6FTC2085-0901-Advanced Artificial Intelligence

NLP - Project

Omar Faisal Amr Mourad - 21026670



Problem Statement:

Mental health issues among students are a growing concern, making it essential to identify those who may need support. This project focuses on developing a system that analyzes students' social media statements to classify them as either **Normal** or **Abnormal**. Statements labeled as **Abnormal** indicate mental health challenges such as **Anxiety**, **Depression**, or **Stress**. The dataset, **"Combined Data.csv"**, includes text statements with corresponding mental health statuses, forming the basis for this classification task.

The project involves cleaning and preparing the text data through preprocessing techniques such as removing punctuation, links, and numbers. Labels are assigned to the data, creating a binary classification problem where the system predicts whether a statement indicates potential mental health issues.

Approach:

To build the system, the text data was converted into numerical form using a **Bag-of-Words (BoW)** model. This allows machine learning models to process and classify the text. Two models were implemented:

1. **Logistic Regression:** A linear model used for binary classification that predicts the probability of a data point belonging to a particular class. It was optimized with regularization to prevent overfitting and evaluated on training, validation, and test data.
2. **Naive Bayes:** A probabilistic model based on Bayes' theorem, assuming independence between features. It was trained after addressing class imbalance using SMOTE (Synthetic Minority Oversampling Technique), which balances the dataset by creating synthetic samples for minority classes.

The models were evaluated using metrics such as accuracy, log loss, confusion matrices, and F1-scores. Additionally, precision-recall curves were used to find the optimal thresholds for predictions. The system also includes a feature to predict the mental health status of new statements interactively.

Performance on Unseen Data:

Logistic Regression:

The Logistic Regression model performed well, achieving a **test accuracy of 92.4%** and a **log loss of 0.3724**. It reliably differentiated between normal and abnormal statements, with only a few misclassifications.

```
Test Accuracy (Logistic Regression): 0.9259
Test Log Loss (Logistic Regression): 0.2235

Test Classification Report (Logistic Regression):
              precision    recall  f1-score   support

           0       0.86      0.94      0.89      2672
           1       0.97      0.92      0.94      5285

    accuracy                           0.93      7957
   macro avg       0.91      0.93      0.92      7957
weighted avg       0.93      0.93      0.93      7957
```

Naive Bayes:

The Naive Bayes model, after applying SMOTE, achieved a **test accuracy of 88.2%** and a **log loss of 0.4785**.

```
Test Accuracy (Naive Bayes): 0.8822
Test Log Loss (Naive Bayes): 0.4785

Test Classification Report (Naive Bayes):
              precision    recall  f1-score   support

           0       0.85      0.79      0.82      2672
           1       0.90      0.93      0.91      5285

    accuracy                           0.88      7957
   macro avg       0.87      0.86      0.87      7957
weighted avg       0.88      0.88      0.88      7957
```

Both models provided detailed insights into their performance through classification reports and confusion matrices. Logistic Regression performed better overall, but Naive Bayes showed promise for improving minority class detection.

Discussion of Results:

The project successfully developed a system for classifying student statements into normal and abnormal categories. Logistic Regression emerged as the stronger model, delivering higher accuracy and better handling of unseen data. Naive Bayes, while effective in detecting less common abnormal cases due to SMOTE, fell slightly behind in overall accuracy. This is likely because Naive Bayes assumes that features are conditionally independent, an assumption that does not hold for text data where words often have contextual and semantic relationships. In contrast, Logistic Regression can better capture these dependencies, making it more suitable for this project.

Future improvements could focus on using more advanced text representation methods like **TF-IDF** or word embeddings (e.g., Word2Vec, GloVe) to capture deeper meaning in the text. Exploring deep learning models, such as LSTMs, could further boost the system's ability to understand complex patterns in the data.