Predicting Mental Health Status in Clinical Patients Using NLP Techniques

NLP for Data Science Final Project
Fall 2024

-Vaishnavi Tamilvanan
(G38942246)

1. Introduction

Project Overview

Mental health issues have become increasingly prevalent in society, impacting individuals' well-being, productivity, and quality of life. Early detection of mental health conditions through automated systems can provide crucial support, enabling timely intervention and appropriate care. This project aims to leverage Natural Language Processing (NLP) techniques to predict mental health conditions based on textual data. The dataset used for this project comprises user-generated text, categorized into various mental health conditions, including Normal, Depression, Suicidal, Anxiety, Stress, Bi-Polar, and Personality Disorder. The primary objective is to develop model that can accurately classify these conditions. However, a key challenge in this task is dealing with class imbalances within the dataset, as certain mental health categories may be underrepresented compared to others.

To address this challenge and achieve accurate classification, the project employs several state-of-the-art NLP models, including LSTM, BERT, and RoBERTa. LSTM (Long Short-Term Memory) is a recurrent neural network architecture capable of learning long-term dependencies, making it suitable for sequential text data. BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based model that captures context from both directions, improving the understanding of contextual nuances in text. RoBERTa (Robustly Optimized BERT) is an optimized version of BERT that enhances performance through better pre-training strategies and more extensive training data. By leveraging these advanced models, the project aims to contribute to the development of effective tools for the early detection of mental health conditions, ultimately supporting mental health professionals and improving outcomes for individuals.

2. Shared Work Outline

To ensure an efficient and thorough approach to the project, our team strategically divided the tasks based on our strengths and the project's requirements. My primary role was to handle the implementation and analysis of the LSTM, BERT-base, and BERT-large models. This involved data preprocessing steps such as tokenization, padding, and cleaning the dataset to ensure compatibility with these models. I then trained each model and analyzed their performance using metrics like accuracy, precision, recall, and F1-score. After identifying the strengths and weaknesses of each model, I documented these findings to provide a basis for comparison with other approaches.

My teammate focused on enhancing the project by integrating RoBERTa-based models, known for their improved performance in NLP tasks. They tackled the challenge of class imbalances by employing techniques such as focal loss, which reduces the impact of easy-to-classify samples, and class weighting to ensure minority classes were adequately represented during training. Together, we collaborated on the final stages of the project, including evaluating all models side-by-side, generating detailed visualizations to highlight key results, and interpreting model outputs. We also worked jointly on developing a Streamlit interface to create a user-friendly platform where end-users can input data and interact with the models easily. This structured approach allowed us to deliver a comprehensive and accessible project, combining robust technical analysis with practical usability.

3. About Dataset

The dataset used in this project consists of 53,043 records, which were later cleaned to 51,060 records. It contains three columns: Unnamed: 0, statement, and status. The statement column contains patient statements or text inputs, while the status column represents the corresponding mental health status, categorized into seven distinct labels: Normal, Depression, Suicidal, Anxiety, Stress, Bi-Polar, and Personality Disorder. The dataset was compiled from a variety of sources, including social media platforms and Reddit, providing diverse and rich text data. The statements within the dataset vary in length, tone, and vocabulary, often containing misspellings, punctuation errors, and varying verbosity.

4. Data Preprocessing

Data preprocessing is a critical step in preparing raw data for model training. In this project, the text data required several preprocessing techniques to ensure the quality and relevance of the input for the machine learning models. First, the dataset was cleaned by removing unnecessary columns, such as the Unnamed: 0, and handling missing values by dropping rows with null entries. The text in the statement column was then cleaned to ensure uniformity: all text was converted to lowercase to eliminate inconsistencies, and punctuation and special characters were removed to reduce noise. Tokenization was performed to split the text into individual words, and stopwords (common words like "the", "is", "and") were removed as they do not contribute meaningfully to the classification task. To further standardize the text, lemmatization was applied to reduce words to their base form (e.g., "running" to "run"). Additionally, sentiment analysis was conducted using TextBlob to extract sentiment polarity from the statements, providing additional features that could help the model identify the emotional tone of the text. Linguistic features such as word count and sentence length were also extracted to add more context to the data. The final step involved splitting the dataset into training, validation, and test sets, ensuring that each subset maintained a representative distribution of the mental health status labels.

5. Data Splitting

To ensure the effective evaluation of the models, the dataset was divided into three distinct subsets: training, validation, and test. The data splitting was performed using a 70-15-15 ratio, meaning 70% of the data was used for training the models, while 15% was allocated for validation and 15% for testing. This split ensured that the models had sufficient data to learn from while also providing a separate set of data to validate model performance during training and evaluate generalization on unseen data. The splitting process also preserved the class distribution, ensuring that the training, validation, and test sets were representative of the overall dataset's imbalances. This was achieved using stratified sampling, which maintains the proportions of each mental health status label across all splits.

6. Model Training

LSTM (Long Short-Term Memory)

The LSTM model was initially used to capture sequential text dependencies. GloVe word embeddings were employed to convert the text into vector representations, which helped the LSTM model process the sequences of words. Basic text preprocessing steps, including stopword removal and lemmatization, were applied to clean and standardize the data before feeding it into the model.

Accuracy: The LSTM model achieved an accuracy of approximately 70% but faced challenges with the recall of minority classes. The LSTM's limited capacity to capture complex dependencies and lack of deep contextual understanding made it difficult to classify certain mental health statuses, particularly those from the minority classes (e.g., Bipolar and Suicidal).

BERT-base (Bidirectional Encoder Representations from Transformers)

In the second phase, the BERT-base model was used to leverage the transformer architecture, which is designed to better capture contextual relationships in text. Unlike LSTM, BERT processes the entire text simultaneously, understanding the context of each word in both directions (left-to-right and right-to-left). The model was fine-tuned with the dataset, and class weights were integrated into the loss function to address class imbalances in the dataset.

Accuracy: The BERT-base model achieved an accuracy of 79% and a Macro F1 score of approximately 76%. While there was a significant improvement in overall accuracy compared to LSTM, the model still struggled with the recall of minority classes. Overfitting was also observed, indicating that the model might have become too specialized in the training data, which impacted its ability to generalize to new, unseen examples.

BERT-large (Bidirectional Encoder Representations from Transformers)

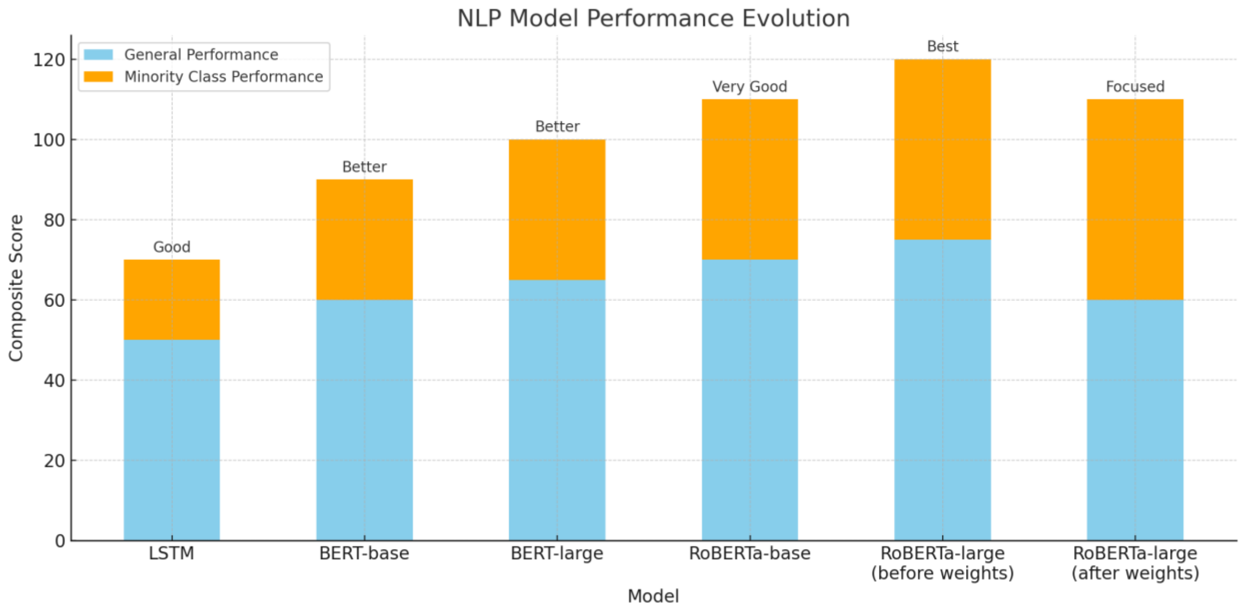
Finally, the BERT-large model was employed for its deeper learning capacity, as it has more layers and parameters compared to BERT-base. This larger architecture allowed the model to better handle complex contextual nuances and improve the representation of the input text.

Accuracy: The BERT-large model achieved an accuracy of 80% and a Macro F1 score of approximately 77%. Although the accuracy improved slightly compared to BERT-base, the improvement in recall for minority classes was modest. The increased capacity of the model allowed for better handling of contextual nuances, but overfitting continued to be a challenge.

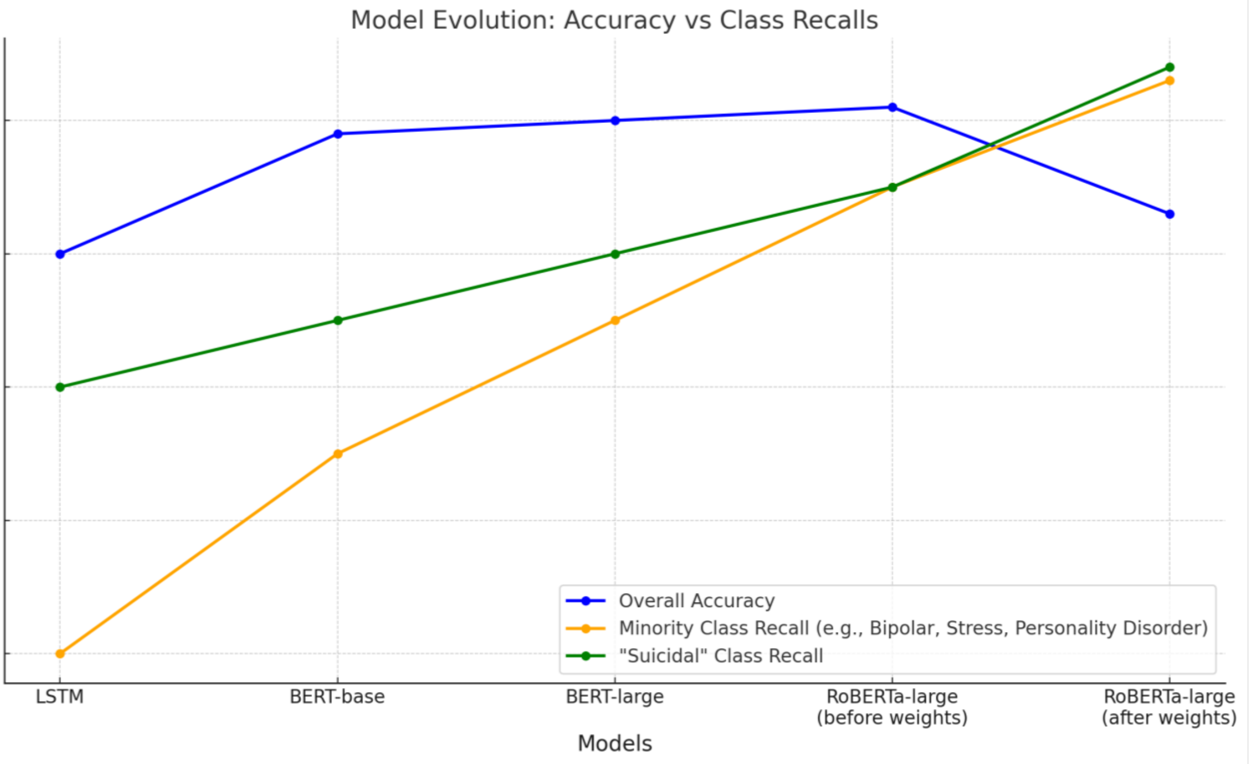
7. Final Model Comparison

After experimenting with multiple models, the performance improved progressively as we moved from LSTM to BERT-base and then to BERT-large. The LSTM model, which used GloVe word embeddings, achieved an accuracy of around 70% but struggled with capturing complex dependencies and understanding context, leading to poor performance on minority classes like Bipolar and Suicidal. Transitioning to BERT-base significantly improved accuracy to 79%, with a Macro F1 score of 76%, thanks to its transformer-based architecture. However, it still struggled with recall for minority classes and showed signs of overfitting. BERT-large further improved accuracy to 80% and Macro F1 score of 77%, particularly handling conditions like Stress and Personality Disorder better, though it still had limitations in recall for the minority classes.

The final models transitioned to RoBERTa, which provided further improvements. RoBERTa-base offered a slight improvement in accuracy (80%) and Macro F1 score of 77%, but the gains in handling minority classes were not transformative. The shift to RoBERTa-large with adjustments like extended training epochs and dropout layers yielded the best balance across all classes, with 81% accuracy, 78% Macro F1, and 82% Weighted F1, showing substantial improvements in minority class performance. However, to further prioritize minority class recall, RoBERTa-large (after class weights) was used, resulting in a decrease in overall accuracy (73%) but significant improvements in recall for minority classes like Bipolar, Personality Disorder, and Stress. This model, with its focus on recall for critical conditions, was selected as the final model, as it provided the best balance between precision and recall, ensuring more equitable and inclusive predictions for mental health statuses.



This stacked bar chart compares NLP model evolution using combined metrics of overall accuracy and minority class F1 scores, highlighting progressive improvements in handling class imbalances.



The chart demonstrates steady improvements in minority class recall (e.g., Bipolar, Stress, Personality Disorder) and mid-sized class recall ("Suicidal") as the models evolved, highlighting the impact of advanced techniques like class weights and focal loss.

While overall accuracy declined from 81% to 73%, significant gains in minority class recall (e.g., Bipolar: 78% → 87%, Personality Disorder: 66% → 79%) and "Suicidal" recall (71% → 84%) showcase the model's balanced approach to inclusivity and fairness.

8. Single-Label Classification and Multi-Label Adaptation

The model developed in this project excels as a single-label classification system, providing an effective solution for predicting a patient's mental health status from their statement. Single-label classification is particularly well-suited for this task, as it focuses on assigning one mental health category (e.g., Normal, Depression, Anxiety) to each patient based on their text input. This approach aligns well with clinical practices, where a diagnosis typically classifies a patient under a single primary condition at a given time. The model's performance, achieved through techniques like fine-tuning BERT-based models (such as BERT-base and RoBERTa), demonstrates a solid understanding of the context and nuances in patient statements, allowing it to accurately classify text into one of the predefined categories. The precision and recall metrics, especially for the majority classes, indicate that this approach is highly effective for identifying conditions that are more clearly defined and commonly observed in the dataset.

However, the model can also be adapted to multi-label classification, which offers several advantages. In real-world healthcare settings, many patients may experience more than one mental health condition simultaneously. For instance, a patient could have both Depression and Anxiety, which is common in clinical practice. Multi-label classification allows the model to predict multiple labels for a single input, enabling a more nuanced and realistic diagnosis. By using binary cross-entropy loss for each label and a sigmoid activation function, the model can output probabilities for each possible mental health condition independently, rather than choosing just one. This adaptation would significantly improve the model's ability to capture the complexity of co-occurring mental health conditions, making it more applicable in practice.

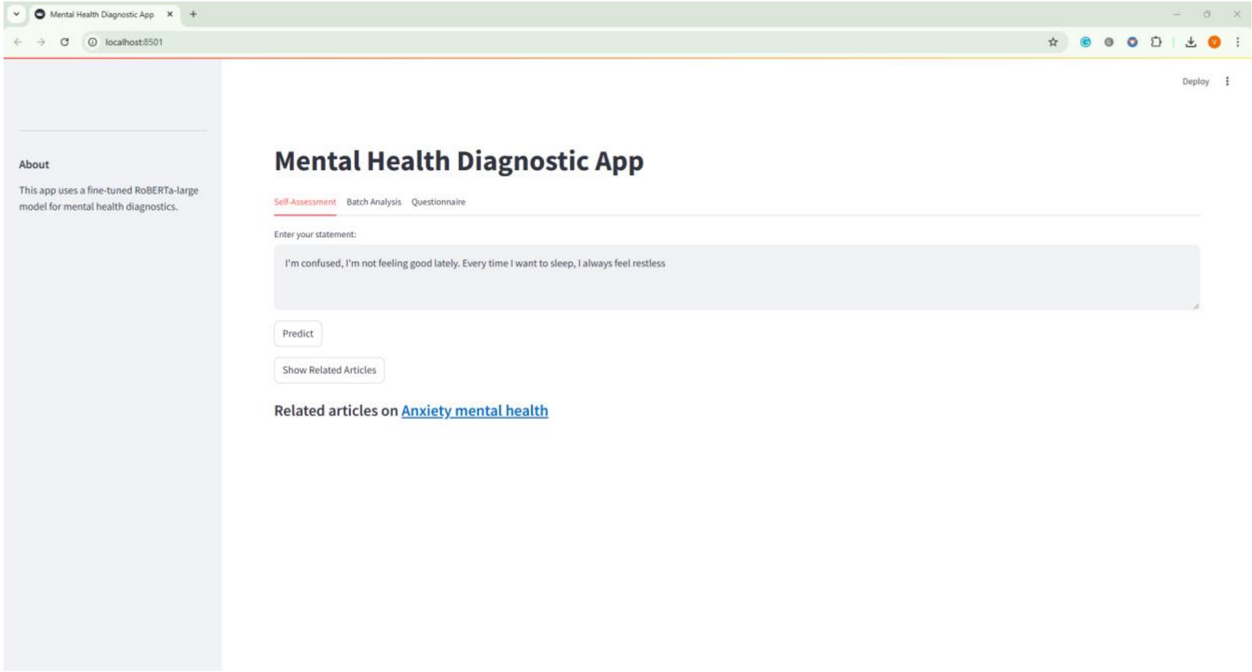
The main advantage of multi-label classification is that it offers a richer representation of a patient's mental health profile. It can provide a more complete view by acknowledging the possibility of multiple diagnoses. Additionally, it can help identify conditions that may not have been captured in single-label classification, thus improving early detection and treatment planning. Moreover, the flexibility of multi-label classification makes the model more adaptable to different use cases, such as personalized healthcare interventions, where patients' symptoms may span multiple categories and need to be addressed simultaneously.

That said, adapting to multi-label classification requires careful consideration of the evaluation metrics, as traditional metrics like accuracy may not be as meaningful. In multi-label settings, metrics like Hamming loss, precision, recall, and F1 score per label would need to be considered to ensure that the model is both precise and inclusive in its predictions. The challenges associated with multi-label classification, such as handling label dependencies and potential for overfitting, can be addressed with further model adjustments, such as the introduction of label correlation techniques or the use of more advanced loss functions.

In conclusion, while the current model provides a strong and effective solution for single-label classification, it holds great potential for adaptation to multi-label classification. This would allow it to better represent the complexity of real-world mental health conditions and further improve its applicability in clinical settings. The ability to predict multiple labels simultaneously offers the advantage of a more comprehensive understanding of patients' mental health, ultimately leading to better and more personalized care.

9. Streamlit – UI

Streamlit was used to build an interactive web application for mental health diagnostics, leveraging a fine-tuned RoBERTa-large model. The app provides users with an easy-to-use interface divided into three main sections: Self-Assessment, Batch Analysis, and Questionnaire. In the Self-Assessment tab, users can input a statement about their mental health, and the app predicts the corresponding mental health status (e.g., Depression, Anxiety) based on the model's classification. The Batch Analysis tab allows users to upload CSV files containing multiple statements and processes them in bulk, providing predictions for each entry along with visualizations like frequency distribution plots and word clouds. The Questionnaire tab engages users by asking random mental health-related questions and combining their responses to predict their mental health status.



About

This app uses a fine-tuned RoBERTa-large model for mental health diagnostics.

Mental Health Diagnostic App

Enter your statement:

I'm confused, I'm not feeling good lately. Every time I want to sleep, I always feel restless

Related articles on Anxiety mental health

The app is designed for ease of use, with interactive features such as text input fields, prediction buttons, and visual analytics that help users understand their mental health status more clearly. The app loads the pre-trained RoBERTa model and a label encoder to process inputs and generate predictions, while also integrating external news sources for further context. Libraries like Matplotlib and Seaborn are used for visualizing prediction distributions, and WordCloud generates insightful visual representations of common terms from input statements. Overall, Streamlit's intuitive interface and real-time predictions make the app a practical tool for personal mental health assessments and batch analyses in healthcare environments.

10. Conclusion

In this project, we developed and evaluated a series of models aimed at predicting mental health statuses based on patient statements. The focus was on single-label classification, where each statement was assigned one mental health status from a predefined set of categories, such as Depression, Anxiety, Bipolar, and Personality Disorder. Through the iterative development of models like LSTM, BERT-base, BERT-large, and RoBERTa, we demonstrated significant improvements in accuracy and recall, particularly for minority classes, such as Bipolar and Suicidal, which are often underrepresented in clinical datasets.

The project successfully showed how transformer-based models like BERT and RoBERTa can effectively capture contextual information from text, allowing for more accurate and nuanced predictions compared to traditional models like LSTM. While the final model, RoBERTa-large with class weights, achieved a balance between accuracy and recall for minority classes, there was a noticeable trade-off in precision for the majority classes. This trade-off highlights the challenge of balancing performance across all classes when dealing with imbalanced datasets.

Looking forward, the model could be adapted for multi-label classification, which would allow it to predict multiple mental health conditions simultaneously, offering a more comprehensive view of a patient's mental health status. This adaptation would improve the model's applicability in clinical settings where patients often experience co-occurring conditions. However, multi-label classification introduces new challenges, such as managing label dependencies and ensuring proper evaluation metrics.

In conclusion, this project provides a robust foundation for developing NLP models that assist in mental health prediction and diagnosis. The results suggest that transformer-based models, particularly RoBERTa, are well-suited for this task. While improvements can be made, especially in addressing the trade-offs between precision and recall, the potential for future research, including the exploration of multi-label classification, offers exciting opportunities for enhancing personalized mental health care.