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

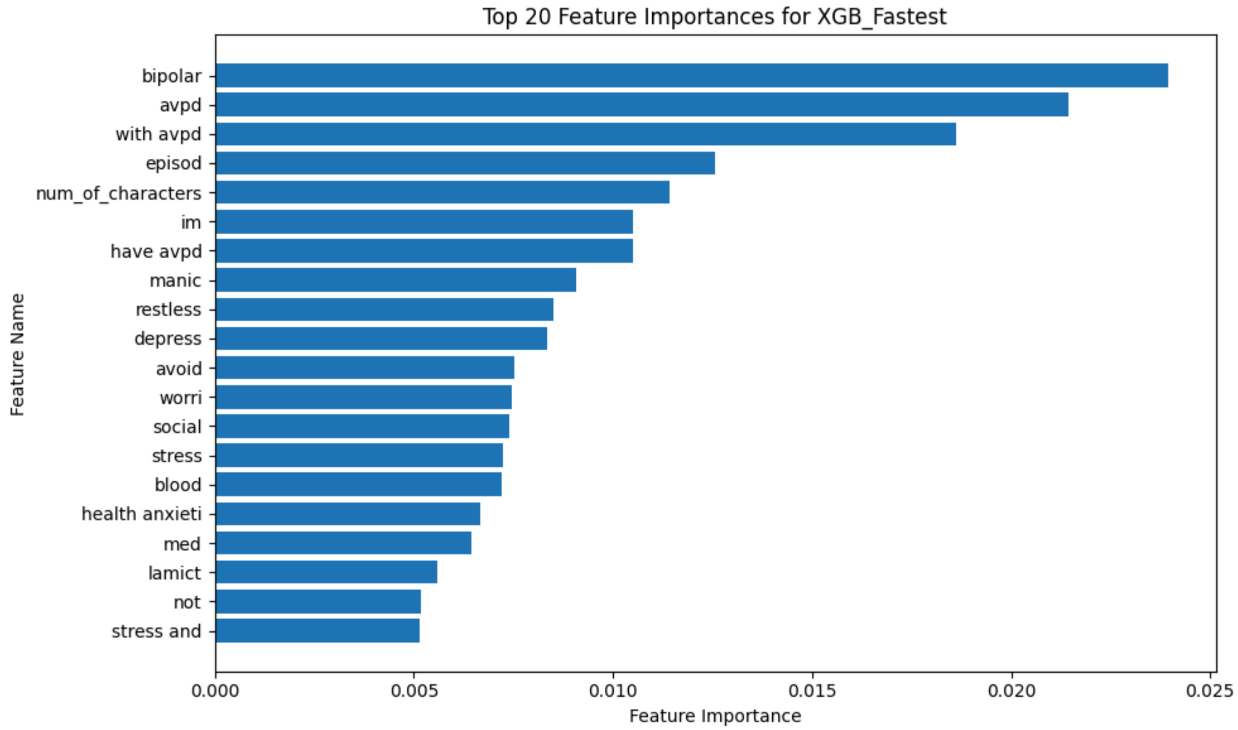
Group Name: Mental Health Warriors

Week 11 Report

Identification of Important Features in the Model

This analysis used the XGBoost model "XGB_Fastest" to classify mental health-related text data. The feature importance analysis shows which words or numerical features were most influential in helping the model make predictions.

According to the feature importance plot generated by the model, the top 10 most important predictors are as follows: bipolar, avpd, with avpd, episod, num_of_characters, im, have avpd, manic, restless, and depress. Among them, "bipolar" has the highest importance score. This means that if the word "bipolar" appears in a text, the model is highly sensitive to it and likely uses it as a strong signal for specific mental health categories, especially bipolar disorder.



Similarly, "avpd" and phrases containing "avpd", such as "with avpd" or "have avpd" are also highly influential. This indicates that the presence of these words strongly impacts the model's prediction related to personality disorders.

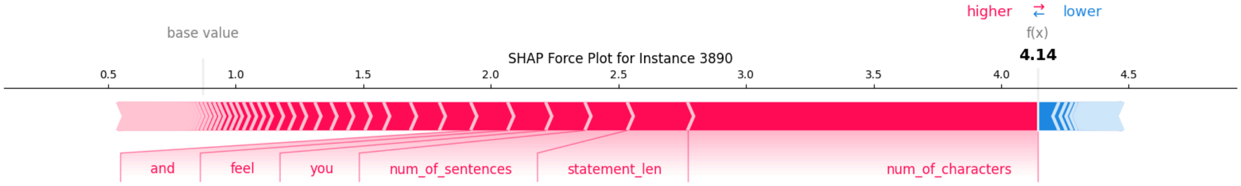
The feature "episod" is also important, probably because it relates to episodes of mood disorders. The numerical feature "num_of_characters" shows that the length of the input text also plays a role in prediction. Other keywords like "manic," "restless," and "depress" reflect common terms associated with mental health conditions and contribute significantly to the model's decisions.

Overall, most of the top features are directly related to mental health conditions or symptoms, suggesting that the model relies heavily on meaningful words and phrases in the text to classify mental health states.

Explanation of Five Random Predictions and Their SHAP Interpretations

We selected 5 random instances from the test dataset to better understand how the XGBoost model made its predictions in our mental health sentiment analysis task. SHAP values were used to explain which features contributed most to each prediction. Through these explanations, we can also explore which features and the value that need to be changed for us to move the output significantly or even flip the prediction to another class.

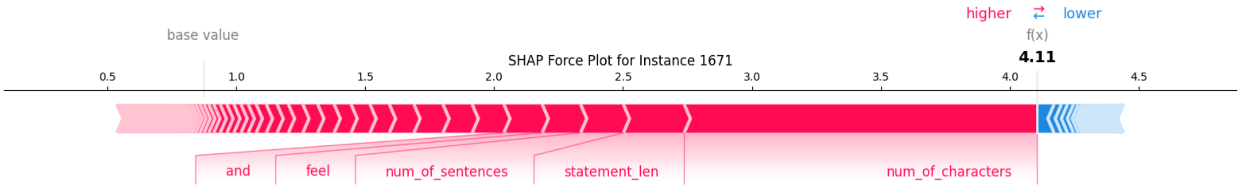
1. Instance 3890 (True Label: Normal, Prediction: Normal)
In this instance, the model correctly predicted "Normal." The most important feature was num_of_characters, which contributed a large positive SHAP value of 1.3718. This indicates that the length of the text strongly pushed the prediction upward. Other positive contributors included statement_len, num_of_sentences, you, and feel.
To flip this prediction from "Normal" to a mental health-related class like "Anxiety" or "Depression," the most effective strategy would be to introduce emotionally charged words like nervou or depress into the text. These words had shown strong positive contributions in other instances associated with mental health conditions. Simply increasing the length of the text is unlikely to be sufficient without these key emotional indicators.



2. Instance 1671 (True Label: Normal, Prediction: Normal)

This instance was again predicted as "Normal," and the pattern of feature contributions was very similar to Instance 3890. The strongest influence came from num_of_characters (1.3693), followed by statement_len, num_of_sentences, and words like feel and and.

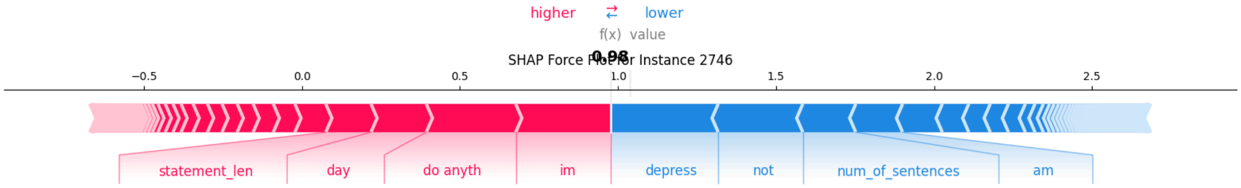
The fact that the model relies so heavily on general linguistic features like length and structure, rather than specific mental health terms, shows why the prediction remained "Normal." To flip this prediction, the most effective approach would be to add mental health-specific words such as depress, anxieti, or worri. Reducing the number of characters or sentence length would likely have only a minor effect.



3. Instance 2746 (True Label: Depression, Prediction: Depression)

In this instance, the true label and prediction were both "Depression." Interestingly, the most influential feature was depress, but its SHAP value was negative (-0.3412). This shows that in this specific context, the model learned that the appearance of depress alone might not always increase the prediction for depression, possibly due to its context.

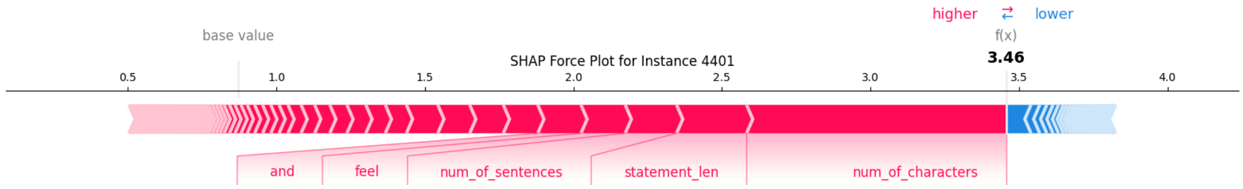
Positive contributors included im (0.2995), do anyth (0.2829), and day (0.1750). To flip this prediction away from "Depression," removing or changing words like im and do anyth would likely lower the prediction score significantly. Additionally, increasing the usage of neutral or positive words might further push the prediction towards "Normal."



4. Instance 4401 (True Label: Normal, Prediction: Normal)

This instance also had a true and predicted label of "Normal." The most important feature was again num_of_characters (0.8739), followed by statement_len, num_of_sentences, feel, and and. These are typical features contributing positively to the prediction score, but since no strong mental health-related terms appeared in this instance, the final prediction stayed in the "Normal" class.

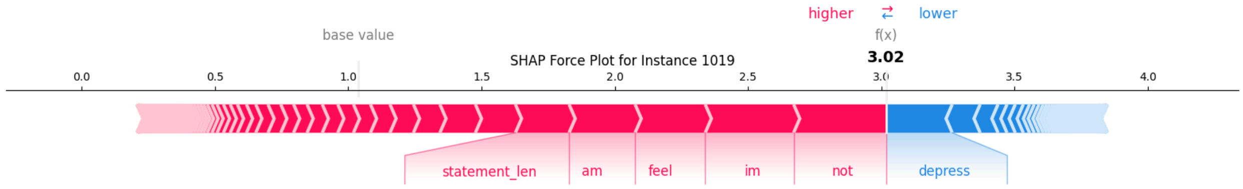
Compared to Instance 2746 ("Depression"), this example lacked features like depress or im that had shown large positive contributions to mental health classes. Therefore, to move this instance's prediction towards a class like "Anxiety" or "Depression," adding specific words such as nervou, worri, or depress would likely have the largest effect.



5. Instance 1019 (True Label: Depression, Prediction: Depression)

In this case, the true label was "Depression," and the model predicted it correctly. The most influential features pushing the prediction upwards were not (0.3465), im (0.3317), and feel (0.2632). On the other hand, the word depress had a negative SHAP value (-0.2486), similar to Instance 2746, showing again that its influence depends on context.

To flip this prediction from "Depression" to "Normal," removing or changing the words not, im, and feel would have the most significant impact. Additionally, increasing neutral or positive words and reducing the length of the text could help reduce the prediction score further.



These SHAP-based explanations demonstrate that the model's predictions are highly sensitive to a small number of key features, especially specific words strongly linked to mental health conditions. In most cases, changing the prediction from one class to another would require either removing these highly influential words or introducing new words that are strongly associated with the target class. While general features like text length or sentence count have some impact, it is ultimately the presence or absence of certain critical words that plays the most decisive role in shaping the model's output.

The Inclusion and Use of Protected Categories

The dataset used in this study does not explicitly include protected categories such as race, gender, religion, or national origin as structured input features. The target variable in our model is mental health status, which can be considered a form of disability under broader legal definitions. However, since this variable serves solely as the outcome to be predicted, and is not used as a training feature, it does not introduce a risk of discriminatory bias during model development.

It is worth noting that the free-text input, particularly the "statement" variable, may occasionally contain implicit references to protected characteristics. For example, certain phrases might mention personal identity, background, or belief systems. Nevertheless, exploratory analyses, including word clouds and feature importance rankings, indicate that the most influential features are primarily clinical or emotional in nature. Terms such as "bipolar," "avpd," "manic," and "depress" consistently appear among the most important predictors. This suggests that the model is primarily learning from linguistic patterns associated with mental health symptoms rather than demographic information. Consequently, while indirect bias cannot be fully ruled out, its impact in the current framework appears limited.

Bias and Fairness Analysis

In Week 11, we focused on evaluating and addressing potential bias in our XGBoost model, particularly concerning protected categories such as mental health diagnoses that may fall under disability classifications.

Bias Evaluation

We began by analyzing SHAP scores for our top five features, which showed that commonly used words such as "feel," "want," and "know" were the most predictive and not tied to any protected categories. This indicates a relatively low bias at the surface level. These top features were mainly derived from classes like Normal, Anxiety, Depression, and Suicidal, which are better represented in the dataset and have more general language.

However, a deeper look at the top 20 features by importance in XGBoost revealed a different story. Keywords such as "bipolar," "AVPD," "anxiety," and "stress" were heavily weighted. While these terms help with classification, they directly reflect diagnoses and conditions that are considered protected under disability law. Their dominance raises concerns about the model relying too much on direct mentions of mental health conditions, especially when such classes—like AVPD and Bipolar—are still underrepresented despite using SMOTE.

This was also visible in the word clouds, where terms like “bipolar” and “AVPD” appear prominently. The model may be learning to associate these specific words with their labels rather than understanding deeper linguistic patterns, which could result in biased predictions, especially when the key term is absent in the input.

Bias in Recall and Model Behavior

The recall scores highlight performance discrepancies across different mental health categories. On both the training and validation sets, the model showed significantly lower recall for Suicidal and Bipolar classes, even after applying SMOTE. For example, recall for Depression and Suicidal was relatively weak due to overlapping vocabulary with other emotional states. This indicates potential bias, as the model fails to detect minority or nuanced cases consistently.

Range of Recall: There is a significant range of recall (0.37) across the categories, with 'Normal' having the highest recall (0.95) and 'Personality disorder' having the lowest (0.58). This wide range indicates that the model is considerably better at identifying some emotional states compared to others, suggesting a bias in its detection capabilities.

Difference in Recall from 'Normal' (Baseline): Several mental health conditions show substantial deficits in recall compared to the 'Normal' category:

- Personality disorder: -0.37
- Stress: -0.32
- Suicidal: -0.28
- Depression: -0.23
- Bipolar: -0.18
- Anxiety: -0.13 These negative differences indicate that the model misses a significant proportion of actual cases for these conditions compared to the 'Normal' state. The large negative differences for 'Personality disorder', 'Stress', and 'Suicidal' are particularly concerning.

Ratio of Recall to Highest Performing ('Normal'): The recall performance relative to 'Normal' is low for several conditions:

- Personality disorder: 0.61 (Recall is only 61% of 'Normal')
- Stress: 0.66
- Suicidal: 0.71
- Depression: 0.76
- Bipolar: 0.81
- Anxiety: 0.86 These low ratios highlight that the model is considerably less effective at recalling instances of these mental health conditions compared to 'Normal'.

Standard Deviation of Recall: The standard deviation of 0.12 in recall scores across categories signifies considerable variability and inconsistency in the model's performance, suggesting bias.
Difference Between Weighted and Macro Average Recall: The weighted average recall (0.79) is higher than the macro average recall (0.73), with a difference of 0.06. This indicates that the model performs better on the more frequent categories in the dataset (likely 'Normal' and 'Depression'), suggesting a potential bias towards the majority classes due to having more training examples.

Bias Mitigation Strategies

To reduce these biases, we recommend the following improvements:

1. **More Advanced SMOTE**: While SMOTE improved minority class balance, applying it more aggressively to underrepresented classes (e.g., Personality Disorder and Bipolar) can reduce the model's dependence on specific keywords. This would allow the model to learn contextual patterns instead of overfitting to terms like “bipolar.” It can also potentially improve the recall for the minority classes. This means the model might become better at correctly identifying instances of these conditions, reducing false negatives, having a more balanced model performance and reducing bias towards majority classes such as normal or anxiety statuses. With more computational resources in the environment and/or the use of GPUs, we could accomplish this.
2. **Adjusting class weights during model training**: Adjusting class weights can directly counter this imbalance by making the model pay more attention to the minority classes during training. For example, the model could be penalized more for misclassifying a 'Personality disorder' instance than a 'Normal' instance if the weight for 'Personality disorder' is set higher. This could potentially improve the recall for these underrepresented categories and might become better at correctly identifying instances of conditions like 'Personality disorder', 'Stress', and 'Suicidal', reducing false negatives. It's important to note that increasing the weights of minority classes might lead to a slight decrease in performance (e.g., precision or recall) for the majority classes.

The model might make more false positives for the majority class as it tries harder not to misclassify the minority classes. However, the overall goal is often to achieve a better balance and improved performance on the minority classes, especially in a sensitive domain like mental health where failing to identify a critical condition can have serious consequences. With XGBoost, we can use scale_pos_weight parameter and the ability to use custom loss functions to focus on minority class performance.

3. **Add Contextual Features**: Incorporating sentence structure, emotional tone, or linguistic style as features may help the model differentiate between mental health classes with similar vocabulary, leading to fairer classifications.

4. **Regularization and Dropout**: Increasing regularization in the model can prevent over-reliance on single features. This helps avoid memorizing protected-category keywords and instead promotes generalized learning.

Conclusion

Our model shows limited surface-level bias, but further analysis reveals potential risks associated with relying on condition-specific words, particularly in underrepresented categories. The recall imbalance confirms this concern, especially in sensitive classes like Suicidal and Bipolar. By refining our SMOTE implementation and incorporating fairness-aware strategies, we can reduce bias and improve the model's real-world reliability for equitable mental health detection.

Risks of Using the Model for Stakeholders

Deploying this sentiment analysis model in real-world settings, especially those involving mental health, raises several ethical and operational risks for various stakeholders. For individuals, the risk of false negatives may lead to serious cases of distress being overlooked, thereby missing opportunities for timely intervention. Conversely, false positives could result in unnecessary concern or stigmatization, especially if the model is used in sensitive contexts such as workplace evaluations, online content moderation, or mental health screening.

The model's performance across mental health categories is not uniform. Notably, it shows lower recall for underrepresented categories such as "personality disorder," "stress," and "suicidal." This inconsistency may lead to unequal treatment of different psychological conditions, thereby reinforcing existing disparities in care or resource allocation. If adopted by public health systems or government institutions, this imbalance may further perpetuate systemic inequities.

Furthermore, the model is trained on English-language social media data, which may reflect cultural and linguistic biases inherent to the source. Without careful localization or fairness assessments, applying the model to broader, more diverse populations could compromise its accuracy and fairness. This could ultimately erode trust in predictive tools used in mental health domains and have negative implications for public perception, regulatory compliance, and ethical responsibility.