Sentiment Analysis Task: IMDb Movie Reviews Polarity Prediction

Sentiment analysis, a subfield of natural language processing, involves determining the sentiment (emotions and polarity) expressed in text data. In this project, I applied machine learning techniques to perform sentiment analysis on IMDb movie reviews. The goal is to classify reviews into positive or negative sentiments based on the expressed opinions.

The input for this project was gotten from Kaggle an online data science platform with numerous of dataset. The specific dataset used for this task is "IMDB Dataset of 50K Movie Reviews" The dataset comprises 50,000 movie reviews, with 40,000 for training and 10,000 for testing. The output is the binary sentiment labels (positive or negative)

The machine learning algorithm used are Logistic regression, Naïve bayes and Decision trees Hyperparameter. The logistic regressions use regularization parameter (C) tuned using GridSearchCV, this was done to find the best value of C that maximizes the model's accuracy on the validation set. While Naïve Bayes and decision trees used default hyperparameters. Packaged used for this project was sklearn. I also included a majority guess strategy for binary classification, a baseline model to establish a trivial performance level.

Accuracy was the measure the success for this binary classification tasks, representing the proportion of correctly classified instances. It provided a straightforward evaluation of model performance.

The result for the following approaches is stated below:

Logistic Regression:

Best Hyperparameters: {'C': 0.1}

Accuracy: 86.62%

Naive Bayes:

Cross-Validation Accuracy: 84.71%

Accuracy: 84.88%

Decision Trees:

Cross-Validation Accuracy: 71.98%

Accuracy: 72.45%

Dummy Classifier:

Accuracy: 49.61%

Overall, Logistic Regression outperformed Naive Bayes and Decision Trees in accuracy. Although, Naive Bayes did demonstrate a competitive performance. Decision Trees, while providing acceptable accuracy, showed a noticeable performance gap compared to logistic regression and naive Bayes. The Dummy Classifier serves as a baseline, the models were able to surpass this minimal accuracy level.

This analysis provides insights into the effectiveness of different machine learning models for sentiment analysis on IMDb movie reviews. The choice of logistic regression, with tuned hyperparameters, resulted in the highest accuracy among the models evaluated.

Acknowledgment:

I would like to acknowledge the authors of the IMDB dataset for sentiment analysis used in this project. The dataset was introduced in the following paper:

Learning Word Vectors for Sentiment Analysis

Authors: Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, Christopher Potts

Published: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, June 2011, Pages 142-150.

Publisher: Association for Computational Linguistics

URL: http://www.aclweb.org/anthology/P11-1015