Analyzing Customer Reviews for Top 5 US Based Airlines

Ajay Iyer

University of Arizona

05/03/2024

Table of Contents

Abstract/Executive Summary

Airline Equality is reviewing website under the Skytrax model where customers are sharing their experience with airlines they take for their travel needs. Customers are allowed to share their view on the airline, airport lounge, whole experience, etc. Following a personal experience of mishandled baggage during trip to India, this study seeks to analyze customer reviews to evaluate the performance of the top 5 airlines in the United States of America. The analysis utilizes sentiment analysis and topic modeling techniques to extract insights from unlabeled text data obtained from the Airline Equality website. The research aim to identify common sentiments and topics among customer reviews, providing valuable insights for future travel decision-making. Through Python programming and AWS for automation, the study explores different models to determine the most effective approach for sentiment and topic analysis of airline reviews along with determining how we can automate this process to make faster and seamless for better decision-making. By examining these trends and patterns provided by the customers, this paper contributes to providing some discoveries for addressing issues with airlines and enhancing transparency accountability with the airline industry ultimately empowering travels to make informed decisions.

Project Plan

Primary Company Details

Established 1989

Based in London, United Kingdom

Skytrax brand associated with Quality Excellence

Company Communication:

Website: airlineequality.com

Business Description:

A fair consumer forum, the Air Travel Review website has grown to become one of the top review sites for traveler reviews related to airlines, airports, and related air travel. The airline and airport star ratings, the World Airline Awards, the Airport Awards, and the Skytrax moniker are all globally recognized as symbols of quality excellence in the aviation business. Users are able to search to for reviews left by previous customers on airlines, airports, and their respective lounges. Users are able to provide information that was related to their travel reviews like their itinerary or route, seat type, type of travel as well ratings of seat comfort, staff service, food & beverages, in-flight entertainment, ground service, Wi-Fi connectivity, value, and if they recommend it.

Business Analysis/Opportunity:

The data will be sourced from this website using a web scrape script developed in Python to ingest the data as a csv file. Each file/dataset will have an X amount of records and consists of 6 columns or attributes which consists of Reviews, Type of Traveler, Seat Type, Route, Date Flown, and Recommended. The last attribute, Recommended, is if the customer would recommend the airline to anyone. After preparing a few research questions, I will be using various supervised and unsupervised techniques to determine the sentiment and the retrieving the topics that associate with each review. These reviews and ratings are there provided by many airports and airlines across the world. The use of this source will provide a lot of knowledge and starting with these research questions will give a better understanding of the airline industry.

Research Questions:

As prices are rising and staff mishandling personal belongings, it becomes a concern to the customer to determine which airline could cause the least number of problems during their respective travel. In order to analyze these problems, we need to have a process using machine learning that will help us analyze the concerns customers have with respective airlines. This information will prove to be very valuable for the airline to determine what fixes are needed. The following questions will be researched in this project.

Research Question #1: **Are there common sentiment patterns in reviews based on specific airlines?**

Studying the sentiment patterns gives the business a better understanding the satisfaction levels from the customers across various airlines. Negative sentiments can determine what needs to be focused on for continuous improvement. Monitoring these sentiment trends enables airlines to address the issues and respond to any customer feedback before it escalates. This is able to protect their brand and building some credibility among customers. Eventually, this can also help with predictive analytics and forecasting by analyzing historical trneds and making predictions on customer behavior and adapt the strategy accordingly.

Research Question #2: **What are the most frequent complaints in passenger reviews, and does it vary by the airline?**

Understanding the most frequent complaints brought to the attention by the customer can help us identify the issue with not being able to meet customers' expectations. This can help decide on the prioritization of what needs to be fixed as soon as possible to alleviate the burden off the customers. Investigating these variations in complaints allows to enhance the customer experience and guiding the strategies that need to be made to strive for excellence. Analyzing this can give an advantage on the airline industry and the market.

Research Question #3: **Are there differences in review sentiments based on the class (economy, business, first class)?**

This is another aspect that can provide us with more information. With the amount of money that needs to be spent for purchasing a any class ticket, there has to be some concern on offering the best service for the customer. In order to provide the best services, there will be an analyzation across the classes to determine where the attention is needed.

Research Question #4: **How has the sentiment of passengers evolved over time by the airline?**

As mentioned before, analyzing the sentiments across time, and tying it with topics helps us understand if the concerns are going to be addressed when making changes to the business model. Along with this, we can introduce predictive analytics in the future to be able to address the concerns sooner.

These questions are beneficial to see where airlines need to improve in their services to get more customers to choose a specific airline. Analyzing the issues/concerns that passengers are facing can help dive deeper. I think airlines themselves can benefit from this study since it is improving their service. The Federal Aviation Administration (FAA) can also benefit from this study to make decisions and policies for passengers' satisfaction and safety. Investors and Stakeholders can also take this study further to gauge customer satisfaction and how it can affect financial performance of an airline.

Hypotheses:

Hypothesis #1: There will be statistically significant differences in sentiment patterns between the customer reviews of various airlines. Passengers will tend to express positive things about the airline when it relates to amenities, on-time, and customer service while negative sentiments will explain the low-quality amenities, delays, food, and poor customer service.

Hypothesis #2: Passengers will frequently mention complaints that are related to itinerary disruptions (delays, cancellations), food, baggage handling, and the cabin experience. The frequency of complaints will emphasize how common airlines across the industry experience the same issues as they tend to opt for lower quality materials to increase profits as per their business model. Since these are the top airlines of the country, expectations are set that there fewer issues on comfort than delays and itinerary changes.

Hypothesis #3: Passengers will tend to spread more positive reviews on the in cabin experience in Business or First compared to those flying in Economy. The difference will be the amenities, quality of service, food, and comfort increase as we get to the higher travel classes.

Hypothesis #4: Sentiments towards airlines will show a complex evolution over time. Airlines are known for striving for high service quality. The impact of social media platforms allows us to raise concerns and make it more vocal. Economic factors and stock market performance can lead to issues with how airline can be perceived after diving into the reasons. Global Events are also something to think about. Health concerns like the Covid-19 Pandemic or airline accidents could play an impact on an airline can be perceived.

Data:

The data is combined with the reviews that were sourced from Airline Equality along with respective statistics from each review ingested as a CSV file. The dataset consists of various breweries that have reviews for their respective beer based on what the reviewers look for. The dataset consists of 6 attributes which includes the ‘review’ column. This column contains the text review that is retrieved from the website. The next attribute is ‘Type of Travel’ which gives more idea on the category of the trip like Couple Leisure, Solo Leisure, Business, Family Leisure, etc. Next is ‘Seat Type’ which shows where their seat was for the trip. Categories that will show up but not limited to are Economy Class, Premium Economy, Business Class, and First Class. The next attribute is the 'Route' attribute that shows their itinerary of the trip. Another attribute is 'Date Flown' which displays the month and year the customer traveled. The last attribute is 'Recommended' which is if the customer would recommend the airline to anyone.

Reviews: Text review provided by the traveler.

Type of Travel: Text category (Couple Leisure, Solo Leisure, Family Leisure, Business, etc)

Seat Type: Text Category (First Class, Business Class, Economy, etc)

Route: Text (Chicago to Phoenix)

Recommended: Text category (Yes/No)

Measurements:

It is very important to identify the need for this type of information and how the use of these customer reviews can help assess the performance of the airline industry. When working with this data, it is important to remember that these reviews come from a consumer and we are storing any personal/confidential information like names, birthdates, etc. It is crucial to keep in mind that users control the website and that their main goals are to promote awareness and educate the airline and other potential customers about respective experiences. Therefore, it is critical that we understand the importance of every review and how it affects the customers' involvement so we would need to gauge the customer engagement across airlines.

Methodology:

This project will use Python & AWS to generate charts and other useful information on a monthly basis for airline companies' better decision making. AWS Lambda will store the script that will web scrape Airline Quality website for customer reviews and basic trip statistics on a monthly basis and convert it into CSV file. S3 will have the web scraped raw data as a backup file. SageMaker will be used to test and perform machine learning algorithms to analyze the text data and evaluating the performance of the models in Python. AWS Athena/Snowflake/DocumentDB will be used to retrieve the updated data with sentiment scoring and/or topic modeling using SQL and connect to any visualization tool like AWS Quicksight. Quicksight/Thirdparty Services will be used to generate the charts/dashboard and have it updated on a monthly basis. Using Python, we will also be developing the machine learning models for sentiment analysis and topic modeling. For sentiment analysis, Logistic Regression, Vader Sentiment Scoring, and BERT, a Transformers Model. For the Topic Modeling, Latent Dirichlet Allocation (LDA) will be evaluated to find the best version while hyper tuning.

Computational Methods:

Combination of supervised and unsupervised techniques will be used for determining which model can accurately give us sentiments and its respective topics. Even though we are working with unlabeled text data, we will manually label and create a train dataset by random for evaluating the model. Accuracy scores, Recall, F1-Score, and Precision will give a better idea of the performance of the model. Same thing will be done to evaluate the unsupervised techniques since now we have a train dataset. For LDA topic modeling technique, we will assess the coherence score for determining which model would be the best for use.

Implementation:

There are so many reviews that are being processed today on various platforms for consolidation. As airlines uncover issues, it can be helpful to determine with the right guidance where improvement is needed. Disregarding these issues can prevent major issues that might disrupt the business. For research question 1, "**Are there common sentiment patterns in reviews based on specific airlines?**", we can use supervised and unsupervised sentiment analysis techniques to give us a better understanding of the data. For research question 2, "**What are the most frequent complaints in passenger reviews?**", topic modeling will come to use after finding the best model for our sentiment related tasks. Using this, we can identify a range of topics that show up during a review and be able to categorize them accordingly to determine if they are frequent complaints. For question 3 "**Are there differences in review sentiments based on the class (economy, business, first class)?**", With the statistics that are ingested along with each review, this will give us a better idea of where complaints are sourced from to target those areas as a need of improvement. For question 4 "**How has the sentiment of passengers evolved over time by the airline?**", time plots can be displayed to show if past solutions were solving the issues customers were bringing up along with any predicting functionality to this solution.

Literature Review

The study was published last year by Aksh Patel, Parita Oza, and Smita Agarwal that has the same idea on analyzing customer reviews for airline services. Data was retrieved from Kaggle as their airline database to perform sentiment analysis. These researchers used multiple machine learning techniques such as Naïve Bayes, Decision Trees, Random Forest, Support Vector Machine, AdaBoost, and comparing them with the BERT model. Pre-processing techniques that were performed included tokenization and the eliminating of stop words. TF-IDF and chi-square algorithms extracted features from a bag of words. These researchers used accuracy, precision, recall, and F-measure for their metrics to compare the performance of the algorithms. This differs from my project since I will be using different natural language processing techniques that I researched on which are VADER Sentiment Scoring and Latent Dirichlet Allocation on unlabeled text data for gaining more insights.

The study was performed by researchers in Korea (Hye-Jin Kwon, Hyun-Jeong Ban, Jae-Kyoon Jun, & Hak-Seon Kim) who wanted to analyze Asian airline reviews. They retrieved data from the same website as my project and performed sentiment analysis and topic modeling using R. They were able to perform topic modeling using Latent Dirichlet Allocation while using TidyText for sentiment analysis on the data retrieved from Skytrax. We have some similarities with how we are going to perform topic modeling, however, I found out that word clouds would be a cool idea on presenting what were the most used words I will be using Python for NLP related tasks and I will be using sentiment scoring which would score each review and we can determine if it is positive, neutral, or negative. Also, I will be automating a portion of the process so we can have updated reviews and changes can be reflected in batches.

Methodology:

Data Ingestion

The data ingestion process includes one Lambda function created using AWS web scrape the data from the customer forum website then ingest it as a csv file in an AWS S3 Bucket. Each airline will be under 1 bucket but separate folders to keep it organized. Once loaded into S3, separate functions are put in place to execute the respective AWS SageMaker notebook that will execute the best model along with providing some basic plots to analyze the data. After final dataset is created with respective transformations and findings along with an S3 load, that dataset is loaded into Athena for querying if needed. Using AWS Athena, we are able to set some basic analyzation queries that can be developed into plots through AWS QuickSight for updating plots and diagrams.

Sentiment Analysis & Topic Modeling

For the sentiment analysis part of the project, multiple algorithms will be tested and analyzed for improvements. The first algorithm that was tested is the Logistic Regression with multiclass classification. The reviews that ingested may have positive, neutral, or negative reviews so it is important to account for those. In this case, Positive will be assigned 1, Neutral: 0, and Negative: -1. The metrics that were used to evaluate for these models were accuracy score and ROC AUC scores. Since the reviews are unlabeled, a decision was made to manually label 100 reviews so there would be some random sample of a dataset that represents our train/test set. Another model that will be used for testing purposes is Vader Sentiment Scoring. Valence Aware Dictionary and sEntiment Reasoner (VADER) is a pretrained sentiment analysis model that provides a score for any given text. It is built on pre-built large dictionary of words that are assigned sentiment scores. Even though scores range from -4 to 4 for each word, it also considers factors like capitalization which will indicate stronger sentiments. The compound score will help us identify the sentiment for a given text. If the compound score is less -0.05 and between 0.05, this is considered neutral. If the score is higher than 0.05, then the review is considered positive. If the review is less than 0.05 then the review is considered as negative. Same train/test will be used to output scores are used to intrepret the model. The final model used for Sentiment Analysis will include the BERT Transformers Model.

BERT stands for Bidirectional Encoder Representations from Transformers which a pre-trained model introduced by Google which revolutionized Natural Language Processing. BERT is pre-trained using a technique that randomly masks words in a sentence and the model predicts each word based on context. This allows for BERT to gain to a contextual understanding of word representation. It also trains the model to understand the relationships even between sentences. For the topic modeling techniques, Latent Dirichlet Allocation (LDA) algorithm will be used to determine the respective topics for each review. This algorithm will be evaluated by a Coherence Score for the following measures. C_V measure is when Normalized Mutual Information (NPMI) and the cosine similarity are used in an inverse measure which is based on a sliding window and one-set segmentation of the top words. C_UCI is an order of points mutual information (PMI) of every word pair of the given top words, along with a sliding window. C_UMASS measure utilizes a one-preceding differentiation, a logarithmic conditional likelihood as a verification metric, and document coexistence counts. C_NPMI is another version which is improved of the C_UCI but it makes use of normalized pointwise mutual information.

Exploratory Data Analysis (Model Selections):

Logistic Regression

Accuracy Score	ROC AUC Score
0.65	0.62901

VADER Sentiment Scoring

Accuracy Score	ROC AUC Score
0.71	0.7250

BERT Transformers

Accuracy Score	ROC AUC Score
0.83	0.7948

For the first part of the project, a sentiment analyzer was built to help us classify the unlabeled text data. Since supervised models were also going to be used to address the complexity, a random sample of 100 reviews were reviewed manually and labeled to train the model that will be used for the analyzer. After manually labeling the dataset, the model that was chosen for handling multi class classification is Logistic Regression. First part of the process was cleaning the text and removing unnecessary spaces, removing the status of whether the review was verified or not, etc. Also, the reviews were converted to lowercase to bring in some consistency and removing any stop words. Then, the words were lemmatized to have each word in its base form which makes it easier for machine learning models to have tasks like sentiment analysis. After the data cleaning process is completed, then feature extraction is performed to find the most occurring features in the dataset.

Count Vectorizer and Term Frequency-Inverse Document Frequency (TF-IDF) were methods used for feature extraction. Count Vectorizer takes a collection of text documents and converts it into a document-term matrix best for its simplicity, efficiency, and interpretability. TF-IDF is applied to determine a word's significance to a document within a group of documents. Next, a train-test split was conducted at the 80-20 level. 80 reviews for the train and 20 reviews for the test. Standard Scaler was used to standardize the features by subtracting the mean for each data point in a feature by the standard deviation and scaling it to unit variance making all features equal. After training and testing the Logistic Regression model, it was evident that the TF-IDF method was not able to classify any review positively through multiple revisions. The Count Vectorizer was also used to train and the test the model and we received significant reviews classified as positive, so this was the model that was used to compare to the rest.

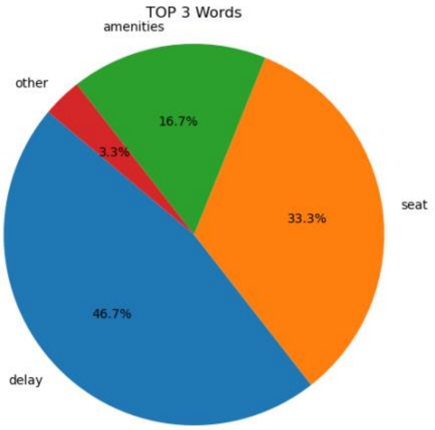
The next model that was used to perform Sentiment Analysis was the VADER Sentiment Scoring unsupervised model. This model is known for not requiring data cleaning and it is able to interpret the raw text. Similar text cleaning techniques were executed to remove the words that were not required. The Sentiment Intensity Analyzer was used to retrieve the sentiments through a compound score. Compound scores are a consolidated metric that gauges the overall sentiment of the text. Next, decision was made for the scores between -0.05 and 0.05 will be categorized as 'Neutral', if it is greater than 0.05 then we classify it as 'Positive', and if it is less than -0.05 then it is classified as 'Negative'. After looking at the accuracy scores and roc auc scores, a conclusion can be made to use the BERT Transformers model is consistently performing better on both measures as our sentiment analyzer.

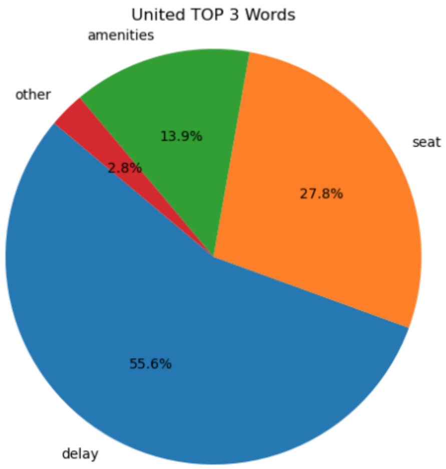
Model	Coherence Score
C_V	0.421
C_UCI	-1.627
U_MASS	0.238
C_NPMI	0.395

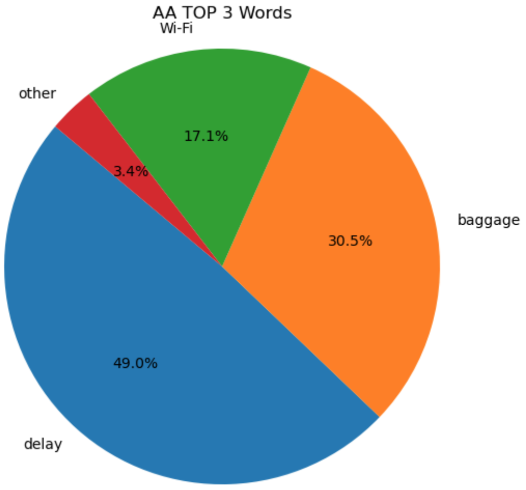
For the topic modeling part of the project, one algorithm was used and that was Latent Dirichlet Allocation (LDA). This technique assumes that documents are composed of various topics, and it is finding the underlying categories to that would not have been thought to be considered. Finding the top 5 topics was necessary for this project and different measures were taken to determine which model is the best. Using similar text cleaning processes and lemmatization, executing, and interpreting these models is key to finding the one that represents the capturing the structures well. Out of all the models that were tested, a conclusion can be made that the C_V measured LDA model is the best representation of capturing the relationships correctly out of the ones tested.

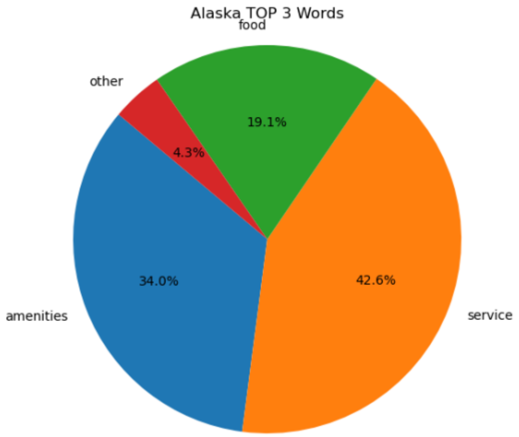
Data Analysis & Visualizations:

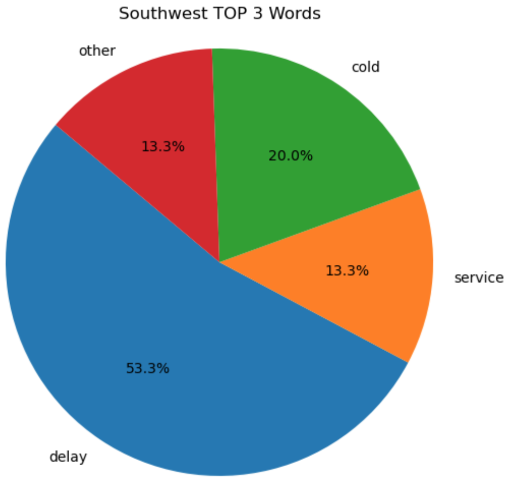
Research Question #2: **What are the most frequent complaints in passenger reviews, and does it vary by the airline?**

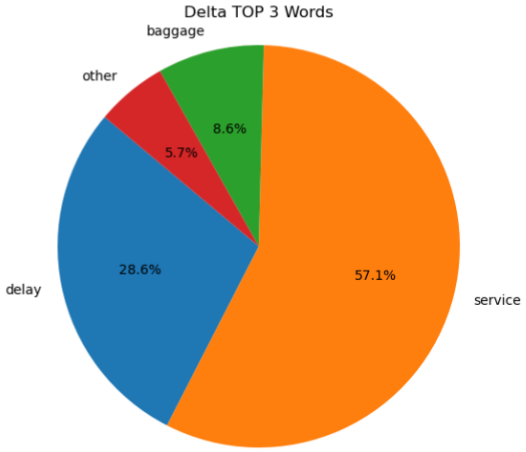






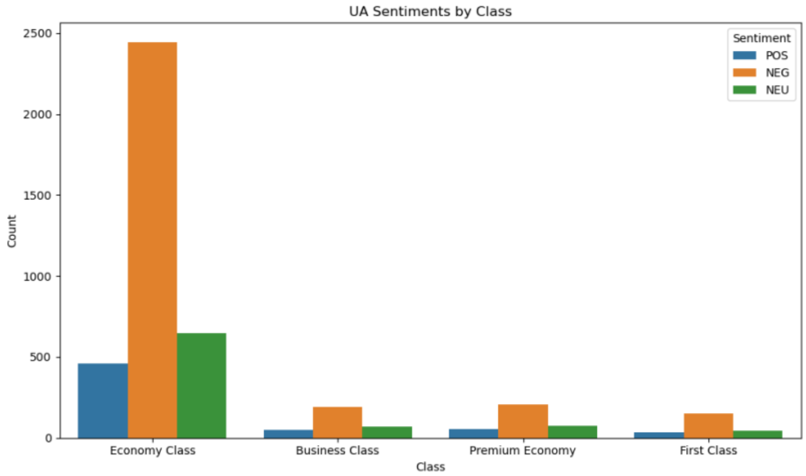

Alaska TOP 3 Words
food
other
19.1%
4.3%
34.0%
42.6%
amenities
service

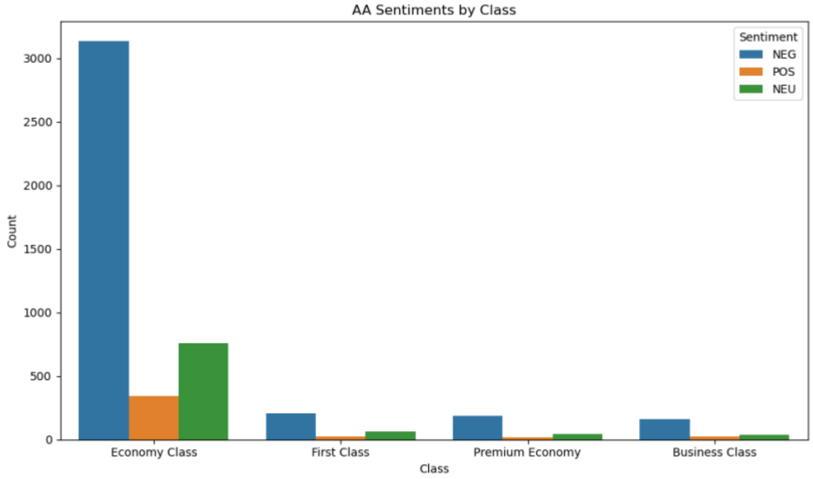

Southwest TOP 3 Words
other
cold
13.3%
20.0%
13.3%
service
53.3%
delay

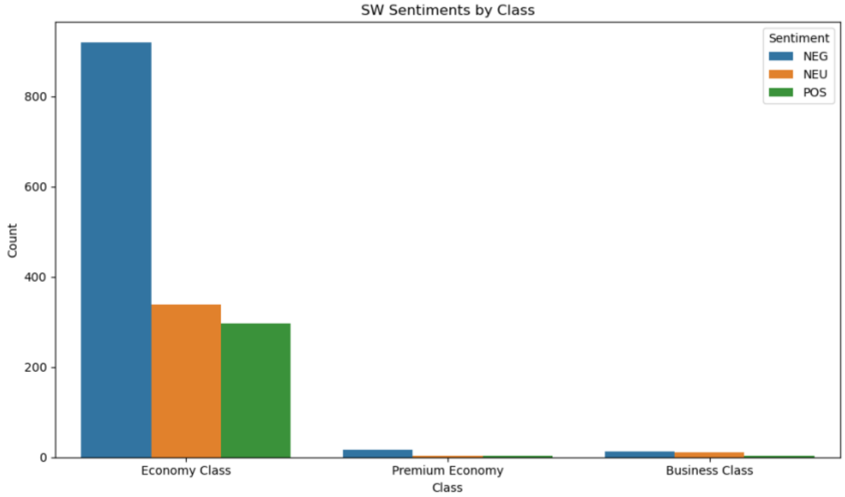

Delta TOP 3 Words
baggage
other
8.6%
5.7%
28.6%
57.1%
delay
service

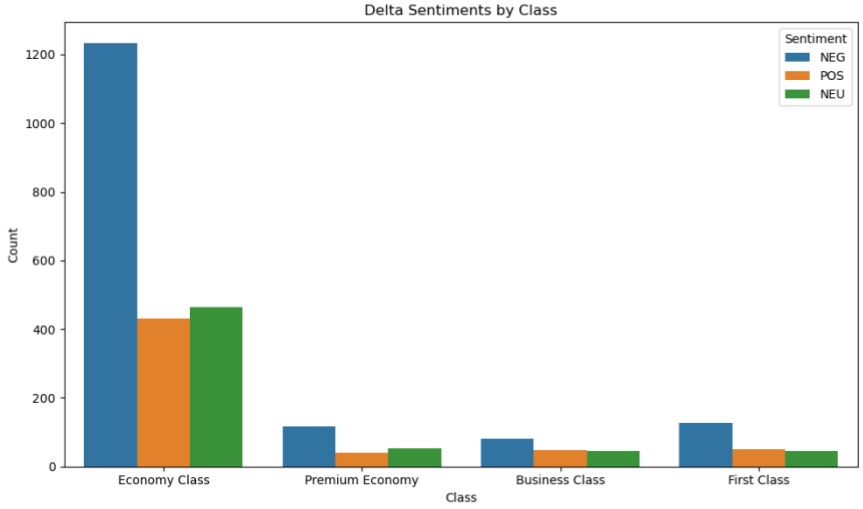
Amenities, delay, and seat are the most common topics that are brought up as passenger complaints. Most of the time, we can say that if a person has complained about an airline, it has to be because of delays that can occur due to multiple different reasons.

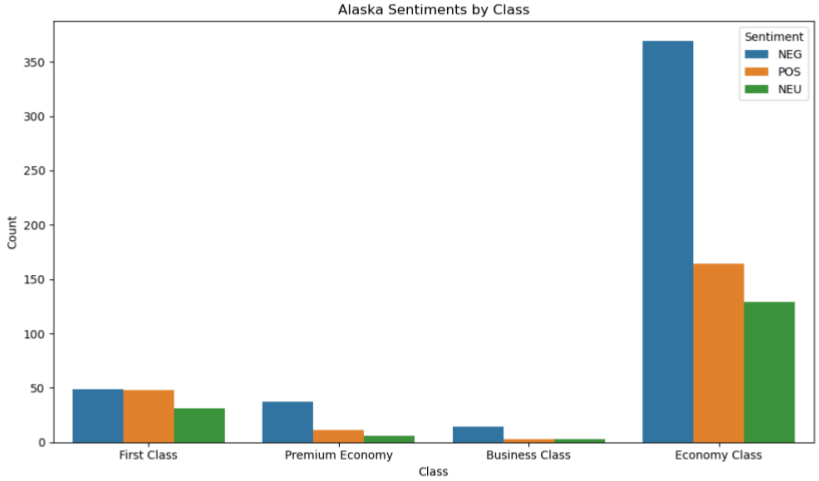
Research Question #3: **Are there differences in review sentiments based on the class (economy, business, first class)?**





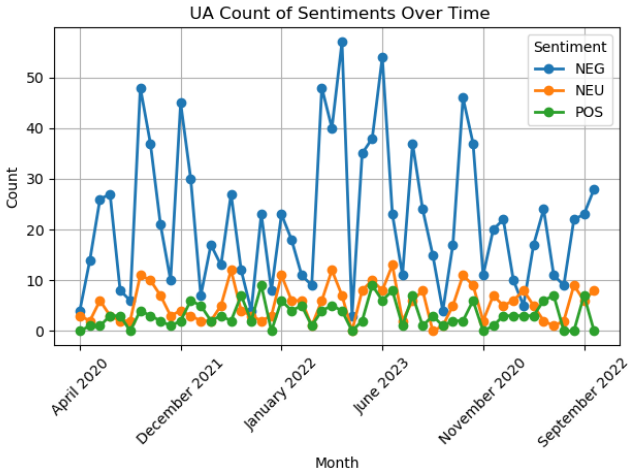

SW Sentiments by Class
Count
0
200
400
600
800
Economy Class
Premium Economy
Business Class
Class
Sentiment
NEG
NEU
POS

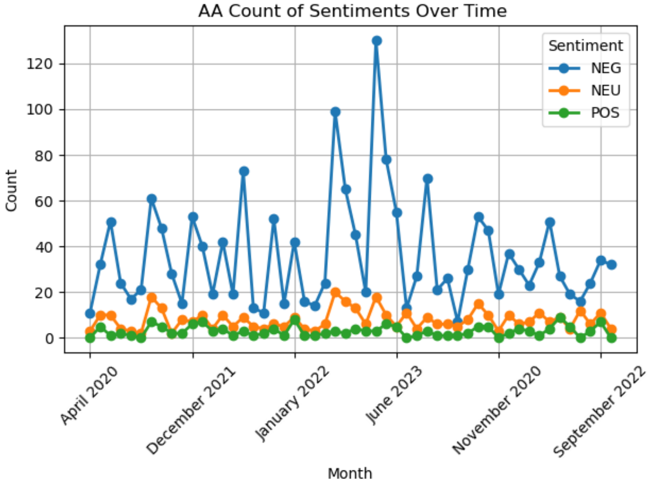

Delta Sentiments by Class
Count
0
200
400
600
800
1000
1200
Economy Class
Premium Economy
Business Class
First Class
Class
Sentiment
NEG
POS
NEU

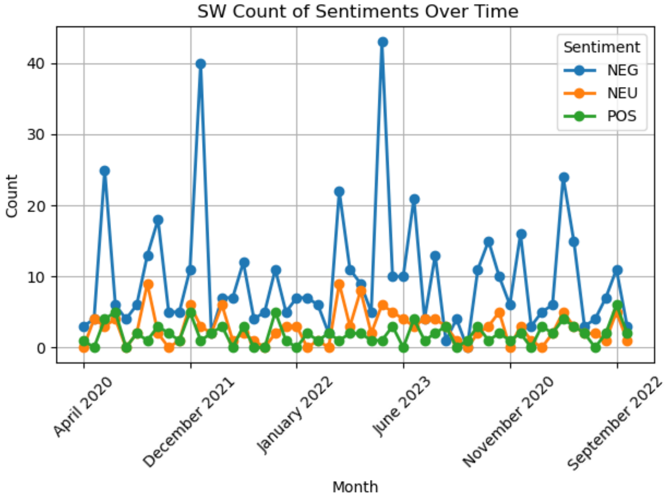

Alaska Sentiments by Class
Count
0
50
100
150
200
250
300
350
First Class
Premium Economy
Business Class
Economy Class
Class
Sentiment
NEG
POS
NEU

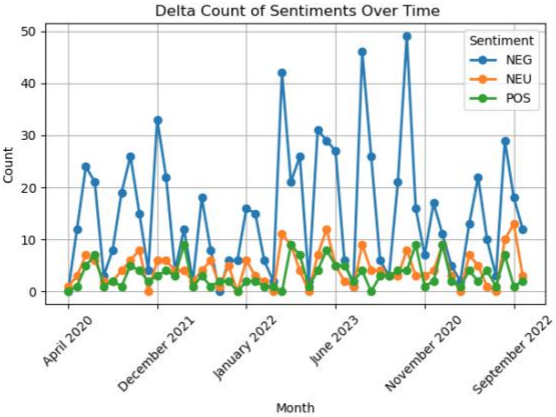
As we look at these graphs, we can see that the common sentiments are negative ones that are reflecting the Economy Class for the most of these airlines. First class has the least number of reviews provided. Economy class has the most amount reviews provided. Economy class is consistently being mentioned in reviews to

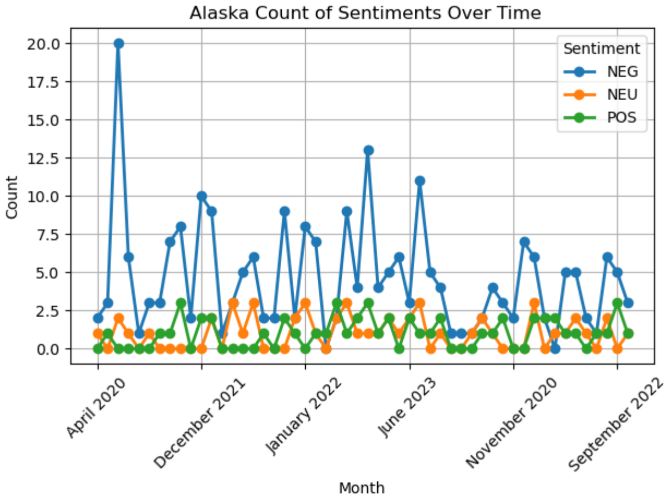
Research Question #4: **How has the sentiment of passengers evolved over time by the airline?**




AA Count of Sentiments Over Time
Count
0
20
40
60
80
100
120
Month
April 2020
December 2021
January 2022
June 2023
November 2020
September 2022
Sentiment
NEG
NEU
POS


SW Count of Sentiments Over Time
Count
0
10
20
30
40
Month
April 2020
December 2021
January 2022
June 2023
November 2020
September 2022
Sentiment
NEG
NEU
POS


Delta Count of Sentiments Over Time
Count
0
10
20
30
40
50
Month
April 2020
December 2021
January 2022
June 2023
November 2020
September 2022
Sentiment
NEG
NEU
POS



We are able to see that there is a significant drop in the number of negative reviews that were provided between past years and compared to 2023. There are still lots of improvement that would be recommended. It will be interesting to see where 2024 will make it for these airlines by the end of the year.

Research Question #1: **Are there common sentiment patterns in reviews based on specific airlines?**

Overall, we are able to see that common sentiment patterns with a lot of negative reviews related with a class but most of them are discussing some delays that could have occurred due weather conditions, etc.

Ethical Recommendations:

Analyzing airline customer reviews through sentiment analysis and topic modeling techniques reveals detailed insights into the dynamics between passengers and the airline. The data extracted from these reviews can impact the customers and even some business decisions that will need to be made to resolve these consistent issues. By retrieving new data and insights from customer feedback, we can identify patterns that drive increased traffic and attention to other airlines in order to give insights if their business model needs to change.

Prioritizing openness in the methods utilized for data collection and analysis is crucial. Clients should be informed that their data is being gathered for analysis but it’s also crucial to clarify that this data is being collected at random. By safeguarding their privacy, confidential information will not be used to evaluate the reviews they made. Best practices are also implemented to avoid misuse and illegal access to this information. To guarantee that there is an equal representation in these models, it is also important to mitigate the bias as much as possible. Regularly assessing and adjusting the models to ensure fair treatment for all categories is a best practice for addressing these potential concerns that may arise with skewed datasets. By following these ethical recommendations, lifelong learners can navigate in this space are able to navigate this side of analysis with accountability and respect for others’ rights and values along with other ethical principles.

Challenges:

When conducting this research, there were always be some unknowns/challenges that pop up during development that were not expected when proposing the idea of the project. In the process of conducting this research, it is important to acknowledge limitations that impacted the study. One issue that was faced was underestimating of efforts that needed to be put for certain tasks. For example, I was stuck on how to develop the Lambda function due to various errors. I had to figure out how to give the right access to S3 and importing the right dependencies in order for the script to execute. Another issue was testing the algorithms that were used. In order to test the algorithms, I had to manually label 100 random reviews with a positive or negative sentiment in order to train a supervised model. This was a whole effort that was not encountered for which led to improper quality checks. Another challenge I faced during the development was when the data had to include the star ratings from each review so the sentiments can also be analyzed with the customer ratings. This issue led to not ingesting those ratings which could give us more information on the customer's experience. Some challenges that were overcome is when working with unlabeled text data made it difficult to train machine learning models to retrieve labels without guidance. In order to train the models, a portion of the data had to be manually labeled. Obtaining the star ratings earlier, which may have helped with comprehension was another challenge. Technical complexities impeded the procedure due to the lack of experience with web scraping. Understanding these issues is crucial because it can be utilized for refining future research from data ingestion to machine learning models development on unlabeled text data.

Recommendations and Next Steps:

Overall, this was a thorough analysis that was completed for analyzing customer reviews for airlines. During the development, there were some concerns that could be addressed for further enhancements. One thing is being able to retrieve the star ratings that were provided by the customer. These ratings can reveal other insights by determining if there are any correlations between the ratings and review. with the models we tested did not work out as expected since there are some accuracy concerns. Training and testing methods can also change along with more hyper tuning models before being able to come to conclusion that we can remove these models from consideration. Another is identifying alternative topic modeling methods compared to the LDA model. Singular Value Decomposition (SVD) is a dimensionality reduction approach called Latent Sentiment Analysis (LSA) or Latent Sentiment Indexing (LSI), which represents documents and phrases as vectors in a high-dimensional space. This identifies the patterns in numerous word combinations to for capturing subjects. Another item I could not get to was displaying the dashboard through one of AWS services. QuickSight would have been useful for making this a fully automated procedure that can help identify issues for the business. There would need to be additional research on these capabilities before deciding on the final version of the architecture.

References

Airline Quality. https://www.airlinequality.com/

Iyer, Ajay. Us-Airline-Customer-Reviews Github Repository. https://github.com/ajiy01/US-Airline-Customer-Reviews-

Manuel Pérez, Juan. Pysentimiento: A Python Toolkit for Sentiment Analysis SocialNLP Tasks. https://huggingface.co/finiteautomata/bertweet-base-sentiment-analysis?text=the+issues+I+am+experiencing+with+the+data+is+ridiculous

Patel Aksh; Oza, Parita; Agarwal, Smita. Sentiment Analysis of Customer Feedback and Reviews for Airline Services using Language Representation model. Link

Hye-Jin Kwon, Hyun-Jeong Ban, Jae-Kyoon Jun, Hak-Seon Kim Topic Modeling and Sentiment Analysis of Online Review for Airlines. Link