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

Group Name: Mental Health Warriors

Week 10 Report

Project Overview

In our ongoing project, we aim to enhance the predictive capabilities of a machine learning model designed to classify mental health statements. This involves refining our approach by addressing data quality, enhancing feature engineering, and balancing the dataset. The model, originally developed in Week 9 using XGBoost, showed promising but improvable results, particularly in handling diverse and imbalanced data. Our focus for Week 10 has been on implementing strategic improvements to the data and the model itself, ensuring it can effectively generalize to new, unseen data while maintaining robust performance across various metrics.

Improvements to the Data

In Week 10, our goal was to enhance the predictive accuracy and reliability of our model through strategic data improvements. Building upon the groundwork laid in Week 9, we implemented three novel approaches to refine our training dataset:

- **Adding Statement Length as a Feature**: The length of each statement (in terms of the number of words) was added as a new numerical feature. This was done because **longer statements might contain more context and emotional depth, potentially improving sentiment classification**, while shorter statements could be more ambiguous. By including this feature, the model can learn patterns between the length of the text and the expressed sentiment.

- **Using SMOTE Instead of Random Oversampling**: The Synthetic Minority Over-sampling Technique (SMOTE) was employed to address the class imbalance in the dataset. **SMOTE creates synthetic examples of the minority classes** instead of simply duplicating existing samples, as was done with random oversampling previously. This technique can be particularly helpful for improving the prediction of less frequent categories like 'Stress' and 'Suicidal'. By generating synthetic data points for the underrepresented classes, **SMOTE aims to reduce bias towards the majority classes and provide the model with sufficient data to learn from all categories**.

- **Removing Outliers from the Data**: Outliers in the statement lengths were identified and removed from the `df_unique` dataframe to improve the consistency of the dataset. Outliers might represent noisy, irrelevant, or misclassified text, which could potentially confuse the model. **Removing these extreme values can lead to a more balanced dataset**, especially since the outliers tended to be from the 'Depression' and 'Normal' categories, which already had a large number of samples. This step helps the model focus on more representative sentiment patterns.

These three techniques were implemented with the goal of enhancing the quality and balance of the dataset, thereby leading to more accurate and fair sentiment predictions for mental health analysis. The results showed that the XGBoost model trained on the data improved using these techniques performed better than the model from the previous week, exhibiting higher validation accuracy, a better balance between training and validation accuracy (less overfitting), and a higher average recall score.

Error Analysis and Data Improvements

The error analysis provided detailed insights into the model's performance across different mental health categories. The model performed very well on the training set for Personality Disorder (TP = 12,599) and Stress (TP = 12,330), showing that it could correctly identify most samples in these classes. However, it struggled more with Depression and Suicidal, where false negatives were higher. Specifically, Depression had 3,871 false negatives, and Suicidal had 2,288, indicating that many true cases were missed by the model.

This trend continued on the validation set. For example, Depression had 536 false negatives and Suicidal had 272. These classes often share similar language with others, making them harder to distinguish. Although the model performed well on more common classes like Normal (Validation TP = 1,482) and Anxiety (Validation TP = 298), identifying less frequent conditions remained a challenge.

To address these issues, we applied several data-centric improvements. One of the main changes was adding statement length as a new feature. Longer statements often carry more context and emotional cues. By including this feature, the model gained more information to better separate categories like Depression and Suicidal, where subtle differences are important.

We also replaced random oversampling with SMOTE, which generated synthetic samples for minority classes. For example, Anxiety was increased to 6,328 samples, Bipolar to 5,229, Personality Disorder to 3,804, and Stress to 5,224. This reduced the class imbalance and helped the model learn from underrepresented categories without simply duplicating existing data.

Another key improvement was removing outliers. These were mostly found in the Normal and Depression classes, which had a large number of samples to begin with. By removing them, we reduced noise and allowed the model to focus on more meaningful examples. This likely helped reduce false positives in some categories. For example, the Bipolar class had 59 false positives on the validation set, and the Stress class had 566—both improved after cleaning the data.

These improvements had a measurable impact. The XGBoost model trained with the new techniques achieved a validation accuracy of 0.80 and a test accuracy of 0.79, both higher than the baseline XGBoost model's validation accuracy of 0.77. The weighted average false negative rate also dropped to 0.21 on the test set, the lowest among all models tested.

By focusing on the data rather than just tuning the model, we were able to significantly improve its ability to detect more challenging mental health conditions while maintaining overall performance.

Refitting the Model

Using the enhanced training dataset, we refitted our XGBoost model from Week 9. The refitting process involved:

- **Procedure:** use sklearn.metrics to compute key performance indicators such as accuracy, precision, recall, and F1-score on both training and validation datasets. We also used the same parameters from the best XGBoost model we selected from Week 9.
- **Tools:** Employ functions like accuracy_score, precision_score, recall_score, classification_report, and f1_score from sklearn.metrics to obtain these metrics on the new model from the new training, validation and test dataset.
- **Data Handling:** Ensure data is appropriately split and preprocessed to reflect true performance on unseen data in the validation set using the three new ways we cleaned the data.

Model Comparison and Final Selection

The enhanced XGBoost model (Week 10) outperformed the best Week 9 model across key metrics:

Metric	Week 9 Model	Week 10 Model	Improvement
Validation Accuracy	0.77	0.80	+0.03
Weighted Recall	0.77	0.80	+0.03

False Negative Rate	0.23	0.20	-0.03

We compared the enhanced XGBoost model from Week 10 to the best-performing model from Week 9 using the updated validation dataset. Across all major metrics, the Week 10 model performed better and demonstrated improved generalization and class balance.

In terms of validation accuracy, the Week 10 model achieved 0.80, compared to 0.77 in Week 9, marking a 0.03 improvement. The weighted recall also increased from 0.77 to 0.80, indicating that the model was able to correctly identify more true cases overall. Additionally, the false negative rate decreased from 0.23 to 0.20, an important gain given the context of mental health, where missing a true case can have serious consequences.

One of the reasons the Week 10 model performed better was its ability to handle class imbalance more effectively. The use of SMOTE allowed us to generate synthetic samples for underrepresented classes like Anxiety, Bipolar, and Stress, improving the model's ability to detect those categories. For example, the number of false negatives for Suicidal decreased from 272 in Week 9 to 222 in Week 10. Similarly, Depression false negatives also dropped from 536 to 222, showing better classification for these critical classes.

The Week 10 model also showed reduced overfitting. In Week 9, the training accuracy was 0.89, while validation accuracy was 0.77, creating a noticeable gap of 0.12. In contrast, the Week 10 model had a training accuracy of 0.88 and validation accuracy of 0.80, reducing the gap to only 0.08. This suggests the model generalized better to unseen data.

Finally, the model was also evaluated on the test set, where it achieved a test accuracy of 0.79, which is consistent with its validation performance. This reinforces the conclusion that the Week 10 model is more stable and reliable.

Given its higher accuracy, lower false negative rate, and improved performance on key mental health categories, the XGBoost model from Week 10 is selected as the final model to be deployed. It offers a stronger balance between precision and recall, and is better suited for real-world applications where identifying all at-risk individuals is a top priority.

Performance Metrics

Using the final model that combines XGBoost with three data-centric AI techniques, we evaluated the model on the test dataset. The following performance metrics were obtained:

Data	Accuracy	Weighted Precision	Weighted Recall	Weighted F1	FPR	FNR
Training	0.88	0.88	0.88	0.88	0.12	0.12
Validation	0.80	0.80	0.80	0.80	0.20	0.20
Test	0.79	0.78	0.79	0.78	0.22	0.21

The test error is slightly higher than the validation error but very close, suggesting good generalization. The gap between training and validation accuracy (8%) and between validation and test accuracy (1%) also confirms that the model is not significantly overfitting and performs reliably on unseen data.

Insights from Training, Validation, and Test Errors

Analyzing the errors across the training, validation, and test datasets reveals several important insights about the performance and reliability of the final model. First, the consistent and moderate drop in performance from training accuracy (0.88) to validation (0.80) and test accuracy (0.79) indicates that the model is not significantly overfitting. This small performance gap suggests that the model generalizes well to unseen data, which is a critical requirement for robust machine learning systems.

Second, the model demonstrates strong and balanced predictive capability. Both the weighted precision and recall remain above 0.78 across all datasets, reflecting a good balance between minimizing false positives and false negatives. This is especially crucial in mental health sentiment analysis, where failing to identify at-risk individuals (false negatives) can have serious consequences, and excessive false positives can reduce trust in the system.

Third, the effectiveness of data-centric AI techniques is evident. The inclusion of SMOTE for synthetic oversampling, removal of outliers, and the addition of statement length as a feature all contributed to improved classification performance, particularly for underrepresented classes like "Suicidal" and "Stress."

Overall, the model's stable performance across datasets, combined with its low false negative rate and improved handling of imbalanced data, indicates that it is well-prepared for deployment in real-world mental health applications. Its ability to generalize, along with the robustness of its predictions, makes it a reliable choice for sensitive sentiment detection tasks.