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

Group Name: Mental Health Warriors

Feature Engineering:

In our feature engineering process, we start with calculating the length of each statement in characters (`num_of_characters`) and the number of sentences (`num_of_sentences`) to capture structural information that might correlate with emotional intensity. The next step is to convert the preprocessed tokens into numerical features using TF-IDF, highlighting the importance of unique words or phrases. Finally, we combined these TF-IDF vectors with the numeric features (sentence length and character count) to form a richer, hybrid feature space, allowing the model to leverage linguistic and structural cues to classify mental health sentiments. Models such as XGBoost are excellent at handling numerical features like sentence and character length and can effectively use them in combination with other features.

Dimensionality Reduction:

- For dimensionality reduction in our analysis, we employed the TF-IDF (Term Frequency-Inverse Document Frequency) method, which, while not strictly reducing the number of features, effectively manages the complexity of our dataset by prioritizing more informative words. TF-IDF works by assigning weights to words based on their importance, which helps in downplaying less informative, common words (which often act as noise) and highlighting critical, sentiment-indicative terms that are less frequent across documents.
- By converting text data into a format where each word's significance is adjusted by its frequency across all documents, TF-IDF allows us to focus our analysis on the features that truly matter. Common words that appear in nearly all documents, such as "the" or "and," receive a lower score due to their high document frequency, reducing their impact on the analysis. Conversely, words that are more unique to certain documents (e.g., "depression" in a mental health context) gain higher scores and thus play a more significant role in our models.
- Additionally, to further reduce dimensionality, we removed redundant features such as the "number of words" column after finding it highly correlated with the sentence length and number of characters. This step helps minimize multicollinearity and enhances model efficiency by focusing on less but more significant features.

Oversampling:

- Class imbalance occurs when the distribution of categories in the target variable is skewed, which can lead to biased predictions from the model. By oversampling the minority class, we ensured that the model would be exposed to more examples of the underrepresented class, thereby improving its ability to learn and predict both classes effectively.
- To address severe class imbalance (e.g., 16,343 "normal" vs. 1,077 "personality disorder" instances) problems in our dataset, we applied random oversampling to the training data. This technique duplicated minority-class samples rather than synthesizing new data, ensuring authentic representation of rare emotional states. After oversampling, all seven classes achieved parity in sample counts (12,000+ instances each), enabling the model to learn distinctive patterns without majority-class bias.
- This approach helps improve performance metrics such as recall and F1-score, particularly for the minority class, leading to more robust and accurate model predictions. More importantly, oversampling was applied only to the training set post-train/validation split to prevent data leakage—validation and test sets retained original distributions for realistic performance evaluation.