Yuqing Qiao
1/28/2024

Text Classification Assignment: Movie Review Sentiment Analysis

Objective:

This assignment's primary goal is to implement and evaluate three text classification algorithms—Naive Bayes, Logistic Regression, and Multilayer Perceptron (MLP)—on the NLTK Movie Reviews dataset. The focus is on exploring the impact of two different feature representations: raw Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF).

1. Data Preparation

Utilized the IMDb movie reviews dataset through the NLTK library, conducted tokenization using NLTK's punkt tokenizer, along with stemming/lemmatization and the removal of stop words.

2. Coverage Analysis Insights

Conduct a coverage analysis to identify the percentage of unique words covered by the preprocessing steps. (coverage percentage = unique processed words count / accumulated unique words count). Belows is the visualization:

Figure 1 on movie reviews documents after shuffle

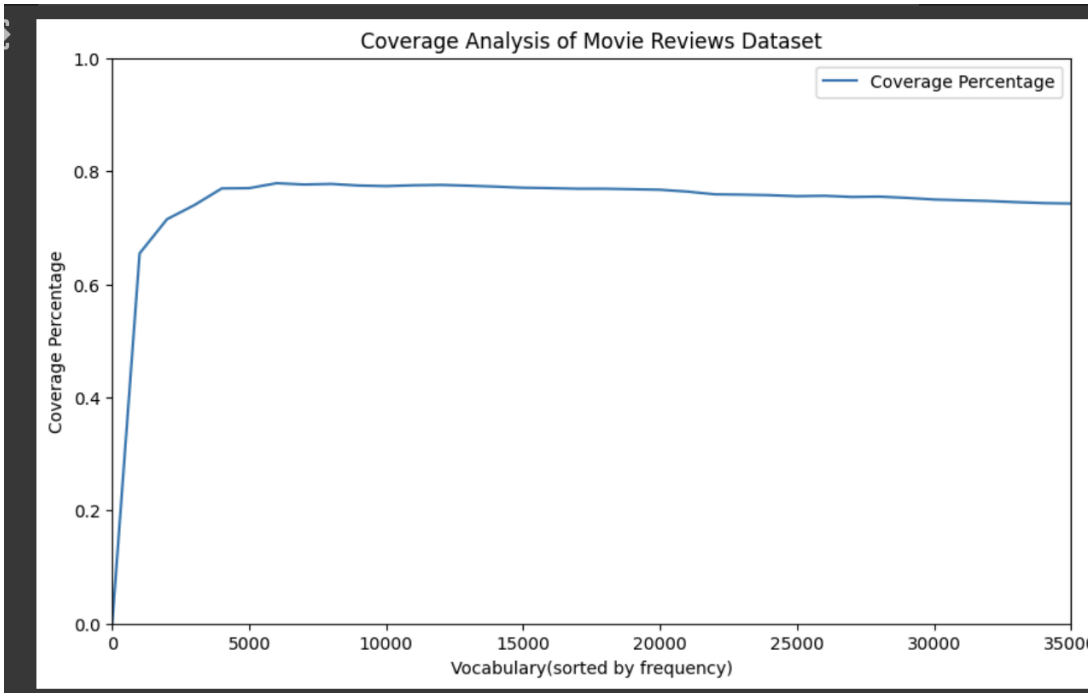
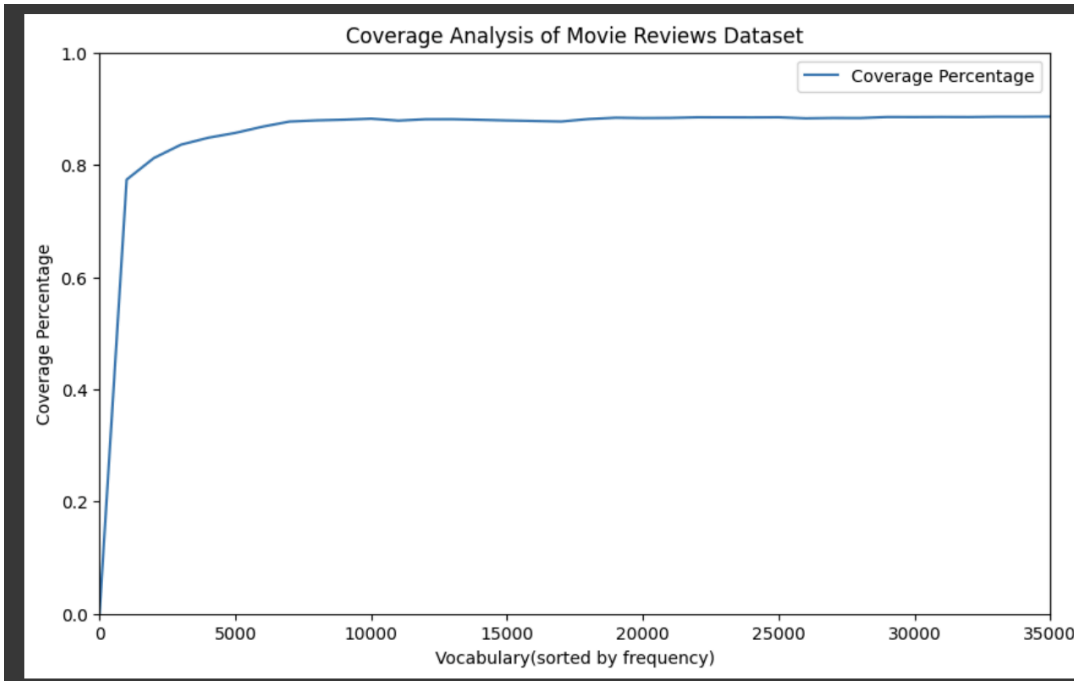


Figure 2 on movies review document after shuffle



The coverage increases as the number of considered tokens grows, eventually reaching a point of stabilization.

As more tokens are considered, less frequent words are encountered. These words add incrementally less to the overall coverage since they are less common or more specialized. Eventually, each new token adds very little to the coverage as I start encountering rare or highly specific words.

Rationalization for Vocabulary Choice:

1. Trade-off Between Larger Vocabulary and Computational Efficiency:
   A larger vocab can capture more info in the data, improving the model's accuracy and process a wider range of sentences. However, it increases memory requirement and computational complexity.

2. Impact of Rare or Very Common Words:
   Rare words can be very informative or irrelevant noise. Too many rare words leads to overfitting. Very common words are not informative based on zipf's law. Model should focus on more meaningful content.

3. Balancing Informativeness and Model Complexity:
   The goal is to find a vocabulary size that captures the most informative features without making the model overly complex.

`Naive Bayes:` This algorithm can work well with large vocabularies but might be sensitive to the presence of rare words which can skew the probability estimates.

`Logistic Regression:` This algorithm can handle sparse data, but a very large vocabulary can lead to long training times and overfitting.

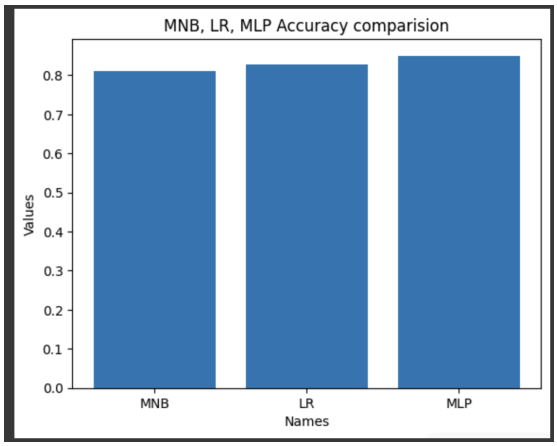
MLP (Multi-Layer Perceptron): Neural network models like MLPs can require substantial computational resources, especially with large vocabularies. They can benefit from a well-optimized vocabulary to reduce training time and memory usage.

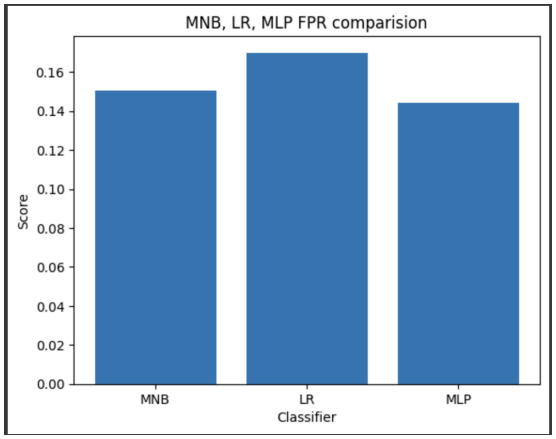
3. Algorithm Implementation

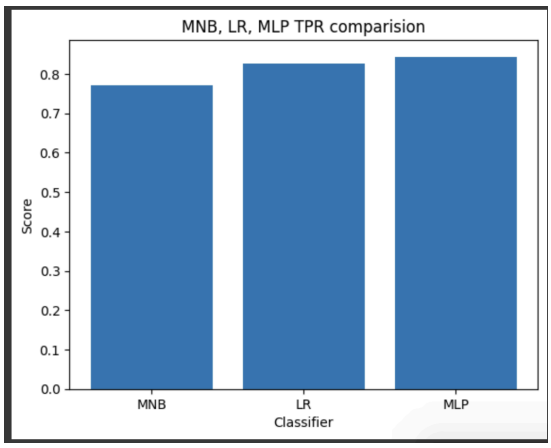
Implement Three classifier Naive Bayes, Logistic Regression, Multilayer Perceptron with both TF and TF-IDF.

4. Training and Evaluation:

Visualize the performance evaluation of each algorithm on the testing set, focusing on accuracy, True Positive Rate (TPR), and False Positive Rate (FPR) as the primary metrics.







Classifiers Comparison

For accuracy, three classifiers achieve almost the same result. For TPR, MLP > LR > MNB. For FPR, LR>MNB = MLP

For accuracy, all three classifiers (Multinomial Naive Bayes (MNB), Logistic Regression (LR), and Multi-Layer Perceptron (MLP)) achieving similar accuracy suggest that they are equally effective overall in classifying the dataset.

For TPR: MLP > LR > MNB: This ranking indicates that the MLP is best at identifying positive cases. The MLP's complex architecture allows it to capture nuanced patterns in the data, leading to better identification of true positives.

For FPR: LR > MNB = MLP: Logistic Regression having a higher FPR suggests that while it's good at identifying positive cases, it also misclassifies more negative cases as positive. This could be due to LR's linear decision boundary being less nuanced in certain scenarios.
The ideal model achieves a high TPR with a low FPR.

The impact of using TF vs. TF-IDF on classification performance

Term Frequency(TF) emphasizes common words that occur frequently in a specific document, under the assumption that frequent words are more relevant to the document's topic.
However, common words might be less informative and potentially dilute the classifier's ability to focus on more meaningful terms.

TF-IDF reduces noise by identifying unique words that are crucial for classifying a movie review. It could enable classifiers to focus more on distinctive words. MLP and LR could be benifited from it.