Group Members: Anita Gee, Siwei Guo, Yingzhu Chen, Nemo (Sorachat) Chavalvechakul

Group Name: Mental Health Warriors

Dataset Partition Strategy:

The most common dataset partition strategy is 70-30 or 80-20. In our project, we process the dataset partition as 80-10-10: 80% for training, 10% for validation, and 10% for testing.

We made this decision based on the size of our dataset. After removing the missing value, there are 52,681 data points left in our dataset. With 80% of the data used for training, it will give the model around 42,144 pieces of data for learning. We want to make sure the model has enough examples to learn from, which is especially important considering the complexity of natural language and nuances in expressing emotional states.

10% of the data (roughly 5268) is used as the validation set during the training process. The validation set can help us to fine-tune hyperparameters and make decisions regarding model architecture. This step is also crucial for mitigating overfitting by providing regular feedback on unseen data.

The last 10% of the data (5268) will be used for the test set. We'll use this 10% data for model performance evaluation. This split can assess the model's generalization capability on a dataset that has not been used during the training process.

From our EDA findings, while the "Normal" class is the most frequent, the dataset still contains a significant portion of negative emotional states. The data split approach of 80-10-10 ensures all subsets remain representative of the overall class distribution, which is important for handling class imbalance during model development.

EDA Analysis and Insights:

In our exploratory data analysis (EDA) part, we processed a series of assessments to evaluate data quality, understand distribution patterns, and try to extract meaningful insight that could help with our model development.

The first step is data quality assessment, where we examined the dataset for missing values and duplicates. We found approximately 0.7% of the 'Statement' column contained missing values. We removed the missing values to maintain data integrity.

Our next step is to perform descriptive statistical analysis and distribution analysis, which can help us to better understand data structure. One key aspect we examined was statement length, measured by word count. We found that the mean of the statement contained approximately 113 words, but the distribution was right-skewed, with some statements exceeding 1,500 words. These long statements were often associated with depression and suicidal statuses, suggesting that individuals experiencing severe emotional distress tend to express themselves with longer and more complex statements. We also analyzed the category frequencies within the 'Status' field and noticed that while "Normal" was the most frequent status, there was still a large proportion of negative emotional statuses in the dataset.

We made several different graphics to visualize our findings to help the audience better understand. Graphics include bar plots and histograms to illustrate the frequency distribution of statement lengths and the count of each emotional status. The histograms confirmed the right-skewed nature of statement length distribution and showed that more severe emotional status is often related to longer text. Furthermore, we generated word clouds for the overall dataset and specific emotional statuses. The overall word cloud highlighted frequently used words such as "feel," "want," and "life," while the status-specific word clouds revealed unique patterns, with certain keywords like "anxiety" being prevalent in anxiety-related statements and "AVPD" frequently appearing in personality disorder-related statements.

To further understand contextual word usage, we conducted an n-gram analysis. We extracted bi-grams and tri-grams using a CountVectorizer with an n-gram range of (2,3). This analysis allowed us to identify common multi-word expressions such as "do not" and "feel like," providing deeper insight into how specific emotional states are conveyed through language. Unlike single-word analysis, n-grams captured important contextual nuances, particularly in expressions involving sentiment changes.

From these analyses, we discovered several key insights. First, even though common words appear across multiple emotional states, the way they are combined varies significantly, as seen in the n-gram analysis. This suggests that context is important in distinguishing between subtle emotional expressions. Second, data shows that statement length can serve as a potential indicator of emotional state, with longer statements often linked to more severe emotional distress. This highlights the importance of preprocessing strategies, such as segmenting lengthy texts, which can potentially better capture emotional nuances. Lastly, our analysis reveals a class imbalance in the dataset, with "Normal" being the dominant category. Although negative emotional statuses are well represented, this imbalance poses challenges for model training. We might need to consider more techniques, such as class weighting, resampling, or data augmentation, to address this issue and ensure the model can accurately recognize minority classes.

Insights from EDA Analysis:

Through EDA, we not only identified the significant characteristics of the data in terms of category distribution, text length, keyword usage, etc., but also discovered the commonalities and differences in users' expressions under different psychological states. Here are some insights from the team:

1. **The Multilayered Nature of Emotional Expression:**
    - From the word clouds and N-gram analysis, we observe that even across different emotional states, certain common words (e.g., "feel," "want") appear frequently. This suggests a shared linguistic pattern in how humans express their inner emotions. However, the key lies in the contextual nuances and collocations of these shared words within different emotional states. For instance, in the **Normal** state, "want" is likely associated with positive aspirations and plans. In contrast, in **Depression** or **Suicidal** states, the same word may carry a stronger sense of helplessness or despair, reflecting a yearning for relief from pain or a desire for redemption. Building on this insight, one potential research direction is to explore the contextual variations of these shared words across emotional states. By doing so, we can develop more refined emotional features that go beyond simple word frequency and capture the deeper contextual and semantic patterns underlying emotional expression.
2. **Semantic Refinement Inspired by N-gram Analysis:**
    - N-gram analysis reveals numerous phrase-level expressions (e.g., "do not," "feel like") that carry richer semantic and emotional information compared to individual words. Building on this insight, we propose an optimization approach similar to the word-level classification but focused on incorporating these phrases as key features. By combining these phrases with contextual embeddings, we aim to achieve more accurate classification. This approach not only captures the emotional tendencies of individual words but also identifies nuanced sentiments such as negation, doubt, or extreme emotions embedded within phrases. As a result, the model’s ability to discern complex emotional states could be significantly enhanced.
3. **Text Length and Emotional Depth Correlation:**
    - EDA reveals that longer texts are predominantly associated with Depression and Suicidal states, suggesting that these texts might convey more intricate or layered emotional expressions. They may encapsulate trajectories of emotional changes—for example, a progression from mild sadness to extreme despair. In the preprocessing and feature

extraction stages, a potential strategy could involve segment-level analysis to capture the dynamic shifts in emotional content within long texts, rather than simply truncating or padding them. This approach could help identify critical emotional turning points, providing richer contextual cues for the model. However, implementing this method poses additional challenges, such as maintaining coherence across segments and handling computational overhead effectively.

4. **The Real-World Implications of Class Imbalance:**
    - While the dataset shows that the "Normal" state accounts for the largest proportion, approximately 70% of user inputs fall into negative emotional categories. This imbalance is not only a technical challenge but also a reflection of the severity of mental health issues in society today. From a model training perspective, class imbalance may lead to suboptimal predictions for certain states. However, it also highlights the need to focus on relatively rare but higher-risk categories, such as Personality Disorder and Suicidal states, during model evaluation and intervention planning. While techniques like data augmentation and class weighting are essential, it is even more critical to design targeted intervention strategies during the application phase to address these high-risk states effectively.

Identification of Data Challenges and Strategic Recommendations for Data Preprocessing:

During the comprehensive exploratory data analysis (EDA), we meticulously identified several critical data integrity issues that necessitate immediate attention to ensure the robustness and reliability of our subsequent analyses:

1. **Presence of Missing Data:** Our preliminary exploration uncovered significant instances of missing entries across multiple variables within the dataset. This phenomenon can lead to substantial biases in predictive modeling, as it may distort the statistical representation of the population under study and impact the performance of the models.
2. **Duplicate Records:** The identification of repeated entries within our dataset presents a considerable challenge. Such redundancies not only inflate the dataset size artificially but also compromise the validity of any inferential statistics derived from the data, potentially leading to overfitting during the model training phase.

Recommendations for Rigorous Data Preprocessing:

In response to the aforementioned issues, we propose a series of methodical and technically sound data preprocessing interventions, aimed at enhancing the dataset's quality and suitability for the intricate task of modeling mental health statuses:

5. **Strategic Imputation of Missing Values:**
    - To counteract the effects of missing data, a robust imputation strategy is recommended. For numerical variables, statistical imputation techniques such as mean, median, or mode imputation are advisable depending on the distribution of the data. For categorical data, employing the most frequent category or a predictive imputation model could preserve the underlying relationships within the data.
6. **Elimination of Duplicate Records:**
    - A rigorous cleaning process should be implemented to detect and remove any duplicate entries. This will purify the dataset, ensuring that each data point uniquely contributes to the insights generated from the models, thus maintaining the integrity of our analytical outcomes.
7. **Advanced Text Normalization Techniques:**
    - Employ advanced text normalization techniques to standardize the textual data thoroughly. This includes converting all text to lowercase, removing non-alphanumeric characters, correcting typographical errors, and standardizing variations of the same words to a single common representation.
8. **Innovative Feature Engineering:**
    - Engage in creative feature engineering to unearth additional insights from the data. This could involve extracting new variables such as text length, word frequency, and sentiment scores, which may offer valuable predictive capabilities for the mental health status classifications.
9. **Data Transformation for Model Readiness:**
    - Transform the textual data into a structured format that is amenable to machine learning algorithms. Techniques such as vectorization through TF-IDF or embeddings from advanced models like BERT should be considered to encapsulate the semantic richness of the text effectively.

By embracing these meticulous preprocessing steps, we aim to fortify our dataset's foundational structure, thereby enhancing the accuracy and effectiveness of our predictive models in the realm of mental health analysis. These efforts are pivotal in transcending the traditional barriers of data quality to establish a

reliable and insightful analytical platform for mental health assessments based on social media interactions.

References

- World Health Organization (WHO). (2021). Mental Health: Strengthening Our Response. [online] Available at: https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response
- National Institute of Mental Health (NIMH). (2022). Mental Illness. [online] Available at: https://www.nimh.nih.gov/health/statistics/mental-illness
- Lancet Psychiatry, The. (2022). Economic Burdens of Mental Health Issues. [online] Available at: https://www.thelancet.com/journals/lanpsy/article/PIIS2215-0366(22)00123-9/fulltext