Title: Analyzing Airline Sentiments

A Review of Machine Learning Models for Predicting Airline Sentiments

Patrick Miller & Ganesh Ramcharan

SUNY Old Westbury

Spring 2022

**Abstract**

This report presents an analysis of Twitter posts scraped from 2015, focusing on various airlines. The objective was to analyze the sentiments expressed in these posts, clean the data, and employ machine-learning techniques to identify the most common issues faced by passengers. Additionally, an attempt was made to develop a predictive model to assess the likelihood of encountering similar issues with an airline. The dataset utilized in this study consisted of Twitter posts, encompassing text data and corresponding sentiment labels. Through data cleaning and machine learning analysis, valuable insights were gained regarding the prevalent issues faced by airline passengers and the potential for predicting such issues in the future.

**Introduction:**

Customer sentiment analysis plays a crucial role in the airline industry, providing valuable insights into passenger experiences and satisfaction. With the rise of social media platforms like Twitter, customers now have a powerful medium to share their opinions and real-time experiences. In this report, we conducted an analysis of the Twitter US Airline Sentiment dataset from February 2015, comprising over 14,000 tweets mentioning major US airlines, including American, Delta, Southwest, United, US Airways, and Virgin America.

The dataset, stored in an SQLite database, contains fifteen columns that provide valuable information about each tweet. Key columns include the tweet text, sentiment labels categorizing tweets as positive, negative, or neutral, and confidence scores reflecting the annotators' certainty in the assigned sentiment.

To ensure the reliability and effectiveness of our analysis, we conducted a comprehensive data cleaning process. We removed irrelevant columns such as "tweet\_id", "negativereason\_gold", "retweet\_count," "name", and others that did not contribute useful information. Additionally, we removed tweets labeled as "neutral" and "positive", as they were not the focus of our analysis. We also filtered out tweets containing generic phrases like "bad flight" and "can't tell", as they did not provide specific reasons or insights into customer experiences.

By focusing on tweets with negative sentiment labels and excluding generic phrases, we refined our dataset to capture genuine issues and concerns expressed by customers. This allowed us to delve deeper into identifying the root causes of dissatisfaction and provide more accurate insights.

Our analysis utilized unsupervised learning algorithms, which do not require labeled data, to explore patterns and cluster similar sentiments within the dataset. By applying clustering algorithms, such as k-means and/or hierarchical clustering, we identified groups of tweets that shared similar sentiments and common issues. This unsupervised approach allowed us to uncover hidden patterns and gain a deeper understanding of the underlying sentiment dynamics.

In addition to clustering, we conducted exploratory data analysis to gain insights into the dataset's characteristics. We visualized the distribution of sentiments and analyzed word frequencies. This analysis provided a comprehensive overview of the data and helped us uncover interesting trends and patterns.

Next, we applied various machine learning algorithms to predict sentiment labels and identify the most common issues faced by customers. While the unsupervised learning algorithms helped us cluster similar sentiments, supervised learning algorithms, such as classification models, were employed to predict sentiment labels for new, unseen data. These models were trained on the labeled portion of the dataset and evaluated based on their performance metrics.

By leveraging the power of social media data and employing unsupervised learning algorithms, our analysis provides valuable information to airlines, enabling them to proactively address customer concerns, enhance satisfaction, and optimize their services. Through this report, we contribute to the continuous improvement of the airline industry by understanding and utilizing customer sentiments.

**Description**

The Twitter US Airline Sentiment dataset used in this analysis comprises over 14,000 tweets from February 2015. The dataset includes tweets mentioning major US airlines, namely American, Delta, Southwest, United, US Airways, and Virgin America.

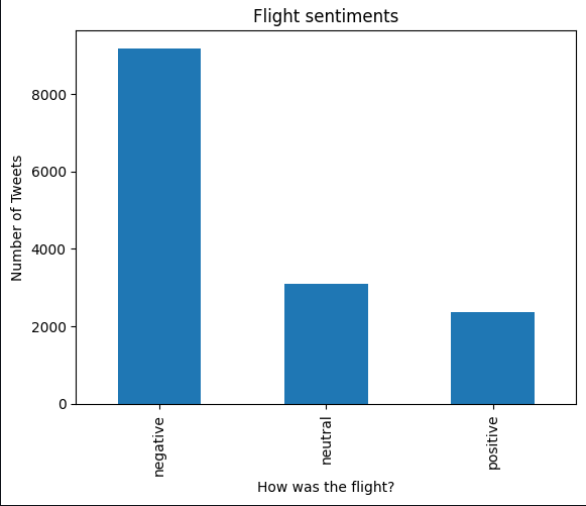
The dataset is stored in an SQLite database and consists of fifteen columns, providing various information about each tweet. The key columns of interest are:

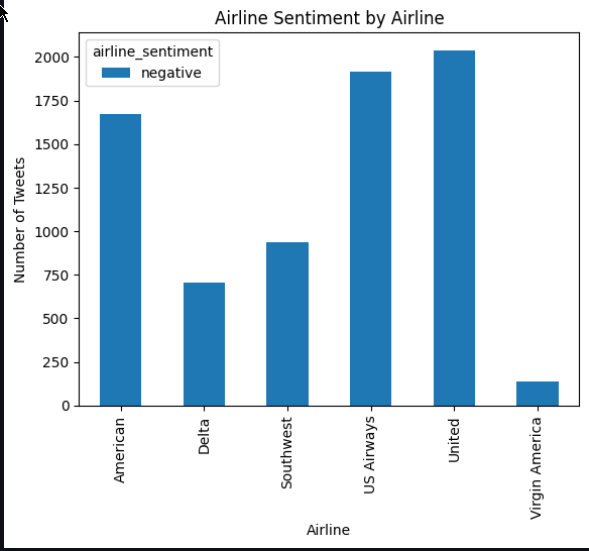
* "negativereason": This column contains the actual tweet text, which serves as the primary source of information for sentiment analysis.
* "airline\_sentiment": This column categorizes each tweet as positive, negative, or neutral, representing the sentiment expressed by the tweet author towards the airline.
* "airline\_sentiment\_confidence": This column provides a measure of confidence in the assigned sentiment label, determined by annotators. It indicates the level of certainty in the sentiment classification.

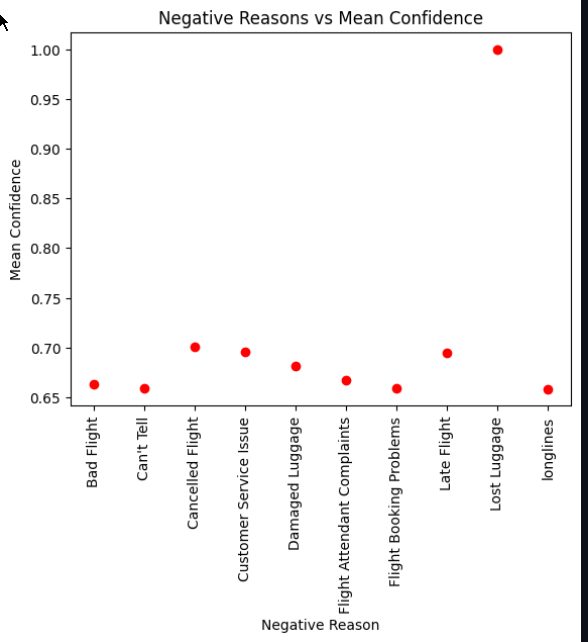
Exploring and understanding the dataset's structure, distribution of sentiments, and other characteristics will be crucial for further analysis and modeling. Next, we will delve into Exploratory Data Analysis (EDA) and visualization techniques to gain insights into the dataset and uncover patterns and trends.

**EDA/Data Cleaning/Preprocessing**

To begin the exploratory data analysis (EDA) process, we first displayed the columns of the untouched data to get an overview of the available information. Then, we used the tweetsDf.head() function to display the first 5 rows and all 15 columns of the dataset.

Next, we filtered the data and selected the 'airline\_sentiment' column to get the unique counts for each sentiment category (positive, negative, neutral). This allowed us to understand the distribution of sentiments in the dataset.

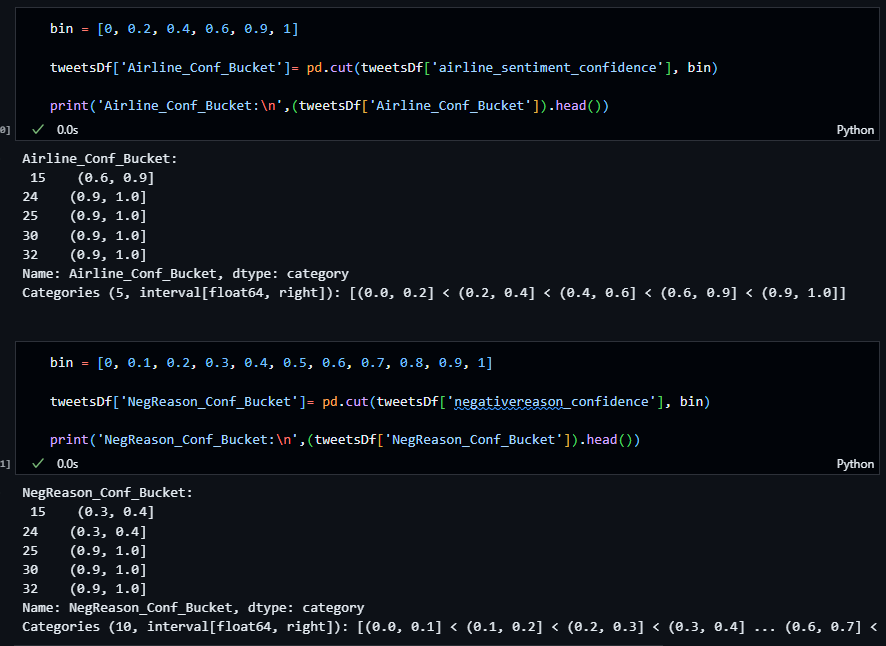
Afterwards, we focused specifically on the negative sentiments and selected the rows that corresponded to negative sentiment tweets. We further grouped the data by airline and sentiment to visualize the distribution of negative sentiments across different airlines using a histogram.

Moving on, we grouped the data by negative reason and computed the median confidence for each negative reason. This information was then displayed in a scatter plot, allowing us to analyze the relationship between the negative reasons and the confidence level in assigned sentiment labels.

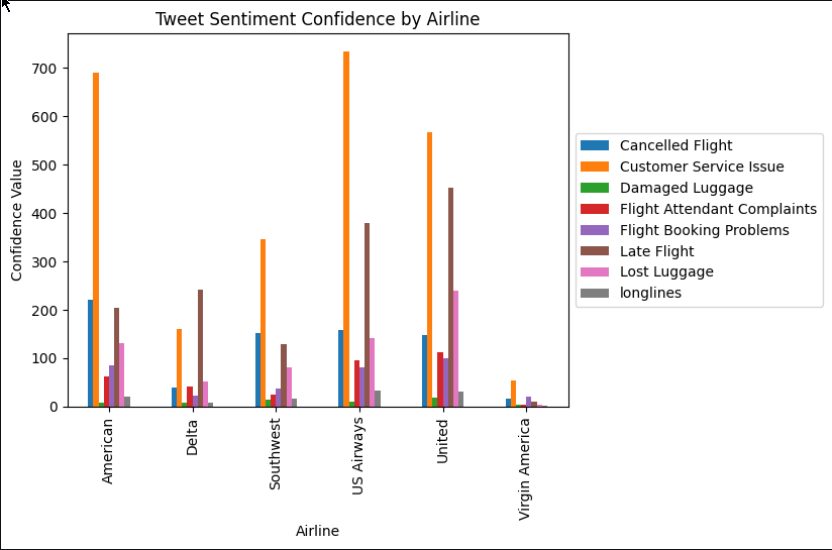
As part of the data cleaning process, we dropped the columns and rows that were not needed for our analysis. This helped streamline the dataset and remove unnecessary information. After making these changes, we displayed the dataframe again to observe the resulting modifications.

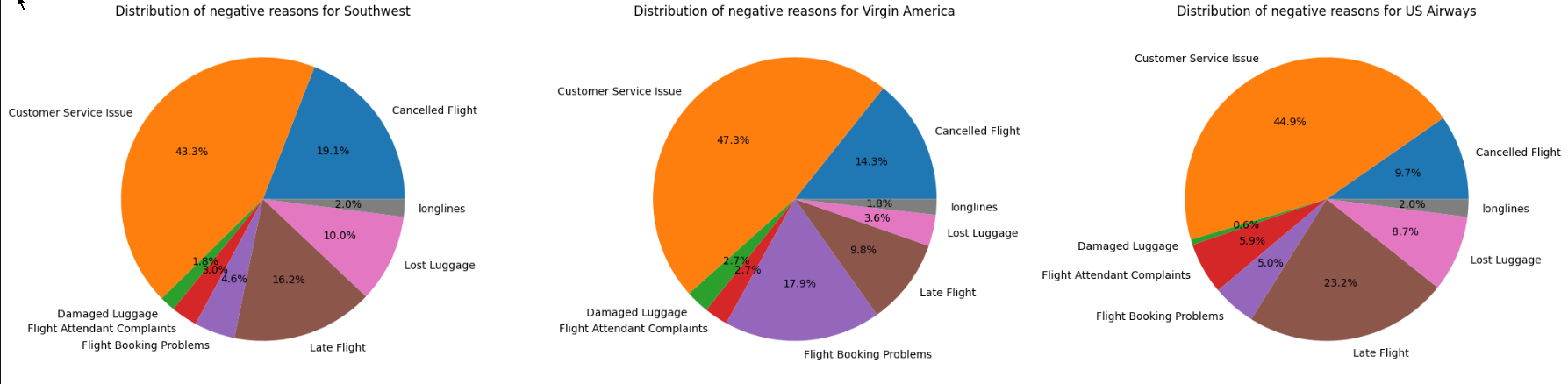
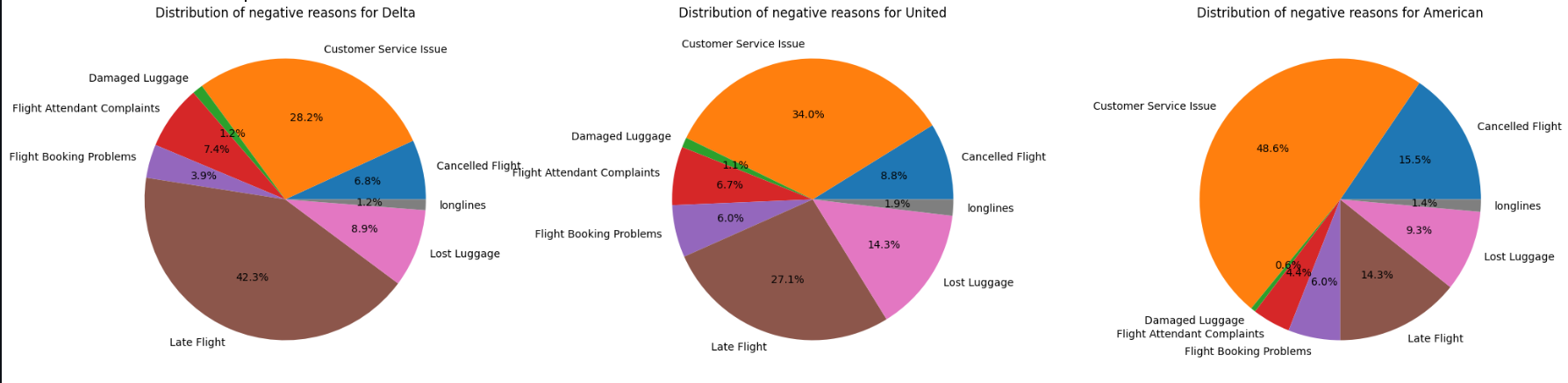
A screen shot of a computer

Description automatically generated with medium confidenceTo ensure data integrity, we checked for empty spaces or missing values in the dataset and confirmed that there were none. This ensured that our subsequent analysis would be based on complete and reliable data.

Next, we bucketized the cleaned data and displayed scatter plots and histograms to visualize the distributions of 'negativereason\_confidence' and 'airline\_sentiment\_confidence'. These visualizations provided insights into the confidence levels associated with different negative reasons and sentiment labels.

Additionally, we displayed a histogram of all the tweets to identify the most frequent negative reasons based on 'negativereason\_confidence'. This helped us understand which issues were more commonly mentioned in the dataset.



To provide a comprehensive view of the negative sentiments by airlines, we created six pie charts to display the distribution of negative reasons for each airline. This allowed for a visual comparison of the prevalent issues faced by different airlines.

Finally, we added a heatmap to the analysis, using colors to represent patterns and relationships in the dataset. The heatmap provided a visual representation of correlations between different variables and helped identify any significant patterns or trends.

A screenshot of a graph

Description automatically generated with low confidence

Overall, the EDA and data cleaning and preprocessing steps provided insights into the structure and characteristics of the dataset, allowing us to understand sentiment distributions, identify prevalent negative reasons, and ensure the reliability and completeness of the data for further analysis.

**Data Modeling and Evaluation**

In the data modeling phase, we applied various clustering algorithms and a classification model to analyze the sentiment data. Here are the algorithms used and their performance:

K-means Clustering and Agglomerative Clustering: We utilized K-means clustering and Agglomerative clustering algorithms to group the data based on the columns 'negativereason\_confidence' and 'airline\_sentiment\_confidence'. The algorithms demonstrated a high accuracy of 96% in identifying the correct clusters.

DBSCAN Clustering: We also employed the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm for clustering. The DBSCAN algorithm produced a score of 0.68, indicating its effectiveness in identifying clusters within the data.

Classification Model: Additionally, we implemented a classification model to predict the sentiment labels. The model achieved a remarkable accuracy of 1.0 on a test dataset, indicating that it correctly predicted the sentiment labels for all instances.

The classification report provided detailed information about the model's performance. Precision, recall, and f1-score were calculated for the "negative" class, as it was the focus of our analysis. The precision and recall values were both 1.0, indicating that all instances predicted as "negative" were indeed negative, and all negative instances were correctly predicted. The f1-score, which represents the balance between precision and recall, was also 1.0, indicating a perfect balance.

Furthermore, the support column displayed the number of instances in the test dataset for the "negative" class, which was 1240 instances.

Overall, the clustering algorithms and classification model demonstrated promising results in analyzing the sentiment data, providing valuable insights into the clustering structure and accurately predicting sentiment labels.