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

Group Name: Mental Health Warriors

Target Variable:

The principal dependent variable in our research is "mental health status," categorized within the dataset through labels such as Depression, Anxiety, Stress, and Normal. Our objective is to harness predictive algorithms to accurately determine these statuses from textual data on social media platforms.

Utilizing Natural Language Processing (NLP) and machine learning techniques, this study seeks to enhance diagnostic accuracy for mental health interventions, improving the efficacy of mental health chatbots and supporting resource redirection for users in need. Precise classification of mental health conditions from text is vital for improving public health monitoring and informing mental health policy.

Furthermore, this research enriches the dialogue on technology's role in mental health, assessing how advanced analytics can detect and interpret emotional expressions online. This contributes to developing models that are not only accurate but also sensitive in identifying complex psychological states.

This exploration underpins the academic and practical goals of a master's project at Boston College, aiming to leverage technology for better mental health understanding and intervention.

Predictors:

The predictors in our analysis include a range of features extracted from the text data, such as word frequency, sentiment scores, presence of specific keywords, and text complexity. These features are chosen because they provide significant insights into the emotional and psychological state conveyed by the text, which is crucial for accurate classification of mental health status. Advanced NLP techniques and machine learning algorithms will be employed to extract and utilize these predictors effectively, aiming to correlate specific textual patterns with particular mental health conditions.

In this project, we use the "Statement" column as the primary predictor variable. This column contains free-text data provided by users, and we aim to infer each individual's corresponding mental health status based on the information it contains. To improve prediction accuracy, we will extract various predictive features from the text. For example, the text length feature—obtained by counting the number of characters or words—can, to some extent, reflect the intensity of emotions or the degree of psychological stress. Additionally, we may employ sentiment analysis techniques to obtain sentiment scores and count

the occurrence frequency of specific keywords (such as "anxiety" and "stress"), among other features. These texts and features will be combined into feature vectors that are input into the prediction model. Once the model learns the associations between the text features and mental health status, it will be able to predict the mental health status of new text data, thereby classifying and assessing individuals' mental health conditions.

Exploration of the Dataset:

Our dataset comprises textual posts collected from various online platforms, labeled with corresponding mental health statuses. The dataset includes:

- **Data Types:** Categorical (labels), Text (posts).
- **Variable Definitions:** Each post is associated with a label indicating the mental health status.
- **General Stats:**
    - **Count of Rows:** 10,000 entries
    - **Count of Columns:** 2 columns (post, label)

Initial data exploration aims to assess the completeness and quality of the data, identify any missing values, and understand the distribution of different categories within the labels. This phase is critical for preparing the dataset for further processing and analysis, ensuring that the inputs into our predictive models are accurate and representative of the real-world scenarios they are meant to emulate.

- From the output, we can see that this dataset includes 2 variables: statement and status.
  The statement variable is a text variable that contains different user inputs.
  The status variable represents different emotional statuses that contain different categories.
- The dataset contains 362 missing values in the 'Statement' column and no missing values for 'Status'.
- The dataset includes 52,681 rows and 2 columns after removing missing values.
- We want to add a column to explore the length of each statement. This can help us quantify the user's input and support further analysis.
- The 'Statement' column contains 51,073 unique values, indicating that most user inputs are unique. The most frequently appeared statement is "What do you mean?" and occurred 22 times in the dataset.
  The 'Status' column contains 7 unique values and represents different emotion statuses. The most common status is "Normal", suggesting that over 30% of the statements in the dataset fall under this category.

- The summary statistics for the 'Statement_len' column show the distribution of statment lengths. The average statement contains 113 words with a standard deviation of 163.5 words. The shortest statement only has 1 word, while the longest contains 6300 words. The most frequent statement length is 5 words, indicating that short phrases are commonly used.
  The following bar plot of the frequency of statement length visualizes the previous statement.
- Here is a plot showing distribution by status. Normal is the most common status and contains 16343 data, followed by depression and suicidal, which are the 2nd and 3rd largest portions of the dataset. Personality disorder is the most rare one, which contains 1077 data.
  The ratio between different statuses suggests about 70% of the user's input falls under the negative status category.
- Here is the word cloud for Statement before data processing, which will be used to compare with the data after processing. The word cloud shows that the most frequently used words are "feel," "want," "know," and "life.". The observation is reasonable considering verbs and similar expressions that reflect personal thoughts would be the biggest part of user inputs. We can also see words like "depression," "tired," and "anxiety" in the word cloud even before data processing, which matches our observation of the status distribution above.

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