Project Report - Sentiment Analysis API

Objective

In order to create an API endpoint that can accept a text and return associated sentiment with it.

Introduction

This project has been done in order to establish my knowledge and analytical skills in Machine Learning. I try to classify airlines travel journey reviews into ‘positive’ and ‘Negative’ sentiment by building a model based on probabilities using given dataset.

Source Dataset link:
https://drive.google.com/file/d/1iHdXv0ex90AT3T2JqFlTRqNtZuATkEJn/view

Text Pre Processing

Dataset contains ‘Sl.No’, ‘airline_sentiment’ which is target variable and ‘text’ which is independent variable in the form of sentence.
By understanding the dataset, Airlines company name present in the first word of the sentence with prefix ‘@’ symbol.
Also, punctuation marks, numerals and unicode characters like â, ð in the sentence.
These text don't particularly add anything meaningful to the sentence. Hence cleansing done by dropping all the above in the text column.

Feature Engineering

By getting the frequencies of the words, we can find many stopwords present in the independent variable. List of words extracted from the dataset and stopwords has been dropped.
Then wordcloud and matplotlib libraries imported to visualize the text and get meaningful insights.
Text is converted into values by vectorization method, so that the ML model can perform.

Model Training & Evaluation

Before fitting the model, given dataset has been split into Train and Test data with test size 0.2. As it is a classification problem, Naïve Bayes and Logistics Regression classification algorithms taken to find out the best prediction outcome. Train data fitted into both algorithms, then test data passed into the trained model.
Evaluation done on the basis of accuracy score and confusion matrix.

Conclusion

Logistics Regression is better than Naïve Bayes which is finest with 90% accuracy score. Hence I chose to go with logistics Regression from sklearn. Final model deployed by Flask and got JSON response from the browser.