Airline Review Classification Using Machine Learning

Pranjal Patel

August 2024

1 Introduction

This project applies a machine learning model to classify airline reviews based on the user ratings given, ranging from 1-5. The implemented model uses NLP techniques on the text for preprocessing purposes. This model has Logistic Regression for the classification of reviews.

2 Data Preprocessing

2.1 Loading of Data

The dataset is loaded into a CSV file through pandas using special encoding. ISO-8859-1 is used here to avoid special characters.

2.2 Missing Value Checking

The data is checked for any missing values to be filled.

2.3 Text Preprocessing

2.3.1 Stemming

The words are reduced to their root form with the Porter Stemmer.

2.3.2 Stopword Removal

Common English stop words are removed because they do not add much value to the context.

3 Class Imbalance Handling

This dataset, due to its nature, contains unequal rating distribution. To deal with this, SMOTE is applied to balance the classes in the training set.

4 Building the Model

4.1 Vectorization

The text data is converted into numerical format using TfidfVectorizer, which is then used to calculate the TF-IDF values.

4.2 Model

A Logistic Regression model is trained using the balanced dataset. The parameters of the model are tuned with max_iter = 1000 and C = 0.5.

5 Model Evaluation

5.1 Accuracy

The accuracy scores are calculated for both training and test datasets. This should give an indication of the model's capacity for generalization to new and unseen data.

6 Saving and Loading the Model

The trained model, along with related variables, is saved using Python's pickle module for easy reloading for future tasks.