51.6%

Table 3: A Summary of BERT and MentalBERT Performance on Binary and Multi-class Classification Task
Notes: All computed values are rounded to the nearest tenths.

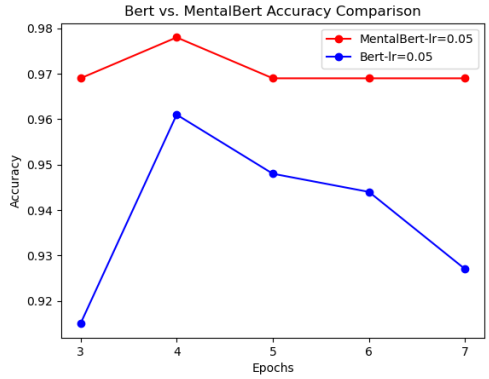

Figure 1: Bert vs. MentalBert Performance on Binary Classification

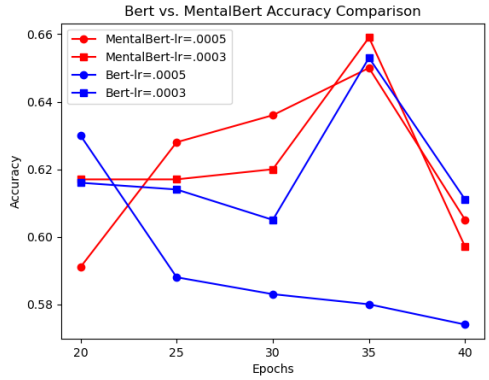

Figure 2: Bert vs. MentalBert Performance on Multi-class Classification

From Figure 1 MentalBert shows high and stable accuracy from epochs 3 to 7 at a learning rate of 0.05 in binary classification tasks, indicating effective learning and minimal evidence of overfitting. Conversely, Bert exhibits a promising initial performance but suffers a decline in accuracy after the fourth epoch, likely due to overfitting. Hence, for future development, Bert could benefit from strategies such as early stopping, regularization, or advanced data augmentation to mitigate such problem.

Both MentalBert and Bert show variability in performance in multiclass classification, as showed in Figure 2, peaking around the 30th epoch. This peak suggests an ideal stopping point to maximize accuracy before potential overfitting or other degradations, such as learning rate decay or learning capacity exhaustion. MentalBert consistently outperforms Bert at both learning rates, likely due to a superior internal architecture or more effective optimization strategies, which better handle the complexities of multiclass classification in mental health contexts. The selection of an appropriate learning rate and the right number of epochs, particularly determining the optimal stopping moment, are thereby crucial to enhance performance and avoid unnecessary resource use.

In summary, our comparative analysis across various configurations and tasks underscores the critical importance of selecting the right model, optimizing learning rates, and managing training epochs to achieve optimal accuracy in NLP tasks. While MentalBert consistently outperforms BERT, demonstrating its better suitability for depression detection, this does not imply that MentalBert is fully ready for real-world application in the mental health field. There is still significant room for improvement and development to enhance these language models' accuracy, reliability and effectiveness in practical scenarios. Further research and development are essential to refine these models, address their limitations, and ensure they can accurately and sensitively handle the complexities of mental health data. Therefore, our findings emphasize the ongoing need for innovation and rigorous testing to bring these advanced language models to a level where they can be confidently used by mental health professionals.

4 Related Work

The rapid advancement of Large Language Models (LLMs) has prompted their integration into diverse areas within mental health care. Notably, there has been a significant emphasis on improving context-sensitive analysis and cultivating authentic human-like interactions, characterized by empathy

split.

and compassion. Numerous research efforts have targeted the enhancement of these abilities, focusing on specialized language usage (Liu et al., 2023) and strategic reasoning (Zhang et al., 2023). Moreover, efforts have also been made to train LLMs in generating empathetic, purposeful text (Hua et al., 2024), aiming to establish a more therapeutic virtual environment and fostering more human-like interactions (akin to mentalizing) for users. Consequently, the utilization of LLMs in psychology and other health domains is poised to serve as a robust ally in clinical evaluation processes, providing more precise feedback with reduced human biases and heuristics.

Furthermore, the emergence of publicly accessible data from social media, although now becoming more restricted, has facilitated the use of Large Language Models (LLMs) as classification tools for comprehensive diagnosis. This application involves binary classification to detect the presence or absence of specific mental health conditions, including PTSD (Dekel et al., 2023), schizophrenia (Mascio et al., 2021), and depression (Lan et al., 2024). Moreover, LLMs are also employed in multiclass classification tasks to assess severity levels or subcategories, such as predicting the severity of depression according to DSM-5 (minimal, mild, moderate, and severe) (Naseem et al., 2022) and identifying suicide risks in various levels (supportive, indicator, ideation, behavior, and attempt) in accordance to the Columbia Suicide Severity Rating Scale (C-SSRS) (Gaur et al., 2019).

Our research builds on the pioneering work of MentalBERT, a groundbreaking contributor in mental health language modeling (Ji et al., 2021). Their research provides a framework for training language models on mental health-related text, enabling the creation of advanced models for both health-condition classification and in-depth analysis and support in the mental health field. We conduct our work with a strong emphasis on ethical considerations, ensuring our dataset is free of personally identifiable information to protect user privacy.

In addition to evaluating model performance, we strive to advance the understanding of how language models can be optimized for mental health applications. This involves not only comparing accuracy and classification metrics but also assessing the models' ability to capture nuanced expressions of mental health conditions. By testing MentalBERT against BERT in binary and multi-class classification tasks, we thereby hope to uncover insights that can inform future improvements in mental health language modeling.

5 Ethical Consideration

The ethical considerations surrounding the use of Large Language Models (LLMs) and other AI models in the mental health field are of utmost concern, particularly due to the sensitive nature of the data involved in mental health care applications. Various studies directly highlighted the necessity for robust data protection and adherence to ethical standards. Specifically, Robillard et al. (2015) delved into the assessment of online tests designed for diagnosing Alzheimer's disease, analyzing aspects such as scientific reliability and privacy protection. Their study revealed that these freely available tests exhibited significant shortcomings in terms of diagnostic accuracy and failed to adequately safeguard the confidentiality of test data. Consequently, while online testing has the potential to facilitate more awareness about certain mental health concerns, the risks it poses might outweigh the potential benefits.

Additionally, in one of the LLM-Based Conversation Agents, there have been reports of models generating harmful content such as violence, drug-related material, and non-consensual sexual content without user initiation (Ma et al., 2023). This issue underscores the imperative for the establishment of more stringent and transparent guidelines aimed at safeguarding users against exposure to inappropriate and inaccurate information.

Another critical aspect to consider is that providing advice to users is not considered ethical in the field of psychiatry and counseling due to the diverse backgrounds and circumstances of individuals. Mental healthcare practices often prioritize promoting patient autonomy rather than offering straightforward solutions (Entwistle et al., 2010). Therefore, the complex dynamics of human psychology and moral ethics cannot be comprehensively learned and replicated by LLMs alone. In reinstating the work of Yao et al. (2023), it is crucial to acknowledge that the usage of LLMs in offering advice is not suitable in digital mental healthcare.

Given the increasing prevalence of mental disorders (Winkler et al., 2020), professionals in mental healthcare may find value in improved identification and intervention assistance offered by large language models. However, the research findings

cited above highlight significant concerns regarding the development of such algorithms, particularly regarding issues of consent, data management, and the validity of results for clinical assessment. The inability of LLMs in addressing these sensitive issues have posed significant challenges for both computer science researchers and clinicians in devising a consensus solution for monitoring and intervention systems that adhere to the highest ethical standards in cyberspace. These studies thereby highlight the importance of employing LLMs responsibly, with a commitment to fostering more reliable and transparent assessment and treatment practices in both computational work and mental health.

References

American Psychological Association. 2013. *Publications Manual*. American Psychological Association, Washington, DC.

American Psychiatric Association. 2022. *Diagnostic and Statistical Manual of Mental Disorders*.

Sharon Dekel, Alon Bartal, Kathleen M. Jagodnik, and Sabrina J. Chan. 2023. ChatGPT demonstrates potential for identifying psychiatric disorders: Application to Childbirth-Related Post-Traumatic Stress Disorder. *Research Square (Research Square)*.

Vikki Entwistle, Stacy M Carter, Alan Cribb, and Kirsten McCaffery. 2010. Supporting patient autonomy: The importance of clinician-patient relationships. *Journal of general internal medicine*, 25(7):741–745.

Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Amit P. Sheth, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit P. Sheth, Randy Welton, and Jyotishman Pathak. 2019. Knowledge-aware assessment of severity of suicide risk for early intervention. *Knowledge-aware assessment of severity of suicide risk for early intervention*.

Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Yi-Han Sheu, Peilin Zhou, Lauren V. Moran, Sophia Ananiadou, and Andrew L. Beam. 2024. Large Language Models in Mental Health Care: a Scoping Review. *arXiv (Cornell University)*.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. MentalBERT: Publicly available Pretrained Language Models for Mental Healthcare. *arXiv (Cornell University)*.

Xiaochong Lan, Yiming Cheng, Sheng Li, Gao Chen, and Yan Hui Li. 2024. Depression Detection on Social Media with Large Language Models. *arXiv (Cornell University)*.

Moria Li. 2024. Depressiondetection$_j sondatasetsathuggingface$.

Siyang Liu, Naihao Deng, Sahand Sabour, Jia Yang, Minlie Huang, and Rada Mihalcea. 2023. Task-Adaptive Tokenization: Enhancing Long-Form Text Generation efficacy in mental health and beyond. *arXiv (Cornell University)*.

Zhenqiang Ma, Yong Ma, and Zhaoyuan Su. 2023. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. *arXiv (Cornell University)*.

Aurelie Mascio, Robert Stewart, Riley Botelle, Marcus Williams, Luwaiza Mirza, Rashmi Patel, Thomas A Pollak, Richard Dobson, and Angus Roberts. 2021. Cognitive impairments in schizophrenia: a study in a large clinical sample using natural language processing. *Frontiers in digital health*, 3.

Danielle Mowery, Craig Bryan, and Mike Conway. 2015. Towards developing an annotation scheme for depressive disorder symptoms: A preliminary study using Twitter data. .

Usman Naseem, Adam G. Dunn, Jin-Man Kim, and Matloob Khushi. 2022. Early identification of depression severity levels on Reddit using ordinal classification. *Proceedings of the ACM Web Conference 2022*.

Julie M. Robillard, Judy Illes, Marcel Arcand, B. Lynn Beattie, Sherri Hayden, P.D. Lawrence, Joanna McGrenere, Peter B. Reiner, Dana Wittenberg, and Claudia Jacova. 2015. Scientific and ethical features of English-language online tests for Alzheimer's disease. *Alzheimer's dementia. Diagnosis, assessment disease monitoring*, 1(3):281–288.

Petr Winkler, Tomáš Formánek, Karolína Mladá, Anna Kågström, Zuzana Mohrova, Pavel Mohr, and Ladislav Csémy. 2020. Increase in prevalence of current mental disorders in the context of COVID-19: analysis of repeated nationwide cross-sectional surveys. *Epidemiology and psychiatric sciences*, 29.

Xuewen Yao, Miriam Mikhelson, S. Craig Watkins, Eunsol Choi, Edison Thomaz, and Kaya De Barbaro. 2023. Development and evaluation of three chatbots for postpartum mood and anxiety disorders. *arXiv (Cornell University)*.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2023. Ask an expert: Leveraging language models to improve strategic Reasoning in Goal-Oriented Dialogue Models. *arXiv (Cornell University)*.