Twitter Airline

1. Project Overview

We analyzed a dataset containing tweets about airlines to determine public sentiment. The goal was to clean the data, visualize it, and build a model to classify sentiments as positive or negative.

The analysis involves:

- Data preprocessing
- Data cleaning and transformation
- Exploration of the most frequently occurring words
- Visualization of data patterns and sentiment distribution

Libraries Used

- **matplotlib.pyplot**: A library for creating basic graphs and visualizations.
- **wordcloud**: A tool for generating word clouds that display frequent words in larger sizes.
- **nltk (Natural Language Toolkit)**: A library for natural language processing tasks, like removing common words and tokenizing text.
- **re**: A library for regular expressions used to find and remove unwanted text, such as URLs and special characters.
- **string**: A module for handling text features, including punctuation and character sets.

- **sklearn.feature_extraction.text.CountVectorizer**: Converts text into numerical format by counting word occurrences for modeling.
- **nltk.tokenize.RegexpTokenizer**: A tokenizer that splits text into words based on specified rules for character inclusion.
- **sklearn.linear_model.LogisticRegression**: A method for predicting sentiment by identifying patterns in text data.
- **sklearn.metrics.accuracy_score**: Measures the percentage of correct predictions made by the model.
- **sklearn.feature_extraction.text.TfidfVectorizer**: Converts text into numerical format while considering the importance of words across documents.

2. Dataset Description

the dataset included many other columns that were not necessary for our analysis. Therefore, these irrelevant columns were removed to focus only on the data pertinent to our sentiment analysis.

- **Data Source**: The dataset is a CSV file named Tweets.csv.
- **Main Columns**:
  - `text`: The tweet content.
  - `airline sentiment`: The sentiment label (positive or negative).

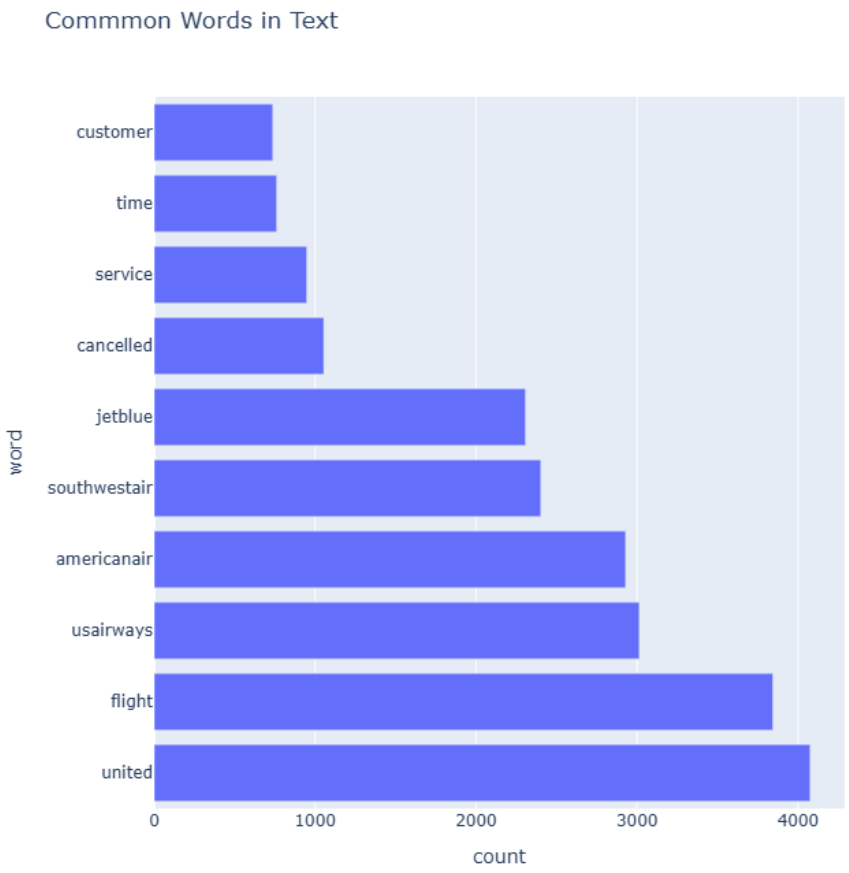
3. Data Cleaning Process

1. **Filtering**: We kept only the tweets with a confidence score above 0.6.
2. **Text Cleaning**:
    - Converted text to lowercase.
    - Removed URLs, numbers, special characters, and emojis.
    - Replaced common contractions (e.g., "isn't" to "is not").
3. **Stop Words Removal**: Removed common words (e.g., "the", "and") that do not add meaning.
4. **Frequent Words Removal**: Excluded the top six most common words to focus on significant terms.
5. **Lemmatization**: Reduced words to their base form (e.g., "running" to "run").

4. Visualization

Word Cloud of Frequent Words

A word cloud provides an intuitive view of the most frequent terms, (e.g flight, united , us airways) .



5.Model Selection

Model Used

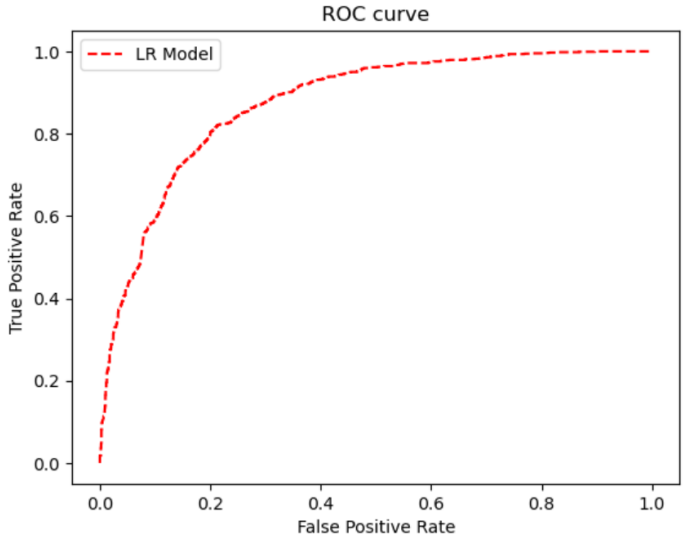
We chose **Logistic Regression** for sentiment classification. This model is suitable for binary outcomes (positive vs. negative) and is easy to interpret.

Data Preparation

- **Text Vectorization**: We transformed the tweet text into numerical format using the **Count Vectorizer**. This method converts text into a matrix of token counts.
- **Train-Test Split**: The dataset was split into training (80%) and testing (20%) sets to evaluate the model's performance.

6.Model Results

- **Accuracy**: The logistic regression model achieved an accuracy of approximately 80%.
- **Evaluation Metrics**:
    - Confusion Matrix: Showed the number of correct and incorrect predictions.
    - ROC Curve: Plotted to visualize the model's performance.
    - AUC Score: Calculated to assess the model's ability to distinguish between positive and negative sentiments.



ROC value of 0.8746 is good, indicating model has strong ability to distinguish between the positive and negative classes. Generally, values closer to 1 represent better performance.