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

Group Name: Mental Health Warriors

Week 8 Report

Model Approach and Its Suitability

For this week's sentiment analysis task in mental health, we employed the XGBoost algorithm as our primary modeling approach. We chose XGBoost due to its effectiveness in handling high-dimensional, sparse data, which is common in text-based sentiment classification. The dataset contained unstructured text from online discussions, where slang, abbreviations, and informal expressions frequently appear. XGBoost's robustness to noisy data made it particularly well-suited for this problem, as it can learn meaningful patterns even with inconsistencies in language use.

Another reason for choosing XGBoost is its capability to handle class imbalance through the scale_pos_weight parameter. Mental health sentiment datasets often have uneven class distributions, with specific sentiments being expressed far more frequently than others. By tuning this parameter based on the inverse ratio of class frequencies, we ensured that underrepresented sentiment classes were not overshadowed by more common ones during model training. Additionally, given that interpretability is a critical factor in mental health applications, XGBoost’s built-in feature importance metrics enabled us to analyze which linguistic patterns had the most influence on sentiment predictions. This transparency is crucial when working with sensitive topics, as it allows us to verify that the model's predictions are aligned with meaningful textual indicators rather than spurious correlations.

Complexity of the Modeling Approach

We carefully balanced the complexity of our model to ensure both efficiency and predictive performance. While deep learning methods such as transformers or LSTMs could improve accuracy, they often require extensive computational resources and large labeled datasets for training. Given our dataset's size and the need for faster iteration cycles, XGBoost provided an ideal alternative, offering strong predictive power without excessive computational overhead.

To enhance XGBoost’s ability to capture relevant patterns in textual data, we incorporated several feature engineering techniques beyond the basic TF-IDF transformations. First, we applied N-gram expansion (up to bi-grams) to preserve local contextual relationships between words, improving the model’s ability to recognize sentiment cues across adjacent terms. Second, we performed dimensionality reduction using Truncated SVD (sklearn.decomposition.TruncatedSVD) to reduce the high-dimensional sparse TF-IDF

matrix while preserving the most important information. This helped prevent overfitting and improved computational efficiency.

Hyperparameter tuning also played an important role in refining the model's complexity. We experimented with different tree depths to balance learning capacity and generalization. By capping the maximum tree depth at 3, we prevented the model from overfitting to minor variations in training data while still allowing it to capture essential sentiment patterns. We also adjusted column and row subsampling rates to introduce randomness in the learning process, thereby improving the model's robustness and reducing its susceptibility to noise.

Hyperparameters Evaluated and Regularization

To optimize performance, we evaluated three variations of the XGBoost classifier: XGB_Conservative, XGB_Faster, and XGB_Fastest. Each variation was designed with different hyperparameter settings to examine their effects on training efficiency and predictive accuracy.

The XGB_Conservative model employed a lower learning rate (0.05), stronger regularization (reg_alpha=2, reg_lambda=2), and a reduced subsample ratio (0.6). These settings prioritized stability and generalization, making the model less prone to overfitting but potentially slower in convergence.

The XGB_Faster model increased the learning rate to 0.1, slightly reduced regularization (reg_alpha=1, reg_lambda=1), and raised the subsample ratio to 0.7. This configuration balanced speed and generalization, allowing for a more efficient training process while maintaining good model robustness.

The XGB_Fastest model used the most aggressive learning rate (0.2) while maintaining similar subsampling and regularization settings as XGB_Faster. This model aimed for the fastest convergence, albeit with a higher risk of overfitting.

Through our evaluation, the XGB_Fastest model demonstrated the highest validation accuracy (0.77), making it the best-performing model. However, we closely examined the difference between training and validation accuracies to assess potential overfitting. Ensuring the discrepancy remained within an acceptable range, we confirmed that the model was neither underfitting nor excessively overfitting.

Performance Metrics and Their Relevance

Accuracy:

- Definition: The proportion of correctly classified instances.

- Relevance: In imbalanced datasets, a high accuracy can be misleading. For instance, if 95% of the data is "non-distressed" and 5% is "distressed," a model predicting all instances as "non-distressed" achieves 95% accuracy, rendering it useless for identifying at-risk individuals.
- Use: More reliable in balanced datasets (achieved through methods like random oversampling), but should not be the sole evaluation metric.

Precision:

- Definition: The proportion of correctly predicted positive instances out of all instances predicted as positive.
- Relevance: High precision for critical negative classes (e.g., suicidal ideation) is crucial. It ensures that when the model flags someone as at risk, it is likely correct, minimizing unnecessary interventions.
- Importance: Prioritizes the reliability of positive predictions.

Recall (Sensitivity):

- Definition: The proportion of correctly predicted positive instances out of all actual positive instances.
- Relevance: High recall is vital for capturing as many true positive cases as possible. In mental health, missing a person at risk (false negative) can have severe consequences.
- Importance: Prioritizes the detection of all at-risk individuals.

F1-Score:

- Definition: The harmonic mean of precision and recall.
- Relevance: Provides a balanced measure of a model's performance, especially in imbalanced datasets, offering a compromise between precision and recall.
- Importance: Offers an overall evaluation of model effectiveness.

Classification Report (Per-Label Metrics):

- Relevance: Provides precision, recall, and F1-score for each class (e.g., "normal," "anxiety," "depression," "suicidal").
- Importance: Essential for understanding how the model performs on different mental health states, particularly critical in imbalanced datasets.

Why These Metrics Matter in Mental Health:

- **Prioritizing Recall:** In mental health, missing a case of severe distress is often worse than a false alarm.
- **Improved Accuracy Relevance:** We did random oversampling which creates a more balanced dataset, making accuracy a more reliable metric.
- **Balancing Precision and Recall:** The F1-score ensures a balance between accuracy and comprehensiveness.
- **Understanding Per-Class Performance:** The classification report enables targeted improvements for specific mental health states.
- **Addressing Imbalanced Data:** Precision, recall, and F1-score are crucial for evaluating models on imbalanced mental health datasets.

XGBoost works well with mental health sentiment analysis for these reasons:

- **Complex Data Handling:** Mental health text data is often nuanced and irregular. XGBoost's ability to capture intricate patterns within this complex data leads to more accurate sentiment classification.
- **Noise Robustness:** Real-world data, especially from online sources, is riddled with noise and inconsistencies. XGBoost's inherent robustness allows it to perform well despite these imperfections.
- **High Predictive Accuracy:** The need for precise sentiment analysis in mental health applications is paramount. XGBoost's high predictive accuracy ensures that subtle shifts in emotional state are detected.
- **Feature Importance Analysis:** By revealing the most influential words and phrases, XGBoost provides valuable insights into the linguistic markers of different mental health states.
- **Gradient Boosting Optimization:** The gradient boosting framework iteratively refines the model, resulting in superior performance.
- **Enhanced by Deep Learning:** XGBoost works extremely well when combined with modern deep learning models such as BERT, which create very accurate text embeddings. This hybrid approach allows for even higher accuracy.

Calculate the Metrics for Training and Validation Datasets:

- **Procedure:** For each of the three model variations, use sklearn.metrics to compute key performance indicators such as accuracy, precision, recall, and F1-score on both training and validation datasets.
- **Tools:** Employ functions like accuracy_score, precision_score, recall_score, classification_report, and f1_score from sklearn.metrics to obtain these metrics.
- **Data Handling:** Ensure data is appropriately split and preprocessed to reflect true performance on unseen data in the validation set.

2. Metrics Calculation Across Three Variations:

The metrics are calculated as follows for both training and validation:

- Training Accuracy: Assesses how well the model learned the patterns in the training data.
- Validation Accuracy: Measures the model's ability to generalize to unseen data.
- Confusion Matrix: Provides a detailed breakdown of true positive, true negative, false positive, and false negative predictions for each class for both training and validation data.
- Classification Report: Includes precision, recall, F1-score, and support for each class, offering insights into the model's performance on individual classes for both training and validation data.

Discussion and Analysis of Metrics Across Variations

Here are the training and validation accuracies, along with the macro average recall:

- **XGB_Conservative**:
    - **Training Accuracy**: 79.65%
    - **Validation Accuracy**: 74.09%
    - **Macro average recall:** 0.76
- **XGB_Faster**:
    - **Training Accuracy**: 84.25%
    - **Validation Accuracy**: 76.03%
    - **Macro average recall:** 0.78
- **XGB_Fastest**:
    - **Training Accuracy**: 89.25%
    - **Validation Accuracy**: 77.36%
    - **Macro average recall Validation:** 0.78

Since we want to focus on recall here are the recall metrics for each label for each XGBoost Model on the validation set:

XGB_Conservative:

- Training Recall:
    - Anxiety: 0.81
    - Bipolar: 0.80
    - Depression: 0.58
    - Normal: 0.90
    - Personality disorder: 0.85
    - Stress: 0.88
    - Suicidal: 0.75
- Validation Recall:
    - Anxiety: 0.81
    - Bipolar: 0.76
    - Depression: 0.56
    - Normal: 0.88
    - Personality disorder: 0.72
    - Stress: 0.85
    - Suicidal: 0.72

XGB_Faster:

- Training Recall:
    - Anxiety: 0.86
    - Bipolar: 0.86
    - Depression: 0.63
    - Normal: 0.91
    - Personality disorder: 0.94
    - Stress: 0.92
    - Suicidal: 0.78
- Validation Recall:
    - Anxiety: 0.82
    - Bipolar: 0.80

    - Depression: 0.61
    - Normal: 0.89
    - Personality disorder: 0.74
    - Stress: 0.83
    - Suicidal: 0.73

XGB_Fastest:

- Training Recall:
    - Anxiety: 0.92
    - Bipolar: 0.94
    - Depression: 0.70
    - Normal: 0.92
    - Personality disorder: 0.99
    - Stress: 0.97
    - Suicidal: 0.82
- Validation Recall:
    - Anxiety: 0.83
    - Bipolar: 0.83
    - Depression: 0.65
    - Normal: 0.89
    - Personality disorder: 0.72
    - Stress: 0.84
    - Suicidal: 0.73

Discussion and Analysis of Metrics Across Variations

XGB_Conservative:

- This model, characterized by a low learning rate (0.05) and strong regularization (reg_alpha=2, reg_lambda=2), generally shows **lower training recall** compared to the other two XGBoost models. For instance, the training recall for 'Anxiety' is 0.81, for 'Depression' is 0.58, and the macro average training recall is 0.80.
- The **validation recall** for XGB_Conservative is also relatively lower, with a macro average of 0.76 and a weighted average of 0.74. Specific validation recalls include 'Anxiety' at 0.81 and 'Depression' at 0.56.

- The **difference between the training and validation recall** is generally smaller for XGB_Conservative compared to the other two models (e.g., macro average difference is 0.04). This suggests that this model might be less prone to overfitting due to the stronger regularization and lower learning rate.

XGB_Faster:

- With a moderate learning rate (0.1) and reduced regularization (reg_alpha=1, reg_lambda=1), XGB_Faster exhibits **higher training recall** across most classes compared to XGB_Conservative. The training recall for 'Anxiety' increases to 0.86, and for 'Depression' to 0.63. The macro average training recall is 0.84.
- The **validation recall** also shows an improvement, with a macro average of 0.78 and a weighted average of 0.76. The validation recall for 'Anxiety' is 0.82 and for 'Depression' is 0.61.
- The **gap between training and validation recall** widens slightly (macro average difference of 0.06), indicating a potential increase in overfitting compared to XGB_Conservative, although still within a reasonable range.

XGB_Fastest:

- This model, with the highest learning rate (0.2) and the same regularization as XGB_Faster, achieves the **highest training recall** among the three models. The training recall for 'Anxiety' reaches 0.92, for 'Depression' 0.70, and the macro average training recall is 0.89.
- The **validation recall** is also the highest, with a macro average of 0.78 and a weighted average of 0.77. The validation recall for 'Anxiety' is 0.83 and for 'Depression' is 0.65.
- However, XGB_Fastest shows the **largest difference between training and validation recall** (macro average difference of 0.11), indicating a higher risk of overfitting. This is expected with a higher learning rate and relatively weaker regularization, as the model might learn the training data too well, including noise, and thus perform relatively worse on unseen data.

Overall Trends and Analysis:

- **Increased Learning Rate and Reduced Regularization Lead to Higher Training Recall:** As we move from XGB_Conservative to XGB_Faster and then to XGB_Fastest, the learning rate increases, and the regularization strength decreases. This generally correlates with an increase in the training recall across most of the status categories. The models learn the training data more effectively with these changes.

- **Validation Recall Improvement Plateaus and Overfitting Risk Increases:** While the validation recall also tends to improve from XGB_Conservative to XGB_Fastest, the magnitude of improvement is smaller than in the training set. Moreover, the gap between training and validation recall widens, particularly for XGB_Fastest, suggesting that the gains in training performance might be coming at the cost of reduced generalization ability due to potential overfitting.
- **Trade-off Between Bias and Variance:** XGB_Conservative, with its lower learning rate and strong regularization, likely has higher bias (underfitting the training data to some extent) but lower variance (better generalization). Conversely, XGB_Fastest likely has lower bias (fitting the training data very well) but higher variance (overfitting, leading to a larger performance drop on the validation set). XGB_Faster appears to strike a better balance between bias and variance.
- **Class-Specific Variations:** The changes in recall are not uniform across all classes. For example, the recall for 'Normal' remains relatively high and consistent across all three models in both the training and validation sets. However, for classes like 'Depression' and 'Personality disorder', the recall values fluctuate more significantly with the model variations, highlighting the difficulty in accurately classifying these statuses.

Compare performance metrics for the validation dataset across all 3 variations

In our evaluation, we compared the validation performance of three XGBoost models with different hyperparameter configurations. The XGB_Conservative model achieved a validation accuracy of 0.74, with a macro-average precision, recall, and F1-score of approximately 0.68, 0.76, and 0.70, respectively. The XGB_Faster model demonstrated an improved validation accuracy of 0.76, with macro-average metrics of around 0.70 (precision), 0.78 (recall), and 0.72 (F1-score). Finally, the XGB_Fastest model attained the highest validation accuracy of 0.77, along with macro-average precision, recall, and F1-score values of approximately 0.71, 0.78, and 0.73. Although the XGB_Fastest model exhibits a larger gap between training and validation performance—suggesting some degree of overfitting—the validation metrics indicate it performs best on unseen data.

Provide a table that shows 3 variations, training and validation values for all metrics you chose

Model	Training Accuracy	Validation Accuracy	MacroAvg Precision (Val)	MacroAvg Recall (Val)	MacroAvg F1-score(Val)
XGB_Conservative	0.80	0.74	0.68	0.76	0.70

XGB_Faster	0.84	0.76	0.70	0.78	0.72
XGB_Fastest	0.89	0.77	0.71	0.78	0.73

Identify the best model for the week and discuss why you selected that model

After analyzing the performance metrics, the XGB_Fastest model emerges as the best-performing model based on the validation dataset. Although the XGB_Fastest model exhibits a higher training accuracy of 0.89 compared to its validation accuracy of 0.77 (yielding an accuracy gap of 0.12), it nevertheless achieves the highest validation accuracy among the three variations. This indicates that, despite a potential risk of overfitting, the model generalizes well enough to produce the best performance on unseen data.

Furthermore, the XGB_Fastest model also demonstrates superior macro-average metrics in terms of precision, recall, and F1-score on the validation set, which are critical for ensuring balanced performance across all classes. In the context of mental health sentiment analysis, where detecting subtle cues is paramount, a high recall is especially important to avoid missing high-risk cases. Thus, even though the model may be slightly overfitted, its high validation accuracy and robust recall make it the optimal choice for our prediction task.

In summary, based on the comprehensive evaluation of training and validation accuracies, as well as the macro-average performance metrics, XGB_Fastest is selected as the best model for this week because it achieves the highest validation accuracy (0.77) and exhibits strong overall performance, making it well-suited for handling the complexities of mental health sentiment analysis.