Twitter Sentiment Prediction of US Airlines

Milestone: Project proposal

Group 12

Student-1 Name : Amitoj Singh Kohli

Student-2 Name : Aryaman Kakad

857-313-0650 (Tel of Student 1)

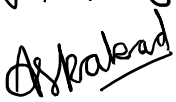
857-313-5234 (Tel of Student 2)

kohli.am@northeastern.edu
kakad.a@northeastern.edu

Percentage of Effort Contributed by Student1: 100%

Percentage of Effort Contributed by Student2: 100%

Signature of Student 1: 

Signature of Student 2: 

Submission Date: 01 / 29 / 2023

Business Problem Definition:

Airline companies in the United States face a significant challenge in understanding and managing customer satisfaction. Negative sentiment towards an airline can lead to decreased bookings and revenue. Social media platforms, such as Twitter, provide extensive customer feedback on airline service quality that can be used to gain insights into customer sentiment toward the airlines. However, the sheer volume of data can make it challenging to manually analyze and understand the overall customer sentiment.

Problem Objective:

This project aims to predict the sentiment of tweets about US airlines using supervised machine learning techniques. It will understand customer perceptions and predict the sentiment of tweets as positive, negative, or neutral. This model will help airlines make data-driven decisions to identify areas of improvement and tailor their services to meet customer needs better.

Data Mining Tasks:

1. Data Collection: The dataset for this project was obtained from Kaggle, and the shape of dataset is (14640, 15).
2. Data Preprocessing: Clean and preprocess the data by removing irrelevant information, handling missing values, removing stop words, and correcting errors.
3. Feature Engineering: Preprocess the text data and generate new features that will be useful for the sentiment analysis model.
4. Exploratory Data Analysis: Use python libraries such as matplotlib, seaborn for data visualization and to gain insights into the distribution of sentiments, the most common reasons for negative sentiment, and the most frequently mentioned airlines.
5. Model Selection: This task will involve selecting the most appropriate supervised machine learning model for the problem like Naive Bayes Classifier, SVM Model, Random Forest.
6. Model Evaluation: Evaluate the model's performance on a test set and make any necessary adjustments and visualizing model’s performance through confusion matrix, F1 score.

Solution Design:

The proposed solution uses supervised machine learning to predict the sentiment of tweets about US airlines. The solution will consist of the following steps:

1. Cleaning and preprocessing the data for various data issues / error pattern.
2. Present the insights clearly and concisely using data visualization techniques.
3. Training a supervised machine learning model using labeled data on sentiment.
4. Using the trained model to predict the sentiment of tweets about US airlines.
5. Continuously evaluate the model's performance and make necessary improvements.

Future Scope:

Deploy the trained model in the production environment and make predictions on new tweets in real-time.