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

Group Name: Mental Health Warriors

Week 6 Report

Model Approach and Its Suitability

This week, we implemented logistic regression as our primary modeling approach for NLP sentiment analysis in mental health.
Logistic regression is suitable to use as a benchmark for this task because it is both simple and efficient, which allows us to train and interpret our models quickly. Logistic regression can offer clear insights into the influence of each feature, which is critical when working with text data that has been transformed into high-dimensional TF-IDF vectors, like the data we had. By leveraging logistic regression, we are able to evaluate the importance of various textual cues and engineered numerical features, such as sentence length and character count, which are useful in predicting sentiment. It also supports regularization techniques (L1 and L2) to prevent overfitting, which is crucial when dealing with high-dimensional text data.

Furthermore, logistic regression naturally extends to multi-class problems using the one-vs-rest (OvR) strategy, where we train separate binary classifiers for each sentiment class. This approach not only provides a robust baseline but also facilitates comparisons between different regularization techniques. Overall, the balance of interpretability, efficiency, and performance offered by logistic regression makes it an ideal choice for our first model used in mental health sentiment analysis.

Complexity of the Modeling Approach

We intentionally keep the first modeling approach simple, this allows us to process fast experimentation and clear interpretability of the results. While more advanced models, such as neural networks or transformer-based architectures, might capture more nuanced patterns in text, logistic regression offers a strong balance between performance and simplicity. Its computational efficiency is particularly beneficial when working with high-dimensional data, such as the TF-IDF features derived from our tokenized and stemmed text data.

In our implementation, we combined both textual and numerical features to enrich the input space. The TF-IDF vectorization transforms the processed text into a sparse, high-dimensional representation that captures the importance of words and bi-grams across the dataset. Besides this, we engineered additional numerical features like sentence length and character count, which provide structural insights that are often lost when just relying on textual data. Merging these two feature sets increases the complexity of the input, but it allows us to model the data effectively in both the linguistic and structural aspects.

Another layer of complexity is introduced through the data augmentation step. We applied random oversampling to address class imbalance in our training set. By duplicating minority class instances, we ensured that the model received a balanced view of all classes during training, which is important for learning robust patterns. This oversampling step is performed after the train/validation split to prevent data leakage, which further underscores our commitment to maintaining the integrity of our evaluation process.

Finally, although the logistic regression model itself is straightforward, the tuning of hyperparameters and regularization methods adds further depth to our approach. We experimented with different regularization strategies (L1 and L2) and varying the strength of regularization via the C parameter. This experimentation helps us control overfitting, especially given the high dimensionality of TF-IDF features. The simplicity of the logistic regression model, combined with these adjustments in data preparation and hyperparameter tuning, allows us to achieve competitive performance while retaining interpretability—a key advantage when dealing with sensitive topics such as mental health sentiment analysis.

Hyperparameters Evaluated and Regularization

We evaluated three logistic regression models, each of them with different regularization strategies. For us to better understand their impact on model performance.

The first model we used is L1 regularization with a C value of 10, which helps in feature selection by shrinking some coefficients to zero. This sparsity-inducing property of L1 regularization is very useful when dealing with high-dimensional data, such as the TF-IDF features derived from text. By eliminating less important features, the model becomes more interpretable and may reduce the risk of overfitting.

The second model we employed is L2 regularization, also with a C value of 10. Unlike L1, L2 regularization shrinks the coefficients but does not force them to be exactly zero. This results in a model where all features contribute to the prediction, albeit with reduced magnitude. Our evaluation showed that Model 2 struck a favorable balance between training performance and generalization on the validation set. The L2 penalty appears to handle the complexity of the TF-IDF feature space more robustly by preventing the model from overly relying on any single feature.

The third model revisits L1 regularization but with a stronger regularization strength, a lower C value of 5. The increased regularization force further reduces the magnitude of coefficients, potentially discarding more features than Model 1. This aggressive regularization can help mitigate overfitting in scenarios where the model might otherwise capture noise from the training data. However, our comparisons indicate that while the stronger L1 regularization in Model 3 reduces training accuracy, it does not necessarily lead to better generalization compared to the L2 approach in Model 2.

The choice to evaluate these specific hyperparameters was driven by the high dimensionality inherent in TF-IDF feature representations and the need to balance model complexity with interpretability. By comparing training and validation metrics, we observed that Model 2, with its L2 regularization, demonstrated less overfitting and better generalization performance. This detailed hyperparameter tuning provides valuable insights into how different regularization strategies impact the model's ability to perform well on unseen data, ensuring that the selected approach is both robust and effective for our mental health sentiment analysis task.

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.

Metric Calculations and Analysis

1. 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.

3. Discuss and Analyze How Metrics Change Across Variations:

To compare the training and validation accuracy for each model and provide observations:

Logistic Regression 1:

- Training Accuracy: 0.998
- Validation Accuracy: 0.739

Logistic Regression 2:

- Training Accuracy: 0.886
- Validation Accuracy: 0.758

Logistic Regression 3:

- Training Accuracy: 0.986
- Validation Accuracy: 0.751

Since we want to focus on recall, we will list the recall scores between the training and validation datasets for each model.

- Logistic Regression 1:
    - Training Recall:

        - All classes (Anxiety, Bipolar, Depression, Normal, Personality disorder, Stress, Suicidal) achieved a recall of approximately 1.00.
    - Validation Recall:
        - Anxiety: 0.80
        - Bipolar: 0.79
        - Depression: 0.66
        - Normal: 0.90
        - Personality disorder: 0.57
        - Stress: 0.51
        - Suicidal: 0.62
- Logistic Regression 2:
    - Training Recall:
        - Anxiety: 0.88
        - Bipolar: 0.91
        - Depression: 0.69
        - Normal: 0.92
        - Personality disorder: 1.00
        - Stress: 0.98
        - Suicidal: 0.82
    - Validation Recall:
        - Anxiety: 0.82
        - Bipolar: 0.83
        - Depression: 0.63
        - Normal: 0.89
        - Personality disorder: 0.63
        - Stress: 0.59
        - Suicidal: 0.73
- Logistic Regression 3:
    - Training Recall:
        - Anxiety: 1.00
        - Bipolar: 1.00
        - Depression: 0.94
        - Normal: 0.99
        - Personality disorder: 1.00

        - Stress: 1.00
        - Suicidal: 0.98
    - Validation Recall:
        - Anxiety: 0.82
        - Bipolar: 0.80
        - Depression: 0.66
        - Normal: 0.91
        - Personality disorder: 0.55
        - Stress: 0.55
        - Suicidal: 0.64

Observations:

- **Overfitting**: Models 1 and 3 have high training accuracy but lower validation accuracy, suggesting overfitting. Models 1 and 3 also show high training recall (often close to 1.00) but lower validation recall, indicating overfitting. This indicates the models learned the training data too well but do not generalize effectively to unseen data.
- **Model 2's Balance**: Model 2 with L2 regularization shows a smaller difference between training and validation accuracy compared to Models 1 and 3. Model 2 also has lower training recall compared to Models 1 and 3 (both have L1 regularization), but its validation recall scores are generally higher than Model 1 and comparable to Model 3.This suggests a better balance between fitting the training data and generalizing to new data.
- **Best Model**: Based on the accuracy scores, **Model 2 appears to be the best** because it achieves a reasonable balance between training and validation performance.
- Class-Specific Performance: All models struggle with "Personality disorder" and "Stress" classes, as indicated by their lower recall scores in the validation sets compared to other classes.
- Best Model: Based on the recall scores, Model 2 appears to be the best because it achieves a reasonable balance between training and validation performance, with better generalization capabilities than Models 1 and 3.

Compare performance metrics for the validation dataset across all 3 variations

Among the three logistic regression variations, Model 2 (with L2 regularization, C=10) achieved the highest validation accuracy (75.76%) and a higher average validation recall (~73.1%), compared to Model 1 (validation accuracy ~73.94% and average recall ~69.0%) and Model 3 (validation accuracy

~75.13% and average recall ~70.4%). Additionally, Model 2 has a smaller gap between training and validation performance, suggesting better generalization and less overfitting.

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

Model Variation	Training Accuracy	Validation Accuracy	Avg. Training Recall	Avg. Validation Recall
Logistic Reg 1	0.998	0.739	1.00	0.69
Logistic Reg 2	0.886	0.758	0.89	0.73
Logistic Reg 3	0.986	0.751	0.99	0.70

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

Based on the validation metrics, Logistic Regression 2 is the best model for this week. It has the highest validation accuracy and the highest average validation recall among the three models, indicating a better balance between fitting the training data and generalizing to unseen data. Additionally, its smaller training-validation gap suggests reduced overfitting, making it the most robust choice for our sentiment analysis task.