**Airline Sentiment Analysis**

**Jason Moore and Mukul Gharpure**

**Fall 2023 - INFO - H516**

**Faculty - Sunandan Chakraborty, Ph.D.**

**SOIC, IUPUI – Indianapolis**

**Introduction**

In an era where air travel is synonymous with global connectivity, understanding and responding to passenger sentiment have become central to the success of airlines. The vast landscape of digital communication platforms provides a rich source of feedback, ranging from social media posts and online reviews to direct customer interactions. This report is dedicated to unraveling the intricacies of airline sentiment analysis, examining the methodologies employed to gauge customer satisfaction and identify key pain points. From sentiment expressed in tweets to the nuances captured in customer service interactions, we explore the diverse facets of passenger sentiment and its implications for the aviation industry. To navigate the complexities of airline sentiment analysis, this report advocates the integration of cloud computing and PySpark, an open-source distributed data processing framework. Cloud computing's scalability and accessibility provide an effective means to manage and process the vast and diverse datasets generated by airline passengers. Concurrently, PySpark's distributed computing capabilities enhance the efficiency and speed of sentiment analysis workflows, allowing for seamless processing of large datasets. By synergizing these technologies, our approach aims to extract actionable insights from the wealth of available data, empowering airlines to make informed decisions, enhance customer satisfaction, and chart a course towards a more responsive and competitive future in the aviation industry.

**Data Description**

We sourced the airline sentiment dataset from Kaggle. The dataset has 15 attributes but we only used 4 in the dataset. Here is a brief description of each attribute that is used for this project:

* **Airline sentiment:** This attribute indicates the sentiment of the tweet, which can be positive, neutral, or negative. For example, a tweet that says “I love flying with @VirginAmerica” would have a positive sentiment, while a tweet that says “I hate flying with @united” would have a negative sentiment.
* **Negative reason:** This attribute indicates the reason for the negative sentiment, if any. The possible reasons are: Bad Flight, Can’t Tell, Cancelled Flight, Customer Service Issue, Damaged Luggage, Flight Attendant Complaints, Flight Booking Problems, Late Flight, Lost Luggage, and longlines. For example, a tweet that says “My flight was cancelled and I had to wait for hours” would have a negative reason of Cancelled Flight.
* **Airline:** This attribute indicates the airline that is being tweeted about. The possible airlines are: American, Delta, Southwest, United, US Airways, and Virgin America. For example, a tweet that says “@SouthwestAir thanks for the free wifi” would have an airline of Southwest.
* **Tweet text:** This attribute indicates the tweet made by a user of flight and their feedback on the flight’s service.

We aimed to analyze the sentiment of airline customers on Twitter and how it affects their future travel choices. However, due to the recent changes in the Twitter API policies, we could not collect enough tweets for our test case (only 1500 tweets were available). Therefore, we decided to create our own web platform to gather user reviews directly from the customers. We stored the collected data in huge files on AWS storage. The data contained the following variables: Tweets, Airlines, and Likelihood of Future Use.

**Methodology**:

Employing PySpark as our tool of choice, we executed three distinct machine learning models—logistic regression, Naïve Bayes, and Support Vector Machines (SVM). These models served as the backbone for our comprehensive analysis, enabling us to delve into the data, predict sentiments, and assess the likelihood of future airline usage. Beyond the realm of predictive modeling, our approach encompassed additional exploratory data analyses, including a nuanced examination of flight moods. To enhance user interaction and gather real-time feedback on flights, we crafted a web interface. This interface not only facilitated the collection of user reviews but also leveraged time series analysis to dynamically update graphs, ensuring that mood and customer sentiment representations evolved in real-time with each review, providing a holistic and responsive perspective on the airline's performance.

**Figure 1 The Framework**:

A diagram of a software application

Description automatically generated

**Figure 2 the web interface:**

A screenshot of a login

Description automatically generated

**Figure 3 graph of customer sentiment:**

A graph with different colored bars

Description automatically generated

**Results:**

**A screenshot of a graph

Description automatically generated**

A close-up of a pink box

Description automatically generated

The visual representation illustrates the performance metrics of our four employed machine learning models: Logistic Regression, Naive Bayes, Support Vector Machine, and Random Forest. The accompanying table provides a comparative analysis of the accuracy achieved by each model and the underlying framework utilized for their training. Notably, Logistic Regression emerges as the leader with a remarkable accuracy of 91.82%, trailed by Support Vector Machine at 88.52%, Naive Bayes at 86.57%, and Random Forest at 85.71%. It is imperative to underscore that all models underwent training using PySpark ML, a robust framework tailored for scalable machine learning on extensive datasets.

Beyond the numerical benchmarks, these models were deployed to scrutinize social media data, unraveling profound insights into customer satisfaction and behavior. The predicted sentiment text, as showcased in the web interface image above, exemplifies a practical application of our machine learning models. In this instance, the sentence analyzed revealed a negative sentiment with a confidence score of 0.52. This negative confidence score serves as an indicator of the model's degree of certainty regarding the negative sentiment expressed in the given text. A lower confidence score may suggest increased ambiguity or nuance in sentiment classification, prompting a closer examination of the specific linguistic cues that influenced the model's prediction.

**Conclusion**:

In this project, we transcended conventional sentiment analysis, delving into the narratives embedded within tweets to unveil a richer tapestry of customer experiences. Leveraging a vast dataset and the prowess of cloud computing, we deciphered the intricacies of service quality in the aviation industry. The project's benefits extend beyond mere analysis, serving as a compass for airlines to navigate toward improved customer experiences. By gaining insights into passenger preferences and expectations, the project lays the foundation for a platform fostering open dialogue and feedback, ensuring a more responsive and customer-centric air travel industry. The impact of this endeavor is twofold: empowering airlines with actionable knowledge and guidance, and amplifying the voices of passengers in shaping the trajectory of the industry. As we conclude, this project not only charts a course for enhanced customer satisfaction but also signifies a transformative step towards fostering a more dynamic, informed, and inclusive air travel ecosystem.

**References**:

Appendix 1: Contribution from team members

Jason Moore:

* Report Writing
* Presentation
* Conducting Project Meetings
* Method analysis

Mukul Gharupre:

* Data preprocessing
* Designing front end and back end
* Spark programming
* Final report editing