Project 5 - Naive Bayes Classifier

Linhao Wu **Jason Jackson**

A02424787 **A02330998**

Github link: [jason-jackson6/CS5830_project_5 (github.com)]

Slides link:
https://docs.google.com/presentation/d/1k2QsdpwdHUH0FBOyHj5-sIsTH_J718379n9irQPPYfw/edit?usp=sharing

Dataset link:
https://www.kaggle.com/datasets/kreeshrajani/human-stress-prediction/data

1. Introduction

In today's information age, textual data on social media platforms is rapidly expanding, containing a wealth of information reflecting people's emotional states and mental health. This project aims to apply a Naive Bayes classifier, a simple yet powerful machine learning algorithm based on probability theory, to analyze social media posts, particularly identifying content related to stress. By analyzing posts from multiple Subreddits on the Reddit platform, this study seeks to understand and identify textual patterns expressing stress and related emotions. The project employs comprehensive text feature extraction methods, including Term Frequency-Inverse Document Frequency (TF-IDF) and word counts, while also considering text length as an auxiliary feature to enhance the performance of the classification model. Through carefully designed text preprocessing and feature engineering steps, along with Naive Bayes classifier optimization via random search, this research aims to provide an effective method for automating sentiment analysis of social media text, especially in the field of mental health.

2. Dataset

This study utilized a dataset comprised of posts from Reddit with the aim of identifying posts related to stress through their textual content (from Kaggle). The dataset consists of 2838 post records covering various Subreddit channels such as 'ptsd', 'assistance', 'relationships', and 'survivorsofabuse', reflecting authentic exchanges and sharing by users when facing stress and challenges. Each record includes information such as the text content of the post, the source Subreddit, post ID, sentence range, labels, confidence scores, and timestamps. Specifically, the 'label' field indicates whether the post is related to stress, serving as the target variable for the classification task in this project.

Characteristics of the dataset include the diversity and complexity of textual data, encompassing personal experiences, feelings, and descriptions of events, which provide rich material for text analysis and sentiment classification. Text data preprocessing involves steps such as removing special characters, lemmatization, and eliminating stopwords, aiming to reduce noise and extract useful information as a foundation for subsequent feature extraction and model training.

This research employed a Multinomial Naive Bayes classifier to analyze and classify the text data, considering its effectiveness in handling textual data. Text features were extracted using a combination of TF-IDF vectorization and count vectorization methods, with the introduction of text length as an additional feature to potentially enhance model classification performance. Additionally, this study utilized random search to optimize the hyperparameters of the model, aiming to find the best model configuration.

3. Analysis technique

The core of this project lies in the application of a Naive Bayes classifier for sentiment analysis on textual data to identify posts related to stress. Text preprocessing is the first step of this project, aimed at converting raw text into a format more suitable for machine learning models. Initially, we attempted to remove special characters using regular expressions to eliminate non-alphanumeric characters, reducing noise. Next, we tokenize the text into word sequences for further processing. Then, we lemmatize the words to reduce the impact of word diversity on the model. Finally, we remove stopwords, such as "the" and "is", which contribute minimally to the meaning of the text.

The second step involves feature extraction, which is the process of converting text data into numerical data that the model can understand. During feature extraction, we first employ TF-IDF vectorization, which involves calculating the importance of words in the document by adjusting the term frequency and inverse document frequency, thereby reducing the weight of high-frequency words and reflecting their importance. Then, we attempt to directly count the occurrences of words in the document, converting the text into a frequency vector. We also extract the text length as an additional feature using a custom transformer, considering that text length may be related to the sentiment of the text.

This project selects a Multinomial Naive Bayes model for classification because it is well-suited to handling the characteristics of textual data. It naturally handles the occurrence counts of variables (vocabulary), which is a typical feature of text data. In text classification tasks, the occurrence counts of each word (or frequencies under TF-IDF weighting) can serve as features, and the multinomial distribution model can adapt well to this count data. By integrating feature extraction and model training through a pipeline, the process from raw data to final predictions is simplified. To further improve the model's performance, this project employs random search (RandomizedSearchCV) to optimize the model's hyperparameters, including parameters for TF-IDF and count vectorization (such as 'max_df' - the maximum document frequency in the vocabulary, 'min_df' - the minimum document frequency, and 'ngram_range' - the range of considered n-grams) and the smoothing parameter ('alpha' value) for Naive Bayes to prevent probability calculation issues due to unseen features in the training samples.

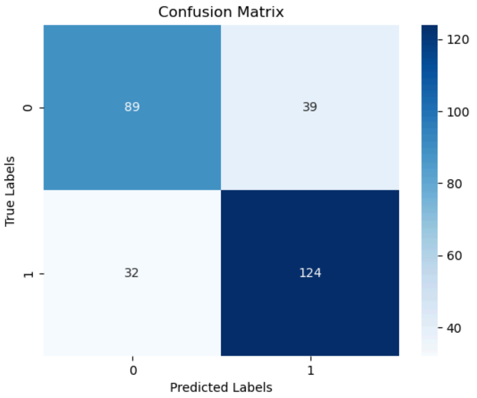
Through this comprehensive series of analysis techniques, this project aims to build an accurate classification model to identify posts related to stress, while also providing a reliable framework for handling and analyzing textual data.

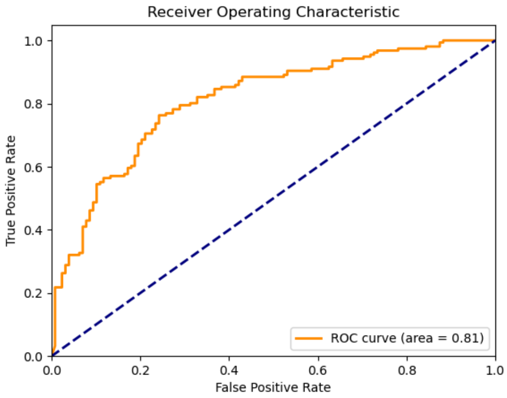
4. Result

In this chapter, we present the results of sentiment analysis on Reddit posts using a Multinomial Naive Bayes classifier. The model is optimized using random search to achieve accurate predictions on the test dataset, and its performance is evaluated using confusion matrices, ROC curves, and classification reports.

For model optimization using random search, we conducted 500 fits on 50 candidate models, each subjected to 10-fold cross-validation. The final best model parameter configuration is determined as follows:
- Multinomial Naive Bayes smoothing parameter 'alpha': 0.1
- TF-IDF vectorization 'max_df': 0.75
- TF-IDF vectorization 'min_df': 2
- TF-IDF vectorization 'ngram_range': (1, 1)
- Count vectorization 'max_df': 0.5
- Count vectorization 'min_df': 2
- Count vectorization 'ngram_range': (1, 2)

These parameters indicate meticulous adjustments to represent textual data to extract maximum information for classification. With the aforementioned parameter configuration, the model achieved a best score of 0.7294 during cross-validation. On the test set, the model attained an accuracy of 0.75, demonstrating good generalization to new data.

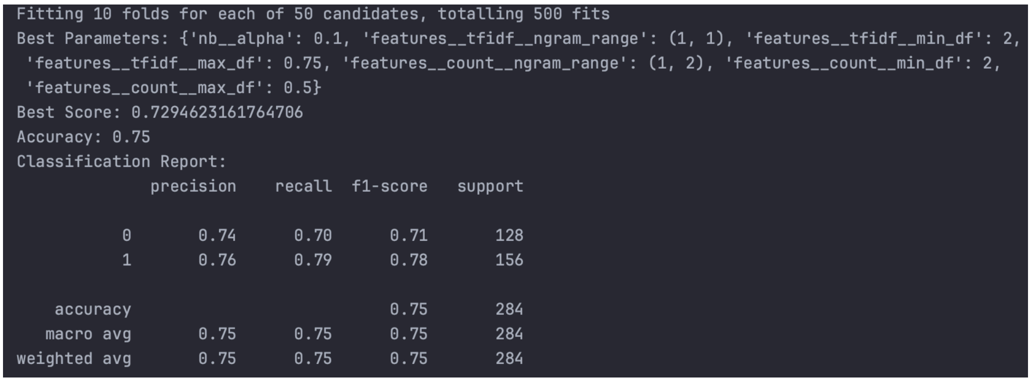




The confusion matrix reveals the model's performance in predictions, as follows:
- Class 0 (Not related to stress): The model correctly predicted 89 samples and misclassified 39 samples.
- Class 1 (Related to stress): The model correctly predicted 124 samples and misclassified 32 samples.

The model demonstrates more accurate performance in identifying posts related to stress, although there is still a proportion of posts misclassified. ROC curves and AUC values are key indicators for evaluating the classification performance of the model. The ROC curve's area under the curve (AUC) is 0.81, indicating the model possesses relatively high classification capability and can distinguish between the two categories of posts with reasonable accuracy.

For Class 0, the model exhibits a precision of 0.74 and a recall of 0.70, with an F1 score of 0.71. For Class 1, the model shows a precision of 0.76 and a recall of 0.79, with an F1 score of 0.78. The overall accuracy is 0.75, with macro-average and weighted-average precision, recall, and F1 scores all at 0.75. The classification report indicates the model performs well-balanced on both classes, with precision, recall, and F1 scores suggesting a good balance in identifying posts related and unrelated to stress.



```
Fitting 10 folds for each of 50 candidates, totalling 500 fits
Best Parameters: {'nb__alpha': 0.1, 'features__tfidf__ngram_range': (1, 1), 'features__tfidf__min_df': 2,
 'features__tfidf__max_df': 0.75, 'features__count__ngram_range': (1, 2), 'features__count__min_df': 2,
 'features__count__max_df': 0.5}
Best Score: 0.7294623161764706
Accuracy: 0.75
Classification Report:
              precision    recall  f1-score   support

           0       0.74      0.70      0.71       128
           1       0.76      0.79      0.78       156

    accuracy                           0.75       284
   macro avg       0.75      0.75      0.75       284
weighted avg       0.75      0.75      0.75       284
```

In summary, our Multinomial Naive Bayes model achieves satisfactory results in identifying stress-relatedness in Reddit posts. The confusion matrix, ROC curve, and classification report all demonstrate the model's excellent performance in distinguishing between post categories. While there is room for improvement in certain aspects of the model, the current results provide a solid foundation for further research and model enhancement. A more advanced model can provide support and contributions to not only mental health research, but also online community support.