Vamsidhar Boddu (G40705113)
DATS6312: Report
Prof. Dr. Nung Rui
Date: 12/12/2024

Predicting Mental Health Status in Clinical Patients Using NLP Techniques

I. Introduction

The timely diagnosis of mental health conditions remains a significant challenge in healthcare due to the subjective nature of assessments and reliance on patients' self-reported statements. These challenges are exacerbated by time constraints faced by mental health professionals and inconsistencies in interpreting patient narratives, often leading to delays in diagnosis and treatment.

This project addresses these issues by leveraging Natural Language Processing (NLP) techniques and transformer-based models to develop a diagnostic application capable of automatically classifying mental health statuses from textual input. By analyzing statements from patients, the application categorizes mental health conditions into relevant categories such as "Anxiety," "Depression," "Normal," and "Stress" etc.

The development pipeline integrates several NLP methodologies, including tokenization, stopword removal, text vectorization, and pre-trained transformer-based models, such as Google's BERT models. Using a dataset sourced from Kaggle, this project builds on state-of-the-art techniques alongside libraries like TensorFlow, PyTorch, and Scikit-Learn to create a classification model capable of identifying patterns indicative of mental health issues.

Finally, the project culminates in the deployment of an intuitive, user-friendly Streamlit-based web application. The web app allows users to input statements, upload files for bulk predictions, or answer dynamic questions to receive a diagnosis. This robust and scalable NLP-driven solution aims to assist clinicians in identifying patients at risk, enabling timely interventions and improving mental health management in both clinical and telemedicine settings.

II. Dataset

The dataset used in this project is a comprehensive compilation of mental health-related textual data, collected and curated to train and evaluate the classification models effectively. The dataset is sourced from Kaggle, consisting of textual statements indicative of various mental health conditions. These statements were either self-reported by individuals or sourced from various social media domains like Facebook, Reddit, X etc.

Key Characteristics

1. Size:

The raw dataset contains 53,045 entries, including various mental health-related textual statements. After thorough data cleaning, the dataset was reduced to 51,060 records.

2. Class Labels:

The dataset is categorized into seven distinct mental health conditions:

- **Anxiety**
- **Depression**
- **Normal**
- **Stress**
- **Bipolar Disorder**
- **Personality Disorder**
- **Suicidal Thoughts**

3. Features:

Statement: A textual representation of the user's mental health condition or self-reported thoughts.

Status: The corresponding mental health label for each statement.

III. Data Preprocessing

During the data cleaning phase, several steps were undertaken to ensure the integrity and quality of the dataset. Initially, the dataset comprised **53,043 rows**, but after identifying and removing **1,588 duplicate rows** and handling missing values, the final cleaned dataset contained **51,093 rows**. This process resulted in the removal of **1,950 rows**, ensuring that only unique and complete records were retained for analysis. The dataset was thoroughly checked for missing values, and it was confirmed that no missing data remained after the cleaning process.

An analysis of the class distribution revealed significant insights into the dataset's composition. Among the mental health categories, "Normal" had the highest representation with **16,040 entries**, followed by "Depression" with **15,094 entries** and "Suicidal" with **10,644 entries**. The remaining categories included "Anxiety" (**3,623 entries**), "Bipolar" (**2,501 entries**), "Stress" (**2,296 entries**), and "Personality Disorder" (**895 entries**). This distribution highlighted potential class imbalances, which were later addressed during model development.

Further linguistic analysis of the cleaned data provided valuable textual insights. The top 10 most frequently occurring words included "like" (**37,107 occurrences**), "feel" (**34,354 occurrences**), and "want" (**28,109 occurrences**), followed by other terms such as "know," "life," "get," and "time." These findings offered a deeper understanding of the common themes and language patterns present in the dataset, serving as a basis for feature

engineering and model training. This comprehensive approach ensured that the data was not only clean and complete but also enriched with critical insights for subsequent processing stages.

IV. Model Experimentation and Analysis

The experimentation process for this project built upon the foundational work by my teammates, who implemented models such as LSTM, BERT-Base, and BERT-Large. These models provided insights into the limitations and strengths of various architectures for addressing mental health text classification. Building on this groundwork, I focused on advanced transformer-based models, specifically RoBERTa-Base and RoBERTa-Large, which offered significant improvements in handling contextual and nuanced data.

The LSTM model employed by my teammates utilized GloVe embeddings to capture sequential patterns in textual data. While this approach demonstrated decent accuracy (~70%), it struggled with minority classes like "Stress" and "Personality Disorder" due to its limited ability to capture complex contextual dependencies. To overcome these limitations, BERT-Base was introduced, leveraging the transformer architecture's bidirectional attention mechanism. BERT-Base achieved an accuracy of ~79% and a macro F1-score of ~76%, showcasing an improvement in overall performance. However, it still faced challenges with recall for minority classes and exhibited signs of overfitting during extended training. Moving forward, BERT-Large was implemented, offering deeper layers and more parameters to enhance representation capacity. This model achieved an accuracy of ~80% and a macro F1-score of ~77%, showing incremental improvements in handling minority classes but falling short of addressing the class imbalance comprehensively.

Building on these insights, I started to explore RoBERTa models, which offered enhanced pretraining strategies compared to BERT.

RoBERTa-Base:

RoBERTa (A Robustly Optimized BERT Pretraining Approach) enhances the standard BERT architecture by improving training efficiency and utilizing more data during pretraining. It removes the Next Sentence Prediction (NSP) task, allowing the model to focus solely on masked language modeling, which better aligns with natural language understanding tasks. For this project, RoBERTa-Base was fine-tuned to classify mental health statuses from textual data.

RoBERTa-Large:

The transition to RoBERTa-Large aimed to utilize its increased parameter count and larger capacity for learning complex patterns and dependencies in the dataset. With deeper layers and a wider architecture, RoBERTa-Large offered better representation of nuanced language features present in mental health-related statements. This iteration was fine-tuned on the same preprocessed dataset without applying class weights, focusing purely on the inherent capacity of the model to adapt to class imbalances.

RoBERTa-Large (without class weights) achieved an accuracy of ~81% and a macro F1-score of ~78%, showcasing improvements in minority class performance. However, the model still displayed biased precision and recall for certain categories, such as "Stress" and "Personality Disorder," which are less represented in the dataset. This result demonstrated that while the enhanced architecture of RoBERTa-Large significantly improved the model's ability to generalize, additional mechanisms were required to balance the learning process for underrepresented classes.

RoBERTa-Large (With Class Weights):

To address the persistent class imbalance issues, class weights were incorporated into the training process for RoBERTa-Large. Class weights were calculated using a balanced weighting scheme, assigning higher weights to minority classes like "Personality Disorder" and "Stress" to penalize misclassifications during training. This approach effectively shifted the focus of the loss function, ensuring that the model allocated more importance to minority classes during optimization.

RoBERTa-Large with class weights achieved significant improvements in recall for underrepresented classes, such as "Bipolar," "Personality Disorder," and "Stress," albeit at a slight cost to the overall accuracy (~73%). This iteration demonstrated the importance of trade-offs between overall performance and minority class representation. Furthermore, the application of focal loss alongside class weights enhanced the model's ability to prioritize harder-to-classify examples, resulting in a more balanced output across all categories. This was a critical step in ensuring that the model aligns with the overarching goal of timely mental health diagnosis, as accurately identifying minority categories can facilitate targeted interventions for at-risk individuals.

The curated dataset, with its seven distinct mental health categories, played a pivotal role in enhancing the model's performance. The diversity of textual data allowed the models to capture nuanced language patterns associated with various mental health conditions. Preprocessing techniques, such as stopword removal, lemmatization, and linguistic feature extraction, ensured that the data was clean and semantically rich, enabling the models to focus on relevant features.

The transformer-based architecture of RoBERTa was particularly well-suited to this task, as it excelled in handling context-dependent language and understanding the subtle distinctions between categories like "Stress" and "Anxiety." The incorporation of class weights further ensured that the model's learning process aligned with the imbalanced nature of the data, emphasizing the importance of minority classes. This synergy between the dataset and the models not only improved classification performance but also enhanced the reliability of the predictions, making the system a robust tool for mental health diagnostics.

Data Splitting:

Data splitting is the foundational step for model training and evaluation. The dataset was divided into three subsets: Training Set (70%), Validation Set (15%), and Test Set (15%). The split was performed in a stratified manner to maintain class distribution across subsets. The stratified split ensured that minority classes, such as "Stress" and "Personality Disorder," were adequately represented, preventing the model from being biased toward the majority classes.

Regularization:

To prevent overfitting, dropout regularization was used in transformer layers of the RoBERTa model. Dropout introduces stochasticity by deactivating random neurons during training, thereby preventing the model from relying heavily on specific features. This technique was crucial for models like **RoBERTa-large**, which have millions of parameters, ensuring better generalization on unseen data.

Optimizer and Learning Rate Scheduler:

Employed the AdamW optimizer, which includes weight decay to reduce overfitting further. A linear learning rate scheduler with a warm-up phase was used to adjust the learning rate dynamically during training.

AdamW: Combines adaptive learning rates with L2 regularization to stabilize training.

Linear Scheduler: Smoothly reduces the learning rate during training, preventing sudden changes that might destabilize optimization.

Focal Loss:

Class imbalance was a significant challenge in the dataset. To address this, we implemented Focal Loss, which focuses more on harder-to-classify examples by adjusting their contribution to the overall loss. Focal Loss was particularly effective in improving recall for minority classes like "Stress" and "Personality Disorder."

Alpha: Controls the balance between positive and negative samples.

Gamma: Emphasizes harder examples by reducing the impact of easy-to-classify samples.

Class Weights:

To further mitigate the impact of class imbalance, calculated class weights based on the frequency of each class in the dataset. These weights were incorporated into the loss function to prioritize minority classes.

This approach ensured that the loss function penalized errors for minority classes more than those for majority classes, leading to a more balanced model.

Hyperparameter Tuning:

Hyperparameter tuning was a critical component of the project. For each model, we experimented with various parameters to optimize performance:

Learning Rate:

Initial experiments with 1e-4 caused unstable training.

Final models used a learning rate of 1e-5, achieving smooth convergence.

Batch Size:

Experimented with sizes 4, 8, and 16.

Optimal results were obtained with a batch size of 8, balancing memory usage and gradient stability.

Epochs:

Tested training for 5, 7, and 10 epochs.

Early stopping was implemented based on validation loss to prevent overfitting.

Dropout Rates:

Evaluated dropout probabilities of 0.1, 0.2, and 0.3.

Selected 0.1 to retain enough information while introducing regularization.

Training Loop:

The training loop was designed to optimize the model's weights iteratively. Each batch of data was passed through the model, and the loss was computed and backpropagated.

Validation Loop:

The validation loop was executed after each epoch to evaluate the model's performance on unseen data. This process helped tune hyperparameters and detect overfitting early.

IV. Results:

RoBERTa-Large:

Validation Results:

	precision	recall	f1-score	support
Anxiety	0.75	0.90	0.82	543
Bipolar	0.88	0.78	0.82	375
Depression	0.77	0.77	0.77	2264
Normal	0.96	0.92	0.94	2396
Personality disorder	0.67	0.66	0.67	134
Stress	0.64	0.76	0.69	344
Suicidal	0.73	0.71	0.72	1597
accuracy			0.81	7653
macro avg	0.77	0.79	0.78	7653
weighted avg	0.82	0.81	0.81	7653

The RoBERTa-Large model, when trained without the inclusion of class weights, demonstrated strong overall performance, particularly for the majority classes, while still exhibiting some challenges in handling minority classes. The overall accuracy of the model was **81%**, with a **macro-average F1-score** of **0.78** and a **weighted F1-score** of **0.81**. These results reflect the model's capacity to perform well for classes that are overrepresented in the dataset but highlight its limitations in balancing predictions across minority and majority classes.

Class-Wise Performance:

Anxiety: The model achieved a **precision of 0.75** and a **recall of 0.90**, resulting in an **F1-score of 0.82**. This suggests that while the model was good at identifying cases of "Anxiety" (high recall), it produced a higher number of false positives, reducing precision.

Bipolar: With an **F1-score of 0.82**, the model showed promise in identifying "Bipolar" cases. However, the relatively low recall (**0.78**) indicates that the model missed some true cases.

Depression: For "Depression," the model achieved a balanced **F1-score of 0.77**, with both precision and recall at **0.77**. This performance reflects its ability to handle this class relatively well, albeit with room for improvement.

Normal: The class "Normal," being the most frequent in the dataset, achieved the highest scores with a **precision of 0.96**, **recall of 0.92**, and an **F1-score of 0.94**. This dominance indicates that the model heavily relied on the majority class for overall accuracy.

Personality Disorder: Despite some improvements, the minority class "Personality Disorder" only achieved an **F1-score of 0.67**, with **precision at 0.67** and **recall at 0.66**, indicating the difficulty of predicting this underrepresented category.

Stress: The model performed moderately for "Stress," with an **F1-score of 0.69**, reflecting challenges in precision and recall.

Suicidal: The model achieved an **F1-score of 0.72** for "Suicidal," with **precision at 0.73** and **recall at 0.71**, showcasing reasonable performance for this critical class.

While the model performed well for majority classes, it struggled to adequately handle minority class predictions. This disparity can be attributed to the imbalance in the dataset, where underrepresented classes received less attention during training. The absence of mechanisms to explicitly address class imbalance, such as class weights or focal loss, resulted in strong performance for well-represented classes at the expense of the minority ones.

RoBERTa-Large with Class Weights:

Validation Results:

	precision	recall	f1-score	support
Anxiety	0.75	0.78	0.76	543
Bipolar	0.67	0.87	0.76	375
Depression	0.85	0.49	0.62	2264
Normal	0.96	0.84	0.90	2396
Personality disorder	0.34	0.79	0.48	134
Stress	0.44	0.82	0.57	344
Suicidal	0.59	0.84	0.70	1597
accuracy			0.73	7653
macro avg	0.66	0.78	0.68	7653
weighted avg	0.79	0.73	0.73	7653

The introduction of class weights in the RoBERTa-Large model training process brought significant changes in performance metrics, especially for minority classes. This adjustment shifted the model's focus from achieving high overall accuracy to balancing predictions across all classes. As a result, the overall accuracy dropped to 73%, and the macro-average F1-score declined to 0.68, while the weighted F1-score reduced to 0.73. However, these declines in aggregate metrics are offset by substantial improvements in the recall and F1-scores of minority classes, reflecting the model's increased sensitivity to underrepresented categories.

Class-Wise Performance

Anxiety: The model exhibited a slight drop in performance for "Anxiety," with an **F1-score of 0.76** compared to the previous 0.82. This was due to a reduction in recall (**0.78**) as the model reallocated its focus to other classes.

Bipolar: The recall for "Bipolar" improved significantly to **0.87**, resulting in an **F1-score of 0.76**, up from 0.67 in the previous model. This reflects the model's enhanced ability to detect cases of this minority class.

Depression: The class "Depression" experienced a decline in **F1-score to 0.62**, with a significant drop in recall (**0.49**). This trade-off illustrates the impact of prioritizing minority classes at the expense of some majority class predictions.

Normal: For "Normal," the model's performance decreased slightly, with an **F1-score of 0.90**, down from 0.94. This reflects the redistribution of the model's focus toward minority classes.

Personality Disorder: This class saw a marked improvement in performance, with recall increasing to **0.79**, leading to an **F1-score of 0.48**, up from 0.34. This highlights the effectiveness of class weighting in addressing imbalance.

Stress: The recall for "Stress" improved to **0.82**, significantly boosting the **F1-score to 0.57**. This improvement underscores the model's better handling of this minority class.

Suicidal: The class "Suicidal" also benefited from the inclusion of class weights, with a recall of **0.84** and an **F1-score of 0.70**, up from the previous 0.72.

The inclusion of class weights addressed the model's earlier bias towards majority classes by assigning higher weights to underrepresented categories. This adjustment significantly improved recall for critical minority classes like "Personality Disorder," "Stress," and "Bipolar," which are vital for mental health diagnostics. However, the trade-offs included a reduction in precision for majority classes and a drop in overall accuracy. This demonstrates the inherent challenge of balancing precision and recall in imbalanced datasets.

The comparison between RoBERTa-Large without class weights and RoBERTa-Large with class weights highlights a pivotal trade-off in model performance. While the model without class weights excelled in achieving high overall accuracy (**81%**) and balanced performance for majority classes, it struggled to effectively identify minority class samples, such as "Stress" and "Personality Disorder." This imbalance left critical gaps in the diagnostic reliability of the system, especially for underrepresented categories.

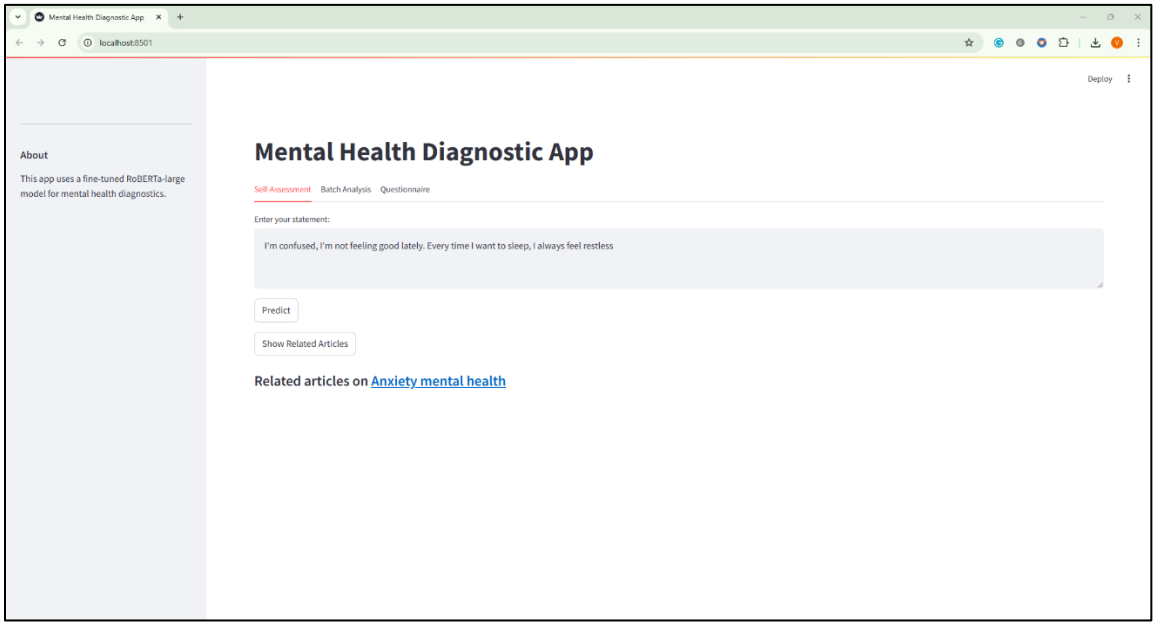
In contrast, the incorporation of class weights in RoBERTa-Large redistributed the model's focus, prioritizing recall for minority classes. Although this approach resulted in a slight decrease in overall accuracy (**73%**), it significantly improved the model's ability to detect minority classes, as evidenced by notable recall gains for "Stress," "Bipolar," and "Personality Disorder." This enhancement ensures a more equitable representation of all

mental health statuses, which is critical in real-world applications where accurate identification of minority classes can have profound implications for timely mental health interventions.

Ultimately, RoBERTa-Large with class weights emerged as the superior model for this project, striking a balance between improving recall for underrepresented classes and maintaining robust overall performance, making it better aligned with the goal of equitable mental health diagnostics.

V. Streamlit APP

As part of the Streamlit-based application, my primary contributions focused on enhancing the Type Input (Tab 1) and Answer Questions (Tab 3) sections. In Tab 1, I designed and implemented the functionality to allow users to input free-form text describing their thoughts or feelings, which the fine-tuned RoBERTa-Large model processes to predict the corresponding mental health status. The system ensures seamless integration with the prediction pipeline, efficiently handling diverse user inputs and providing actionable insights in real time. While the Google News API integration, which fetches topic-relevant articles based on predictions, was led by my teammate, I collaborated to align this feature with the prediction outputs to ensure consistency and usability.



Streamlit Web App

In Tab 3, I developed a framework for engaging users through a randomized selection of five questions from a predefined pool. These questions aim to gather nuanced details about the user's mental health status. I ensured that the answers were dynamically combined into a cohesive input for the RoBERTa-Large model, allowing for accurate and context-aware predictions. To enhance user experience, I implemented robust validation to handle incomplete or invalid answers and created mechanisms to ensure predictions align with multi-sentence inputs.

Both tabs were optimized for prediction accuracy, user interactivity, and seamless functionality. Tab 1 offers a quick assessment tool with real-time results and resource links, while Tab 3 provides a more in-depth diagnostic process that encourages self-reflection. My contributions to these sections ensured the app's usability, accuracy, and overall value for individuals seeking mental health support.

VI. Conclusion

This project has been an enriching journey, providing me with profound insights into the intricate process of building a robust AI-driven mental health diagnostic application. At the outset, my primary focus was understanding the dataset, exploring preprocessing techniques, and addressing challenges such as missing data, duplicate entries, and class imbalance. These initial steps laid the foundation for a well-structured dataset, enabling accurate and meaningful predictions.

The experimentation process across multiple models ultimately led to the selection of **RoBERTa-Large with class weights** as the optimal solution. While models like **LSTM**, **BERT-Base**, and **BERT-Large** showed incremental improvements in handling sequential text and contextual nuances, they struggled with achieving balanced performance across all mental health categories. The introduction of **class weights** in **RoBERTa-Large** addressed the inherent imbalance in the dataset, substantially improving recall for minority classes such as "Stress" and "Personality Disorder" while maintaining high accuracy and F1 scores across other classes.

The implementation of **focal loss**, coupled with hyperparameter tuning and an extended training loop, further enhanced the model's capability to prioritize harder-to-classify examples. This refined approach enabled the model to strike a balance between overall accuracy and the critical recall metric, making it well-suited for real-world applications where the correct identification of minority classes is vital.

VII. Future Work

This project, while a significant milestone, opens up several avenues for future enhancements:

- **Domain-Specific Model Training**: To further improve accuracy and relevance, training domain-specific models like **MentalBERT** or **PsychBERT** on curated datasets specific to mental health contexts could provide better insights and predictions. These models, designed with a deep understanding of mental health terminologies and patterns, would add significant value.
- **Expanding the Dataset**: Increasing the dataset size by curating textual data from diverse sources such as mental health forums, social media, and clinical records (while ensuring data privacy and ethics) could enrich the model's ability to generalize across varied user inputs.
- **Scalable Application Deployment**: Future iterations of the application could focus on deploying a scalable version that serves two distinct audiences:
    - A **patient version** with simplified features for self-assessment and immediate resources.

    - A **clinician version** with detailed analytics, patient history tracking, and actionable insights to aid in clinical decision-making.
- **AI-Driven Enhancements**: The application could evolve into a fully AI-driven solution with features like automated therapeutic recommendations, interactive conversational agents for mental health support, and dynamic learning capabilities to adapt to new trends in mental health research.
- **Integration with Telemedicine Platforms**: Deploying this tool as part of telemedicine platforms could enable seamless integration into existing healthcare workflows, providing clinicians and patients with an accessible, efficient, and reliable tool for mental health diagnostics.

Reference

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need*. Retrieved from https://arxiv.org/abs/1706.03762

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. Retrieved from https://arxiv.org/abs/1810.04805

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. Retrieved from https://arxiv.org/abs/1907.11692