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

Group Name: Mental Health Warriors

Week 7 Report

Model Approach and Its Suitability

This week, we implemented Bernoulli Naive Bayes as our primary modeling approach for NLP sentiment analysis in mental health. Bernoulli Naive Bayes is suitable for this task due to its design, which handles binary feature representations effectively. This makes it ideal for text classification when working with word presence indicators rather than word frequency. In mental health contexts, the presence of keywords like "anxious," "hopeless," and "overwhelmed" can indicate distress, regardless of frequency.

The model is also computationally efficient, capable of quickly training even on high-dimensional text data. This efficiency is particularly valuable for large, sparse datasets, which are common in text classification. Furthermore, Bernoulli Naive Bayes works well with imbalanced datasets, such as those in mental health, where some sentiment categories may be underrepresented. Its simplicity and interpretability allow us to easily identify which words contribute most to classification decisions.

To optimize performance, we evaluated the alpha hyperparameter, which controls Laplace smoothing. Smoothing prevents the model from assigning zero probabilities to unseen words. We tested alpha values of 0.1, 1.0, and 10.0. Lower alpha values make the model more sensitive to rare words, while higher values prevent overfitting by distributing probability mass more evenly. Additionally, we set binarize=0.0 to ensure the model focuses on word presence rather than frequency, aligning with Bernoulli Naive Bayes' assumptions.

Complexity of the Modeling Approach

We chose a straightforward modeling approach to allow rapid experimentation and clear interpretation of results. While more complex models like deep learning architectures could potentially capture intricate linguistic patterns, the Bernoulli Naive Bayes model offers a good balance between simplicity and effectiveness. Its computational efficiency is essential when dealing with high-dimensional text data, as it requires minimal resources.

We transformed the text data into binary features for Bernoulli Naive Bayes. Unlike TF-IDF, which accounts for word frequency, this method focuses strictly on word presence, simplifying the input representation while preserving sentiment-indicating terms. Since the model assumes feature independence, it does not capture contextual relationships between words, reducing computational complexity but limiting its ability to discern deeper semantic patterns.

Data preprocessing involved standard text-cleaning techniques, such as lowercase and stopword removal, to enhance feature extraction. Unlike more advanced NLP methods that use word embeddings or contextual models, Bernoulli Naive Bayes relies on presence-based features, making it a lightweight and interpretable choice.

Hyperparameter tuning added complexity. By experimenting with different alpha values, we aimed to find the optimal balance between overfitting and underfitting. Although Bernoulli Naive Bayes lacks regularization features like logistic regression, adjusting alpha helps fine-tune the model's sensitivity to infrequent terms.

Hyperparameters Evaluated and Regularization

This week, we explored the effect of alpha on Laplace smoothing in the Naive Bayes classifier. We tested three alpha values: 0.1, 1.0, and 10.0.

Alpha = 0.1: minimal smoothing, allowing the model to be more sensitive to rare words. This can be useful when certain mental health-related terms are infrequent but important. However, it may lead to overfitting if rare words dominate.

Alpha = 1.0: Default smoothing, offering a balance between sensitivity and generalization. This value reduces overfitting while capturing important patterns.

Alpha = 10.0: Stronger smoothing, reducing the influence of rare words and distributing probabilities more evenly. While preventing overfitting, it may cause the model to overlook important, rare terms.

We also set binarize=0.0, focusing on word presence rather than frequency, which aligns with Bernoulli Naive Bayes' assumptions. By testing different alpha values, we aimed to strike a balance between preserving the significance of rare words and avoiding 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.

Bernoulli Naive Bayes (BNB) works well with the mentioned performance metrics in mental health:

- Binary Data Fit: BNB handles yes/no symptoms effectively.
- Recall Priority: Its probability output allows threshold adjustments to prioritize finding all at-risk individuals (high recall).
- Imbalanced Data: BNB can use adjusted prior probabilities to mitigate the effects of imbalanced datasets.
- F1-Score Balance: Threshold tuning helps balance precision and recall, optimizing the F1-score.
- Classification Reports: BNB provides probabilities for detailed per-class performance analysis.
- Accuracy Limitations: It goes beyond simple accuracy by providing probability based outputs, and allowing for threshold tuning.
- Speed and efficiency: BNB is fast, which is important for timely interventions

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.

Discussion and Analysis of Metrics Across Variations

Here are the training and validation accuracies, along with the macro average recall, for each Bernoulli Naive Bayes classifier:

- **Bernoulli NB (alpha=0.1)**:
    - **Training Accuracy**: 0.749739393599731
    - **Validation Accuracy**: 0.6389269629919718
    - **Macro Average Recall**: 0.75
- **Bernoulli NB (alpha=1.0)**:
    - **Training Accuracy**: 0.7137476881690299
    - **Validation Accuracy**: 0.6316820050910515
    - **Macro Average Recall**: 0.71
- **Bernoulli NB (alpha=10.0)**:
    - **Training Accuracy**: 0.6401950344672981
    - **Validation Accuracy**: 0.6052476992363423
    - **Macro Average Recall**: 0.64

Bernoulli NB (alpha=0.1) has the highest validation accuracy (0.6389) and that it had the highest recall scores for more of the classifiers (4 highest recall values).

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

- **Bernoulli NB (alpha=0.1)**:
    - **Training Recall**:
        - **Anxiety**: 0.70
        - **Bipolar**: 0.79
        - **Depression**: 0.54
        - **Normal**: 0.93
        - **Personality disorder**: 0.79
        - **Stress**: 0.88
        - **Suicidal**: 0.63
    - **Validation Recall**:
        - **Anxiety**: 0.58
        - **Bipolar**: 0.57
        - **Depression**: 0.49
        - **Normal**: 0.92
        - **Personality disorder**: 0.10
        - **Stress**: 0.46
        - **Suicidal**: 0.52
- **Bernoulli NB (alpha=1.0)**:
    - **Training Recall**:
        - **Anxiety**: 0.65
        - **Bipolar**: 0.73
        - **Depression**: 0.49
        - **Normal**: 0.93
        - **Personality disorder**: 0.75
        - **Stress**: 0.85
        - **Suicidal**: 0.60
    - **Validation Recall**:
        - **Anxiety**: 0.57
        - **Bipolar**: 0.61

- **Depression**: 0.45
- **Normal**: 0.92
- **Personality disorder**: 0.28
- **Stress**: 0.56
- **Suicidal**: 0.51

- **Bernoulli NB (alpha=10.0)**:
    - **Training Recall**:
        - **Anxiety**: 0.57
        - **Bipolar**: 0.61
        - **Depression**: 0.39
        - **Normal**: 0.94
        - **Personality disorder**: 0.66
        - **Stress**: 0.77
        - **Suicidal**: 0.54
    - **Validation Recall**:
        - **Anxiety**: 0.52
        - **Bipolar**: 0.51
        - **Depression**: 0.37
        - **Normal**: 0.94
        - **Personality disorder**: 0.42
        - **Stress**: 0.66
        - **Suicidal**: 0.48

Observations and Analysis:

- **Impact of Alpha on Training Recall**: As the alpha value increases (from 0.1 to 1.0 to 10.0), the training recall generally tends to decrease for most of the negative sentiment classes (Anxiety, Bipolar, Depression, Personality disorder, Stress, Suicidal). This suggests that **higher smoothing (larger alpha) can lead to a model that is less certain about the presence of specific words as indicators for these classes during training**. For the "Normal" class, the training recall remains high and relatively stable across all alpha values.
- **Impact of Alpha on Validation Recall**: Similarly, the validation recall for most of the negative sentiment classes also tends to decrease as the alpha value increases. This indicates that **higher smoothing might hinder the model's ability to correctly identify these classes on unseen data**. Again, the "Normal" class shows high and stable validation recall.

- **Generalization Performance (Training vs. Validation Recall)**:
    - For **alpha=0.1**, we observe a noticeable drop in recall from the training set to the validation set for several classes, particularly Personality disorder (0.79 to 0.10) and Stress (0.88 to 0.46), suggesting **potential overfitting** where the model has learned the training data too well and struggles to generalize to new data for these less frequent classes.
    - For **alpha=1.0**, the drop in recall from training to validation is less extreme compared to alpha=0.1 for most classes, especially Personality disorder (0.75 to 0.28) and Stress (0.85 to 0.56), indicating **better generalization than with lower smoothing**.
    - For **alpha=10.0**, the training recall is generally lower, and the validation recall is also among the lowest for many negative sentiment classes. While the difference between training and validation recall is not as large as with alpha=0.1, the overall performance on the validation set is weaker, suggesting that **high smoothing might lead to underfitting** where the model is too generalized and fails to capture the specific characteristics of each class.
- **Performance on "Normal" Class**: The recall for the "Normal" class remains consistently high (around 0.92-0.94) in both training and validation sets across all alpha values. This could be due to the fact that "Normal" is the most frequent status in the dataset, which might make it easier for the model to learn and predict correctly regardless of the smoothing parameter.

- **Performance on Less Frequent Classes**: The recall values for less frequent classes like Personality disorder and Stress are generally lower and more sensitive to the alpha value, particularly on the validation set. This highlights the challenge of accurately classifying minority classes, as mentioned in the source regarding the imbalanced dataset.

Conclusion:

The choice of the alpha value in Bernoulli Naive Bayes significantly impacts the model's performance. A very low alpha (like 0.1) can lead to overfitting on the training data, especially for less frequent classes, resulting in poor generalization. A very high alpha (like 10.0) can lead to underfitting, where the model is too smooth and fails to capture the nuances of different sentiment classes, leading to lower recall on both training and validation sets. An intermediate alpha value (like 1.0) appears to strike a better balance between overfitting and underfitting for most classes, leading to a more robust generalization performance, which identifies Bernoulli NB (alpha=0.1) as the best model based on the highest validation accuracy and recall for more classifiers. However, our analysis of the recall values suggests that while alpha=0.1 has the highest overall validation accuracy, it also shows signs of more significant overfitting for certain classes compared to alpha=1.0

Model Comparison Table:

Model Alpha	Training Accuracy	Validation Accuracy	Precision	Recall	F1-score
0.1	0.75	0.64	0.62	0.52	0.53
1.0	0.71	0.63	0.61	0.56	0.56
10.0	0.64	0.61	0.59	0.56	0.54

Insights from the Comparison:

- Bernoulli NB (alpha=0.1) has the highest validation accuracy (63.89%), meaning it performs best at predicting unseen data.
- Bernoulli NB (alpha=10.0) has the lowest validation accuracy (60.52%), likely due to excessive smoothing.

Precision, Recall, and F1-score:

- Bernoulli NB (alpha=0.1) has the highest recall (0.52), meaning it is better at capturing the minority classes.
- Bernoulli NB (alpha=1.0) has the highest precision (0.61) and an F1-score of 0.56, suggesting a better balance between false positives and false negatives.

Analysis and Conclusion:

The model with Alpha=0.1 performs best on the validation set, but the performance on the training set is significantly higher than that on the validation set, which may be at risk of overfitting. The model with Alpha=10.0 has good generalization, but the overall accuracy is low, which may not fully learn text features. The alpha=1.0 model delivered the best validation performance overall. Its accuracy on the validation set was the highest, and both precision and recall were strong across the board, leading to the top F1-score of the three models. Importantly, the training accuracy was very close to the validation accuracy, indicating that the model is not overfitting but generalizing well to new data. Although the accuracy of the validation set is slightly lower than alpha=0.1, it is better in generalization ability and suitable for practical applications.