Sentiment Analysis

Mudit Arora

April 26, 2025

1 Sentiment Analysis on Stanford's total movie review corpus

1.1 Introduction

This problem focuses on predicting the sentiments expressed in movie reviews using the Stanford movie review corpus, which offers a diverse array of reviews labeled as either positive or negative. The primary challenge is to accurately classify sentiments, considering the nuances of human expression, where the same words can convey different emotions depending on their context.

The dataset features a rich variety of reviews, enabling an exploration of sentiment expressions across various genres and styles. By employing multiple machine learning models, we aim to evaluate their performance in accurately identifying sentiments and compare their effectiveness based on validation accuracy.

Through this analysis, we seek to identify not only the most effective models for sentiment classification but also the factors that influence their performance. This exploration will enhance the understanding of sentiment analysis methodologies and their practical applications in real-world situations.

1.2 Dataset

The dataset utilized in this study is the Stanford movie review corpus, comprising movie reviews that are labeled as either positive or negative. This dataset features a structured format, including text reviews alongside their corresponding sentiment labels. To facilitate effective analysis, we implemented various methods for loading and preprocessing the dataset.

1.3 Feature Extraction

We utilized an n-gram approach combined with Count Vectorizer. The dataset was split into training and validation sets to facilitate model evaluation.

1.4 Models

We evaluated the following models in our analysis, each with specific hyperparameters tuned for optimal performance:

1.4.1 Naive Bayes

The Naive Bayes model is based on applying Bayes' theorem with strong independence assumptions between the features. We utilized the Multinomial Naive Bayes variant, suitable for discrete data like text.

1.4.2 Linear Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. It works by finding the hyperplane that best separates data points of different classes in a high-dimensional space, maximizing the margin between them.

1.4.3 Decision Tree

A decision tree is a supervised machine learning algorithm used for classification and regression tasks that models decisions and their possible consequences as a tree-like structure. It splits the dataset into subsets based on feature values, creating branches that lead to outcomes (leaves) while maximizing information gain or minimizing impurity at each node, enabling clear and interpretable decision-making paths.

Table 1: Hyperparameters for Decision Tree

Hyperparameter	Value
Criterion	gini
Max Depth	5

1.4.4 Logistic Regression

Logistic regression is a statistical method for predicting binary classes. It applies the logistic function to model the probability that a given input belongs to a particular category.

1.4.5 K-Nearest Neighbors (KNN)

KNN is a simple yet effective algorithm that classifies instances based on the majority class among its k nearest neighbors in the feature space.

Table 2: Hyperparameters for KNN

Hyperparameter	Value
Number of Neighbors	5
Power	2
Metric	minkowski

1.5 Result on the Training Dataset

The validation accuracies obtained from our experiments are as follows:

- **Naive Bayes**: 0.88
- **SVM**: 0.90
- **Decision Tree**: 0.69
- **Logistic Regression**: 0.90
- **KNN**: 0.58

Both SVM and Logistic Regression are giving 0.90 accuracy so we'll be using these models on the testing dataset

1.6 Result on the Testing Dataset

- **SVM**: 0.89
- **Logistic Regression**: 0.90

Therefore, the best model for sentiment analysis for movie reviews seems to be Logistic Regression beating SVM by just 0.1

1.7 Conclusion

In summary, we showcased the effectiveness of different sentiment analysis models applied to the Stanford movie review corpus. The findings highlight opportunities for enhancement through additional experimentation and the exploration of alternative methodologies.