Airline Reviews Sentiment Analysis using Machine Learning

Hemant Nikam

M.Tech, Aerospace Engineering

Indian Institute of Technology, Bombay

Introduction

In this report, we present a detailed analysis of sentiment prediction for airline reviews using a logistic regression model. The goal of this project was to classify reviews as either positive or negative based on the sentiments expressed by passengers. The analysis involved comprehensive data preprocessing, feature engineering, model development, and performance evaluation.

Data Preprocessing

We began by importing the necessary libraries, including Pandas for data manipulation and Scikit-Learn for machine learning tasks. The dataset, containing airline reviews, was loaded using Pandas. We conducted initial exploration to understand the structure and content of the data.

After loading the data, we performed data cleaning, which included:

- Dropping unnecessary columns: We removed the 'Unnamed: 0' column that held no valuable information.
- Handling duplicates: Duplicate reviews were identified and removed from the dataset.
- Imputing missing values: We applied forward-fill imputation to columns like 'Type Of Traveller' and 'Seat Type' to maintain data integrity.
- Numeric conversion: The 'Overall_Rating' column contained non-numeric values, which were replaced and converted to integers.
- Feature engineering: We engineered the 'Recommended' feature by mapping string values ('yes' and 'no') to binary integers (1 and 0).

Text Data Processing

Text data processing was crucial for preparing the textual content of reviews for analysis. The following steps were taken:

- Lowercasing: All text was converted to lowercase to ensure uniformity and eliminate case-based variations.
- Punctuation and digit removal: Punctuation and digits were removed from the text using custom functions.
- Stopword removal: Common stopwords were removed to reduce noise and improve feature quality.
- Lemmatization: Text was lemmatized to normalize words and improve text representation.

Feature Engineering and Model Development

We then moved to feature engineering and model development:

- Encoding categorical variables: We used one-hot encoding to transform categorical variables into numerical features suitable for analysis.
- TF-IDF vectorization: Processed text data was transformed into numerical features using TF-IDF vectorization.
- Data splitting: The dataset was split into training and testing sets using the train_test_split function.

A logistic regression model was chosen for sentiment prediction due to its simplicity and effectiveness for binary classification tasks. The logistic regression model was built using Scikit-Learn's LogisticRegression class.

Model Evaluation

The logistic regression model was evaluated using various performance metrics:

- Accuracy: The accuracy of the model in predicting sentiment labels was calculated.
- Confusion Matrix: The confusion matrix provided insights into true positive, true negative, false positive, and false negative predictions.
- Precision, Recall, and F1-score: These metrics offered a balanced assessment of the model's precision and recall capabilities.
- Classification Report: A comprehensive classification report summarized key metrics for both positive and negative classes.

Conclusion

In conclusion, the sentiment analysis of airline reviews using a logistic regression model demonstrated the effectiveness of this approach in classifying reviews as positive or negative. The meticulous data preprocessing, feature engineering, and model development contributed to the model's success in accurately predicting sentiment labels. Further enhancements and refinements could be explored, such as more advanced text processing techniques and experimenting with different machine learning algorithms.