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. 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.

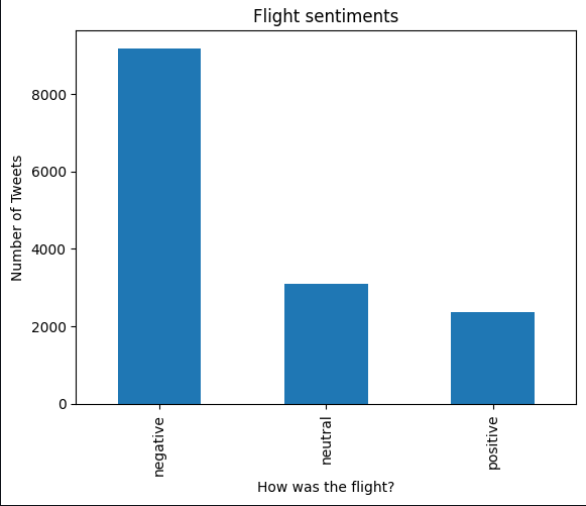
The dataset is stored in an SQLite database and consists of fifteen columns, providing various information about each tweet. The key columns of interest are:

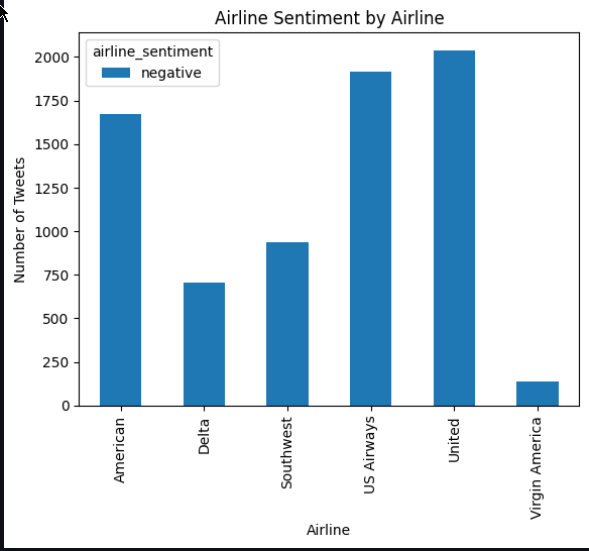
* "negativereason": This column contains the actual tweet text, which serves as the primary source of information for sentiment analysis.
* "airline\_sentiment": This column categorizes each tweet as positive, negative, or neutral, representing the sentiment expressed by the tweet author towards the 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.

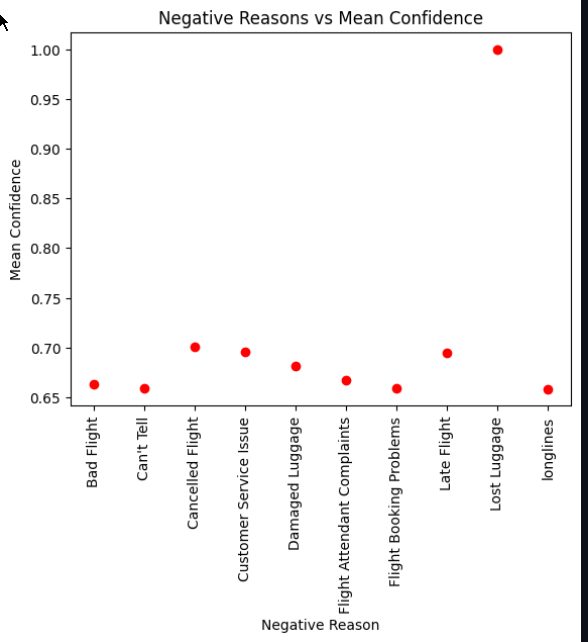
Exploring and understanding the dataset's structure, distribution of sentiments, and other characteristics will be crucial for further analysis and modeling. Next, we will delve into Exploratory Data Analysis (EDA) and visualization techniques to gain insights into the dataset and uncover patterns and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

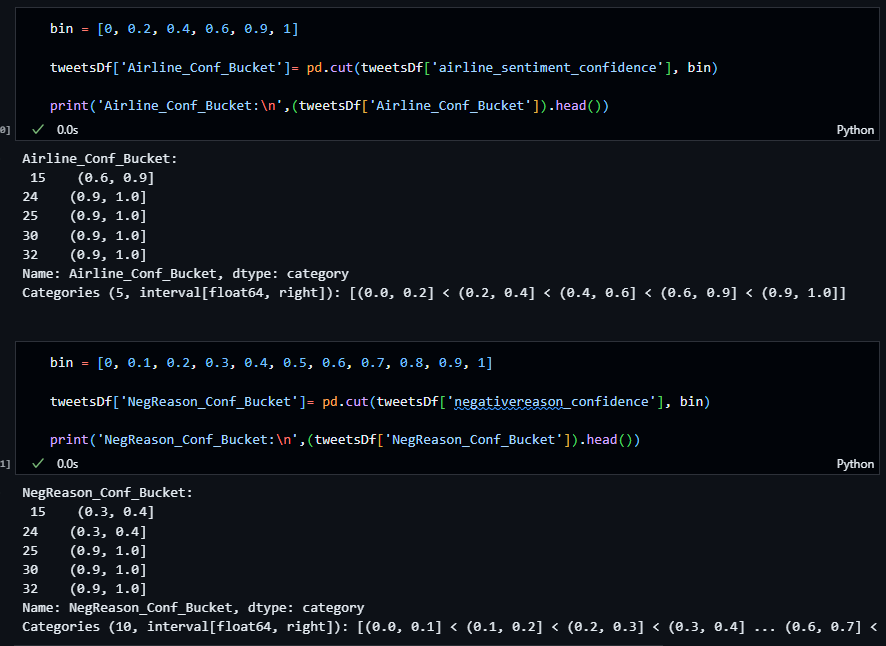
Afterwards, we focused specifically on the negative sentiments and selected the rows that corresponded to negative sentiment tweets. We further grouped the data by airline and sentiment to visualize the distribution of negative sentiments across different airlines using a histogram.

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.

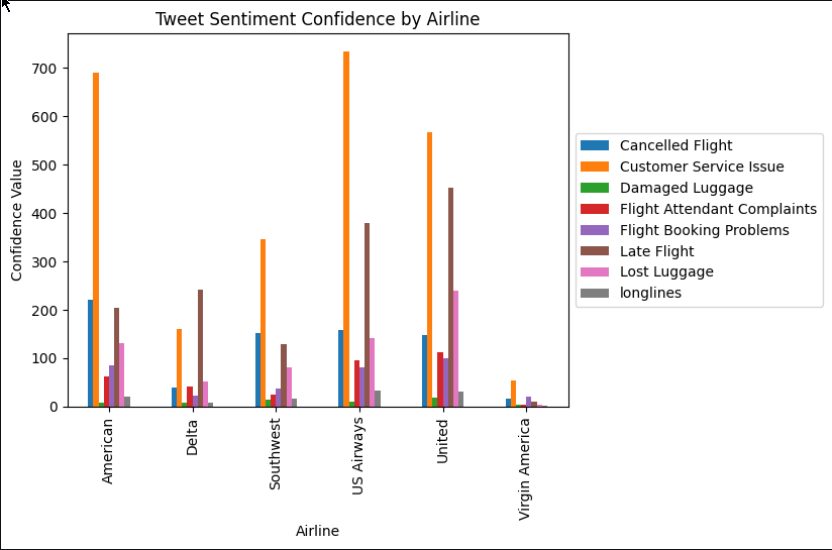
As part of the data cleaning process, we dropped the columns and rows that were not needed for our analysis. This helped streamline the dataset and remove unnecessary information. After making these changes, we displayed the dataframe again to observe the resulting modifications.

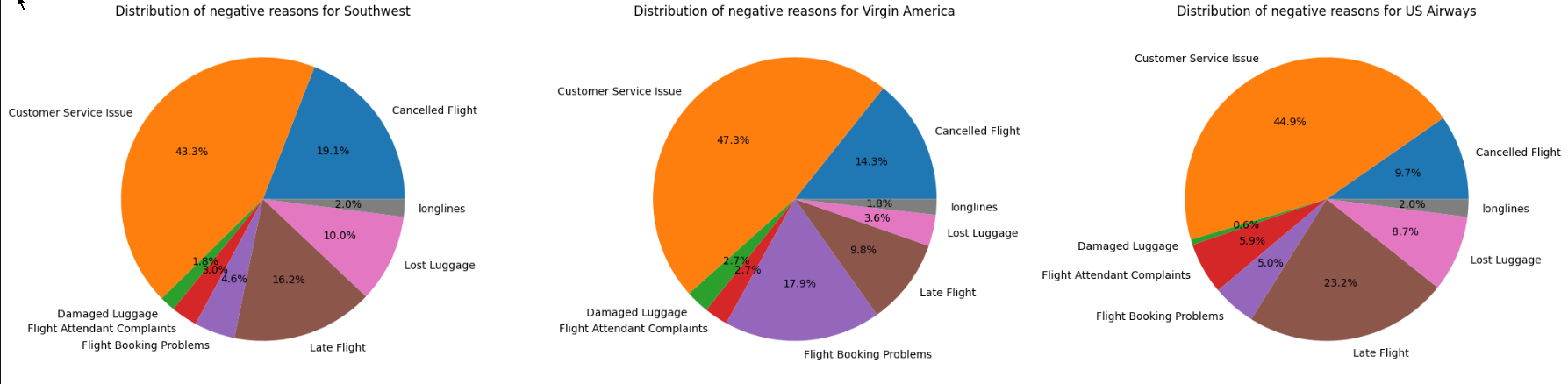
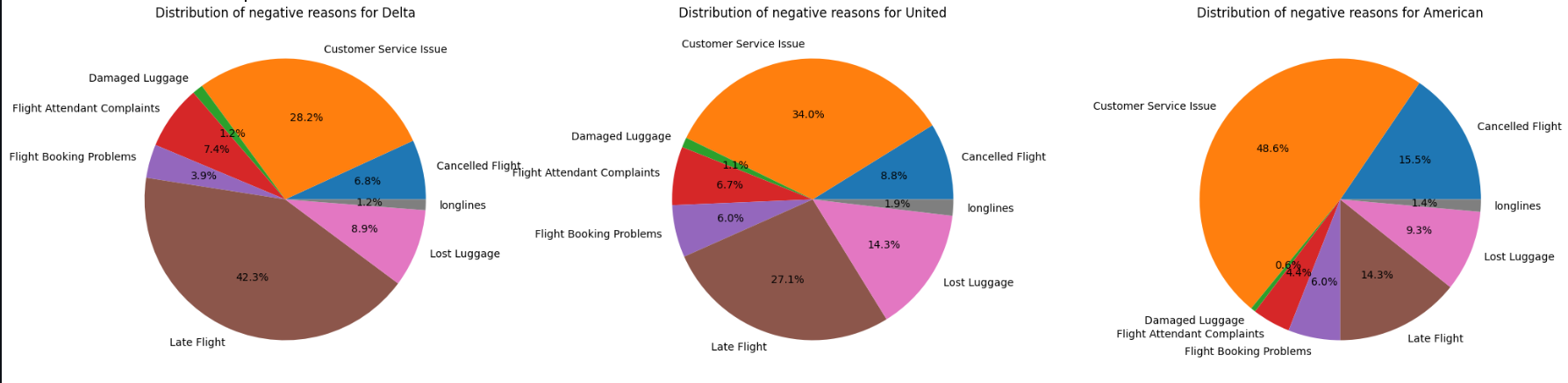
A screen shot of a computer

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**

In the data modeling phase, we applied various clustering algorithms to analyze the airline sentiment data obtained from Twitter. Specifically, we utilized k-means clustering, agglomerative clustering, and DBSCAN. These algorithms aimed to identify patterns and group similar data points together based on the features of negativereason\_confidence and airline\_sentiment\_confidence.

Evaluation:

Based on the evaluation metrics, we compared the performance of three clustering models: k-means, agglomerative clustering, and DBSCAN. Here are the results:

K-means Clustering:

Silhouette Score: 0.5671

Calinski-Harabasz Index: 10.0281

Davies-Bouldin Index: 0.2622

Agglomerative Clustering:

Silhouette Score: 0.5343

Calinski-Harabasz Index: 10.7066

Davies-Bouldin Index: 11.8348

DBSCAN:

Silhouette Score: 0.6052

Calinski-Harabasz Index: 4.9518

Davies-Bouldin Index: 2.2967

Key Findings and Insights:

DBSCAN demonstrated the highest Silhouette Score among the three models, indicating better clustering fit. This suggests that DBSCAN was able to identify cohesive clusters in the Twitter sentiment data related to the airline industry.

K-means clustering exhibited the lowest Davies-Bouldin Index, suggesting well-separated clusters. This indicates that the clusters identified by K-means were distinct and separated from each other.

Agglomerative clustering had a higher Davies-Bouldin Index and lower Silhouette Score compared to the other models. This implies that the clusters generated by agglomerative clustering were less well-separated and cohesive.

The results suggest that DBSCAN is a suitable choice for sentiment analysis on Twitter data in the airline industry. Its ability to handle arbitrary-shaped clusters and its robustness against outliers make it a valuable tool for capturing sentiment patterns.

**Discussion**

Through the application of clustering algorithms to the airline sentiment data, we were able to uncover meaningful insights regarding common sentiment patterns in the industry. Our analysis revealed distinct clusters associated with negative sentiments, including customer service issues, late flights, and flight cancellations. These findings offer actionable recommendations for airlines to address these specific areas and enhance their overall service quality.

By understanding the trends and patterns in customer sentiment, airlines can tailor their services and communication strategies accordingly, resulting in a more satisfying customer experience. This data-driven approach empowers airlines to make informed decisions and drive improvements in customer satisfaction.

Among the clustering models evaluated, DBSCAN stood out as the most effective, as it demonstrated a higher Silhouette Score and successfully captured the sentiment patterns. The insights gained from these clustering algorithms highlight the most prevalent negative sentiment issues, providing valuable guidance for airlines to refine their services and effectively address customer concerns.

Furthermore, our exploration of classification algorithms revealed that Decision Tree and Multinomial Naive Bayes models were not suitable for this dataset. The dataset's predominant focus on negative sentiments. Specifically, instances with positive sentiment exhibited atypical characteristics, including 'airline\_sentiment\_confidence' values falling within the range of 0 to 1, a blank 'negativereason' field, and 'negativereason\_confidence' values either being blank or 0. These non-standard attributes deviate from the typical classification criteria, rendering traditional classification algorithms ineffective for this subset of instances.

These findings highlight the importance of exploring alternative approaches, such as clustering, to analyze and understand sentiment patterns within the airline industry dataset. By leveraging clustering algorithms and excluding attributes that hinder classification, we were able to gain significant insights into the sentiment landscape. These insights have the potential to drive meaningful improvements in customer satisfaction and guide airlines in addressing the specific issues highlighted by the negative sentiment clusters. Overall, our analysis combining clustering techniques and attribute exclusion has provided valuable information for understanding the sentiment dynamics in this specific context.

**Conclusion:**

In conclusion, our analysis of airline sentiment data from Twitter using clustering algorithms has provided valuable insights into the patterns and common negative sentiments expressed by users. We applied k-means clustering, agglomerative clustering, and DBSCAN to group similar data points based on negativereason\_confidence and airline\_sentiment\_confidence.

Among the clustering algorithms, DBSCAN demonstrated the best performance with a higher Silhouette Score, indicating cohesive clusters, and a lower Davies-Bouldin Index, suggesting well-separated clusters. This highlights the effectiveness of DBSCAN in identifying and grouping tweets with similar negative sentiments.

Through the clustering analysis, we identified prevalent issues such as customer service problems, late flights, and canceled flights as the most common sources of negative sentiment in the airline industry. These findings can help airlines focus on addressing these specific issues and improving their services in these areas.

Possible Future Directions:

* Sentiment Analysis Refinement: Further refine the sentiment analysis by exploring advanced natural language processing techniques, including sentiment lexicons, deep learning models, or incorporating contextual information to enhance the accuracy and granularity of sentiment classification.
* Real-time Monitoring: Implement a real-time monitoring system to analyze and classify tweets as they are posted. This would enable airlines to promptly address customer concerns, identify emerging issues, and take immediate actions to resolve them, thereby enhancing customer satisfaction and loyalty.
* Social Media Engagement: Leverage the insights gained from sentiment analysis to improve social media engagement strategies. Airlines can proactively engage with customers, respond to their concerns, and provide timely assistance, ultimately enhancing customer experience and building a positive brand image.
* Predictive Analytics: Explore predictive analytics techniques to forecast potential negative sentiment trends and identify factors that may lead to negative experiences. This can help airlines take proactive measures to prevent or mitigate issues, leading to improved customer satisfaction and loyalty.
* Integration with Customer Feedback: Integrate sentiment analysis results from Twitter with other customer feedback channels, such as surveys and reviews, to gain a comprehensive understanding of customer sentiments and opinions across different touchpoints. This holistic approach can provide a deeper insight into customer preferences and expectations.

By pursuing these future directions, the airline industry can leverage sentiment analysis and clustering techniques to gain a competitive edge, enhance customer satisfaction, and drive continuous improvements in their services and overall customer experience.