U.S. Airlines Twitter Sentiment Analysis

Marlene Aviles

Department of Data Science

DSC 680: Applied Data Science

Catherine Williams

July 10, 2020

Abstract

Businesses and organizations during this modern age in time manage to rely on the analysis of data created by their customers as a way of improving their products and services. By intertwining the businesses' customer data and data science methodologies such as sentiment analysis, valuable information can be gained for the company. The term sentiment analysis refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information (Wikipedia, 2020). This type of analysis will be applied to this project given that it is widely applied to voice of the customer materials such as reviews, survey responses, and online social media post for different applications from marketing to customer service (Wikipedia, 2020). This aim of this analysis will be to analyze the sentiment of U.S. airline customer tweets published on the social media platform Twitter regarding their experience. The field of Natural Language Processing (NLP) and sentiment analysis will be utilized for this particular project as it involves working with textual data. Classifiers will also be implemented to attempt and predict the sentiment behind the published tweets. This analysis and the yielded results will assist the U.S. airlines in identifying what areas of the business require the most improvement according to the customer's feedback via their feedback in the tweets.

Keywords: twitter sentiment analysis, U.S airlines sentiment analysis, natural language processing (NLP)

U.S. Airlines Twitter Sentiment Analysis

The airlines American, Delta, Southwest, United, U.S. Airways, and Virgin America all share a common goal, which is to provide the best customer experience and services to their customers. The customer is an essential part of their business given that they drive profits and can ultimately help a business succeed or fail. The feedback and information that a customer leaves regarding the business can ultimately cause a positive or negative impact. For this analysis, the tweets addressed to U.S. airlines on the social media platform Twitter will be analyzed, via sentiment analysis. Sentiment analysis is a type of natural language processing problem that can help determine the sentiment or emotion of a piece of text (Lim, 2019). The tweets will be classified as either positive or negative for this project and will be utilized as a way to learn more about the airline's products and services and customer feedback. By being able to identify these factors and the sentiment that is being generated on Twitter for these airlines, the airlines can make the appropriate changes or modifications that the customers are requesting or voicing via their tweets.

During recent years, the increased use of social media has managed to provide customers with a new platform where they can share their opinions, experiences, and feedback via real-time. The use of sentiment analysis and the textual data generated by the customers will provide a general framework for these airlines to focus on and overall help in the improvement of their products and services.

Methods

Technical Approach:

The following project will focus on analyzing the sentiment behind the customer's tweets directed at the major U.S. airline companies. The initial step of this project will entitle the use of Exploratory Data

Analysis (EDA). Exploratory Data Analysis will be implemented to the data as this method allows for a greater understanding of the data points that make up the dataset. Visuals will be executed with the use of EDA which allows us to look at factors such as counts and distributions within the tweets. EDA will assist in conducting the proper pre-processing actions on the textual data and natural language processing (NLP) will be used to analyze the sentiment of the tweets from the dataset. Natural language processing (NLP) is best described as a field of artificial intelligence that involves computers understanding and processing human language, in this case, the tweets by the customers (Lim, 2019). Lastly, classifiers will be implemented to the pre-processed data to attempt and predict the sentiment of the tweets.

Data:

The following data was collected by Crowdflower's Data for Everyone library and retrieved via the Kaggle website. The data was scraped from Twitter in 2015, and the contributors classified the tweets as either positive, negative, and neutral (Kaggle, 2019). A second classification was applied to the extracted data, to categorize the negative reasons that were included in the data. These negative reasons that were found within the U.S. airlines included factors such as late flights or rude service. The data was most frequently updated in late 2019 and an SQLite database was also incorporated. The objective behind the creation of this dataset was to achieve a sentiment analysis job regarding the problems that each major U.S. airline is facing (Kaggle, 2019).

Analysis

Problem Overview:

Airline companies in the United States and the increased use of social media platforms have created a space for customers to freely express their opinions and experiences. Twitter serves as the ideal platform to collect and analyze the experiences that the customers are facing when they conduct business with these airlines. The use of customer tweets is best suited for this type of problem, given that insightful information is included in these tweets and if required they can be retrieved in real-time. Being able to extract data in real-time is a great advantage for businesses as it can provide them with issues that occurring presently and not irrelevant ones. This project will use Natural Language Processing as a way to analyze the sentiment behind the textual data collected via the customer's tweets. The project will focus on what improvements need to be made by the airlines given the sentiment found in the data.

Data Understanding:

The dataset collected provides insightful information for this project as it includes information such as the airline sentiment, the reason why the sentiment was negative, user time zone, and the total number of re-tweets. The dataset is easy to understand given that no out of the ordinary variables or data types are present in it. The dataset can be considered to be large as it succeeds to contain information about a total of 14,640 published tweets. It is made up of 15 columns representing the different variables that can be measured and 14,641 rows including the header of the file. The first column in the data can be ignored or dropped given that it represents the internal identification number used by twitter to identify each individual tweet. This first column contains little to no value for this analysis while the remaining columns can help gain insightful trends and patterns.

Variables in the dataset:

tweet_id – Internal identification code for tweet
airline_sentiment- sentiment of airline (positive, negative, neutral)
airline_sentiment_confidence- confidence of sentiment (0 -1)
negativereason- extracts why tweet was negative (bad flight, customer service issue)
negativereason_confidence- confidence of negativereason sentiment (0-1)
airline- name of airline

airline_sentiment_gold - sentiment of airline

name- username of individual on Twitter

negativereason_gold- sentiment of negativereason

retweet_count- the total number of retweets

text- text making up the published tweet

tweet_coord- the coordination of where the tweet was published

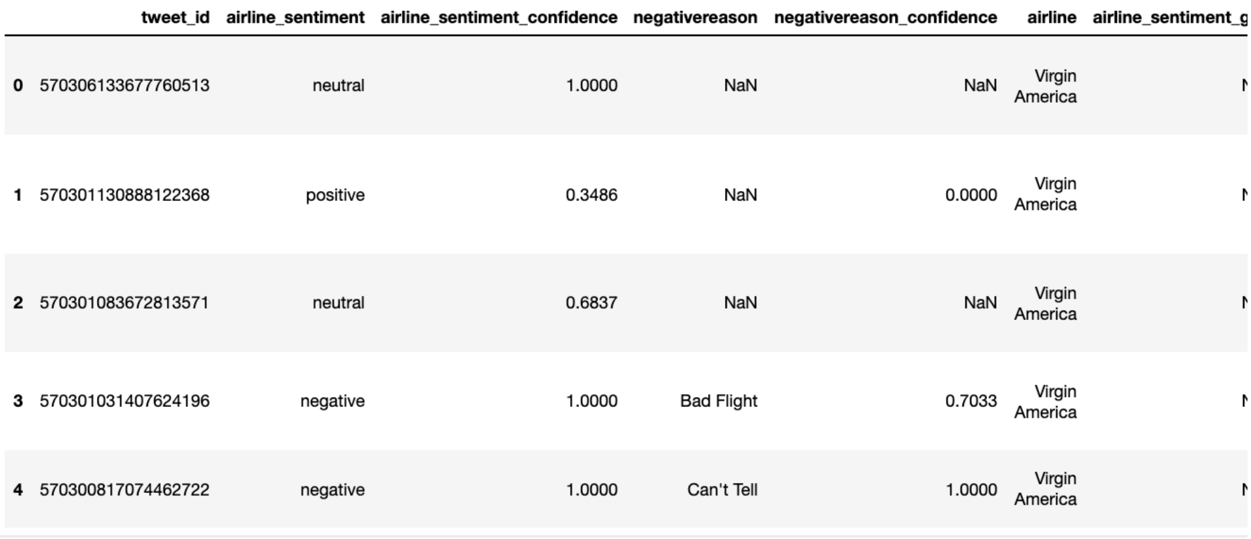
tweet_created- the data and time the tweet was published

tweet_location- the location of the tweet (e.g. San Francisco)

user_timezone- the timezone of the user when the tweet was published

Figure 1.

A snapshot of the first seven variables in the dataset.



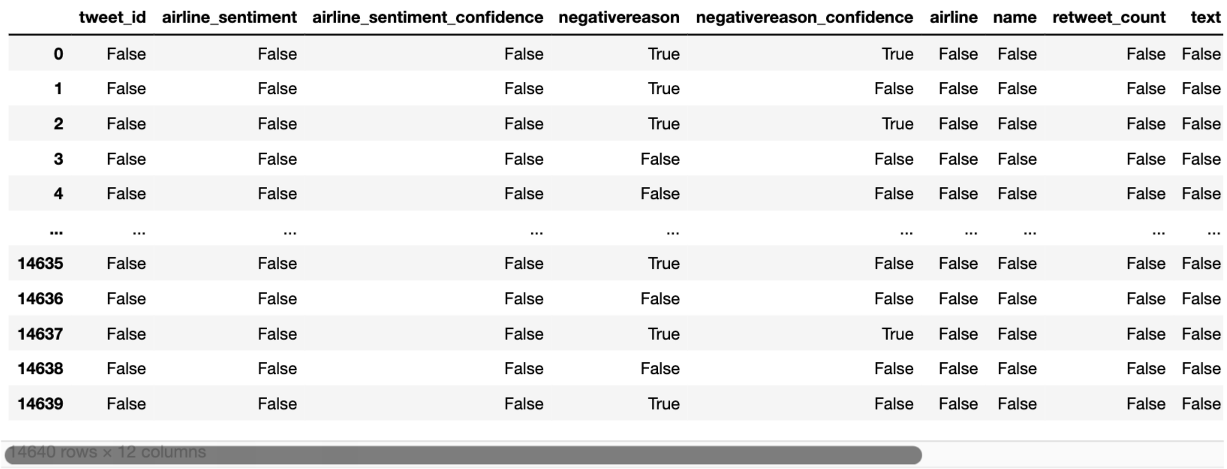
	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_g
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America	N
1	570301130888122368	positive	0.3486	NaN	0.0000	Virgin America	N
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America	N
3	570301031407624196	negative	1.0000	Bad Flight	0.7033	Virgin America	N
4	570300817074462722	negative	1.0000	Can't Tell	1.0000	Virgin America	N

Data Preparation:

During the data preparation phase of this project, several steps needed to be taken to make sure the data was in the correct format for the classification models. The data was first converted into a Pandas data frame for a more comfortable administration. The summary statistics and visualizations of the variables were analyzed during this phase to identify missing values, unique values, and outliers. During this phase, the unnecessary columns were dropped, and Exploratory Data Analysis was then conducted. After the Exploratory Data Analysis phase, the data were then pre-processed to remove common words in the tweets that have no significance to the sentiment along with the punctuation. These steps allowed for the data to be ready to be inputted into different predictive models.

Figure 2.

A screenshot of the data frame showing the missing values.



	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	name	retweet_count	text
0	False	False	False	True	True	False	False	False	False
1	False	False	False	True	False	False	False	False	False
2	False	False	False	True	True	False	False	False	False
3	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...
14635	False	False	False	True	False	False	False	False	False
14636	False	False	False	False	False	False	False	False	False
14637	False	False	False	True	True	False	False	False	False
14638	False	False	False	False	False	False	False	False	False
14639	False	False	False	True	False	False	False	False	False

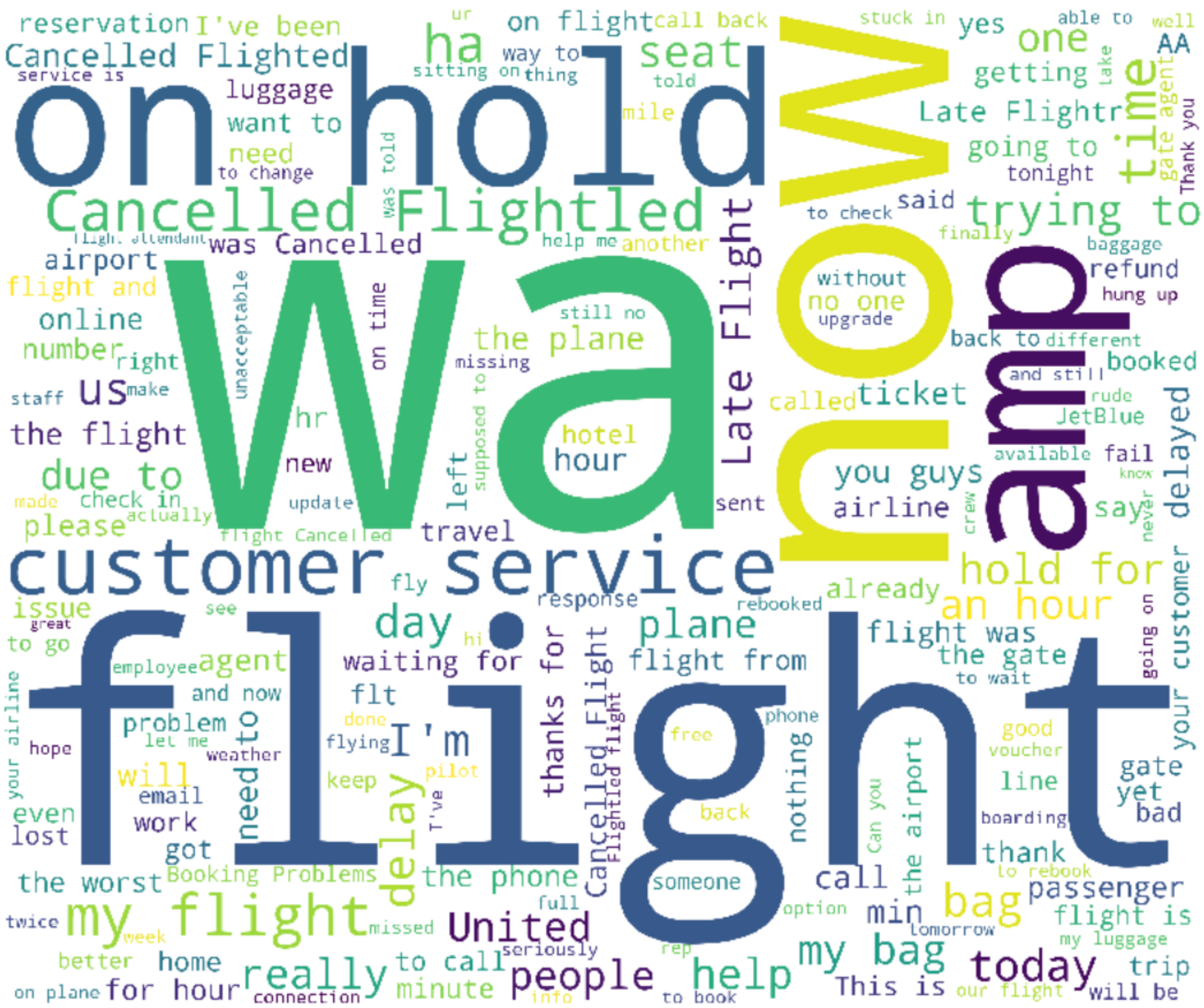
14640 rows × 12 columns

Exploratory Data Analysis

Exploratory Data Analysis was implemented on this dataset to look at the type of information contained in the tweets extracted. EDA assisted at visualizing the count of the sentiment found regarding the airlines, followed by the distributions of the reasons why the airlines received that particular sentiment. Visualizations containing the most popular words were also created to gain a better understanding of these sentiments behind these tweets. The counts of the tweets were also plotted to look at the different trends and patterns included in the dataset.

Figure 3.

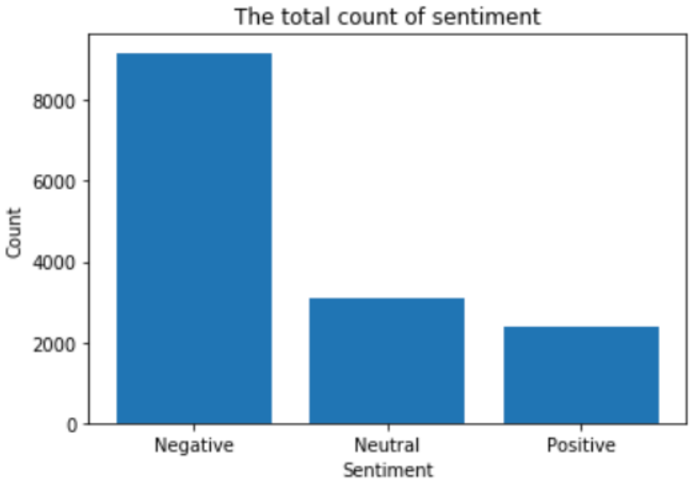
A visualization from the negative words in the tweets.



The visualization represented in Figure 3 managed to show the most common words found in the tweets directed at U.S. airlines. The most common words that appeared are on hold, customer service, canceled flight, and flight.

Figure 4.

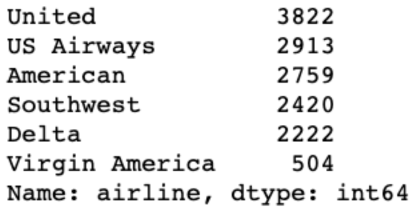
The total count of the sentiment in the tweets was analyzed.



The total count histogram shows that most of the tweets manage to represent a negative sentiment followed by a neutral sentiment, and finally a positive one.

Figure 5.

The total count per airline was established.


```
United              3822
US Airways          2913
American            2759
Southwest           2420
Delta               2222
Virgin America       504
Name: airline, dtype: int64
```

Results

During this project, the sentiment of the tweets in the dataset was initially analyzed via exploratory data analysis. Exploratory data analysis managed to provide numerous information about the trends and patterns found in the data points. Figure 6. was executed via EDA and showed that the distribution between the airlines U.S. Airways, United, and American are skewed more towards the negative sentiment portion of the graphs. Figure 7. looked at the remaining U.S. airlines of Southwest, Delta, and Virgin America. These distributions were more evenly distributed when compared with the initial three airlines from Figure 6.

Figure 6.

The distribution of sentiments of U.S. Airways, United, and American.

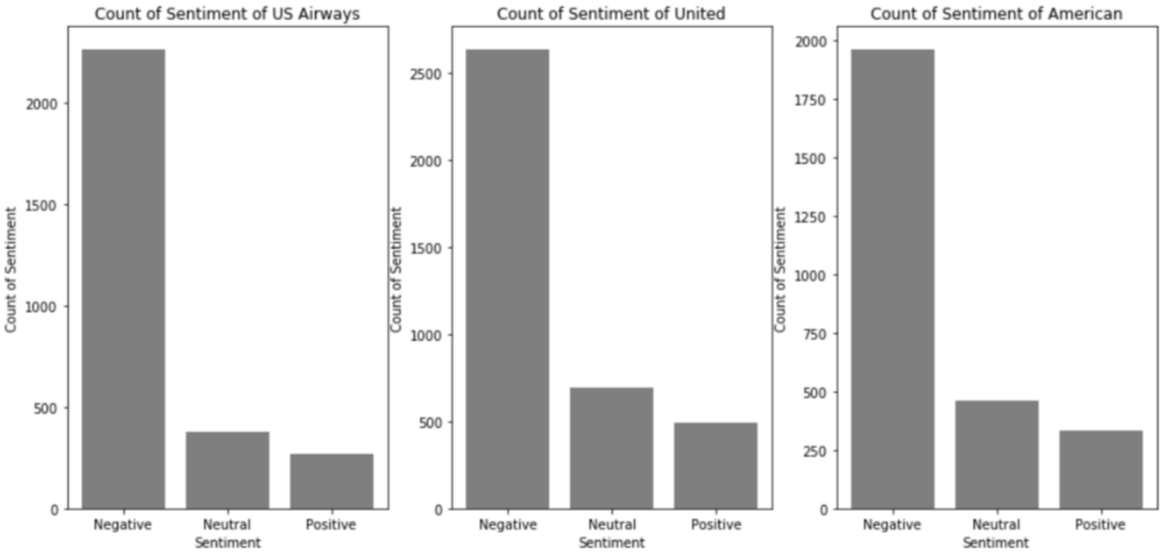


Figure 7.

The distribution of sentiments of Southwest, Delta, and Virgin America.

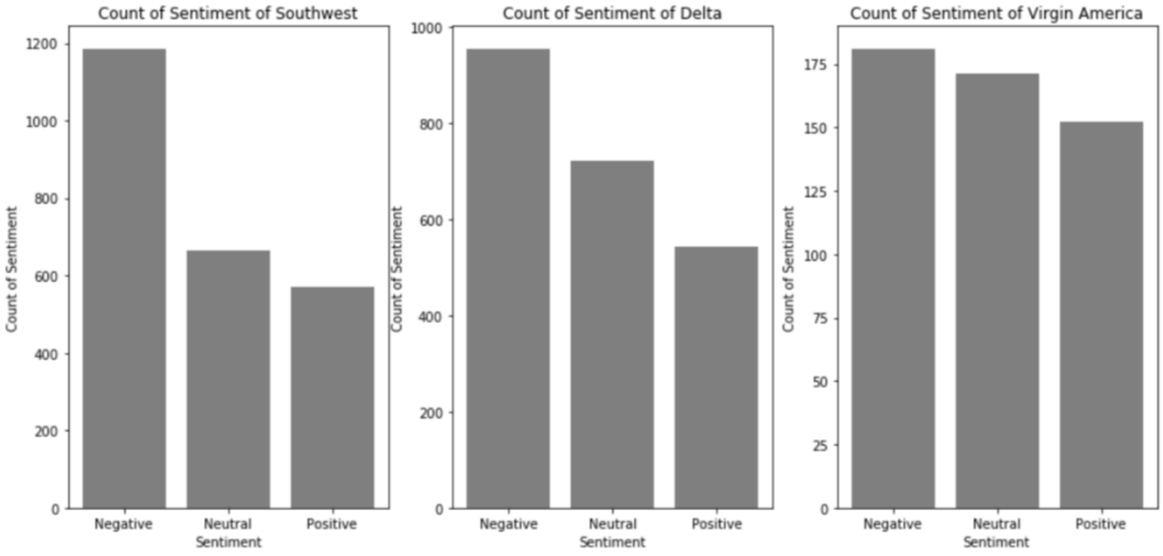
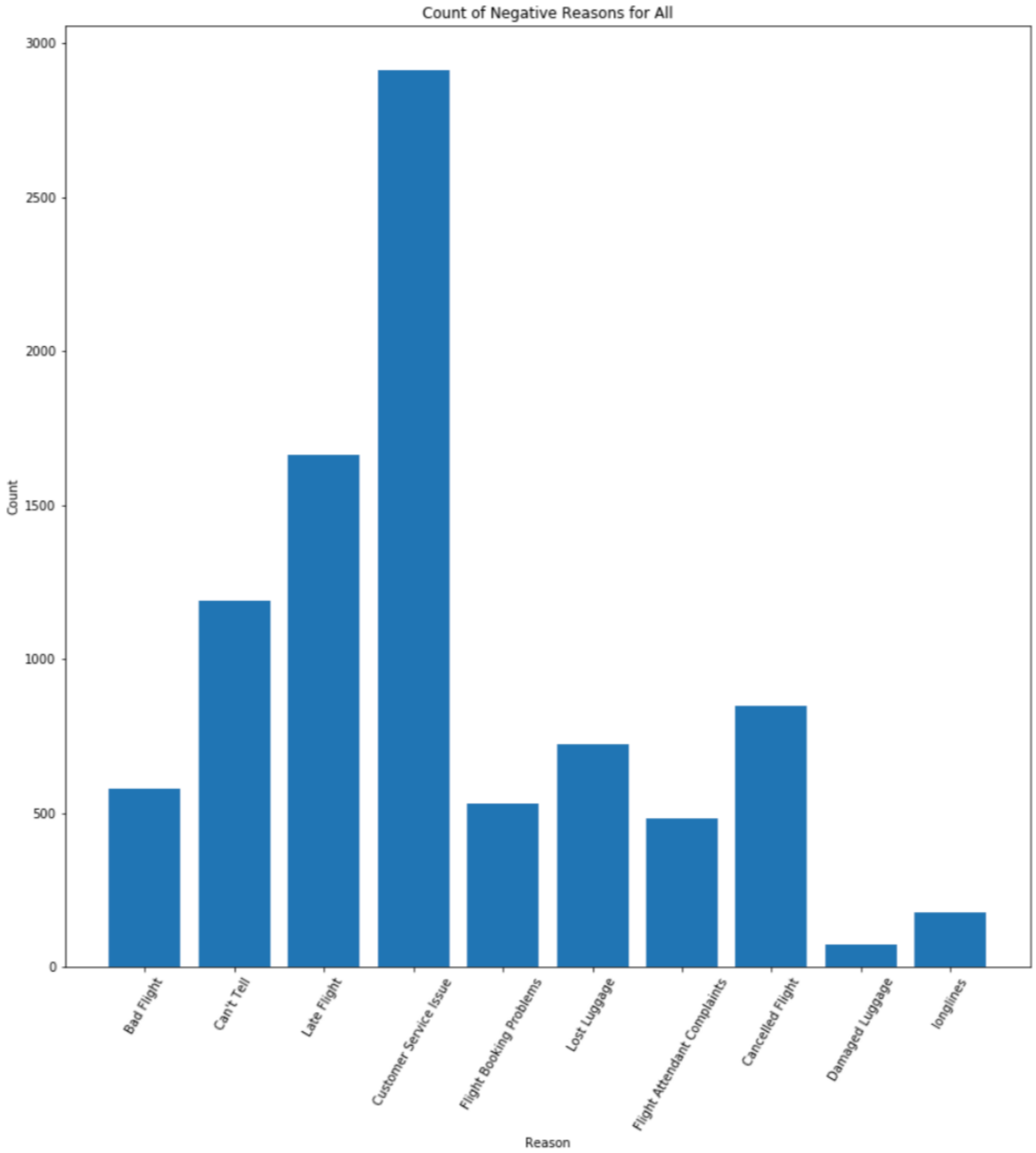


Figure 8.

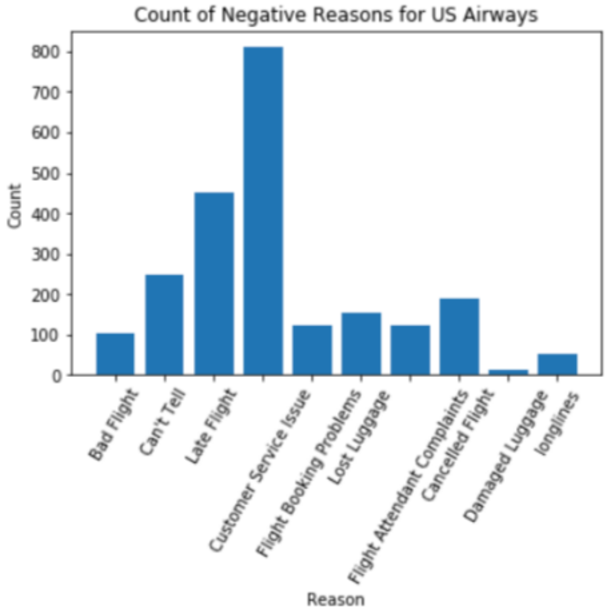
A bar plot showing the total count of the negative sentiments between all of the airlines.

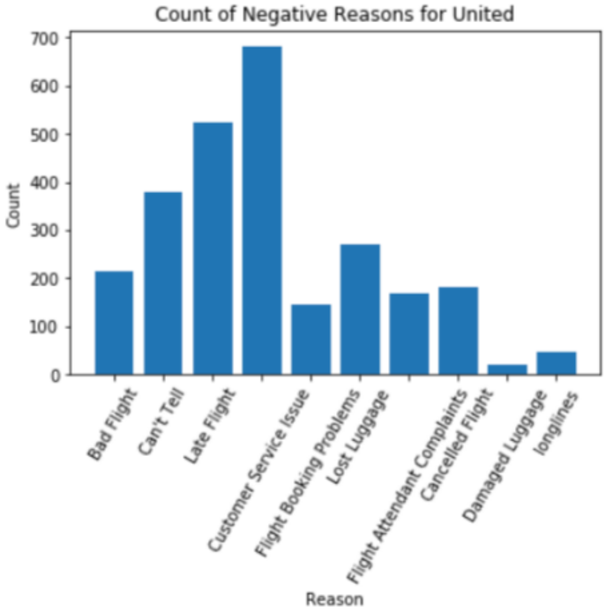


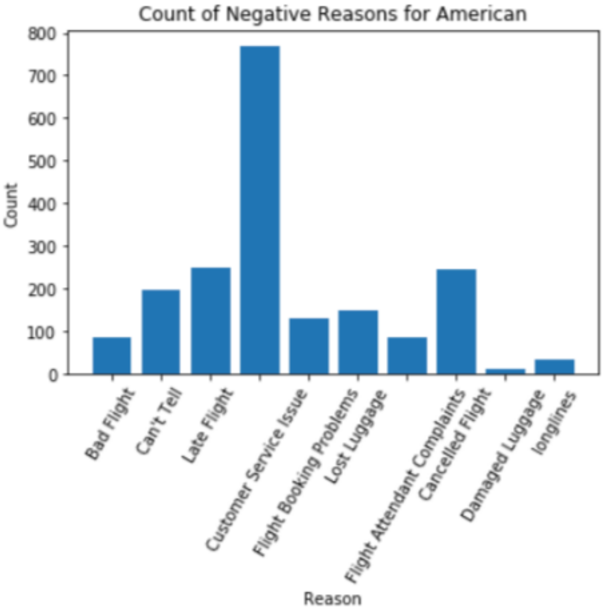
A specialized function was created with the use of programming language Python to manage to look into the details behind the total count of the negative sentiments directed at U.S. airlines. When looking at the airlines combined, we can see that the negative sentiment represented at the airlines is mostly represented by words such as customer service issues, late flight, bad flight, and even lost luggage. These types of phrases and words can help the airlines identify the types of issues that their customers are currently facing and which areas of the business need improvement.

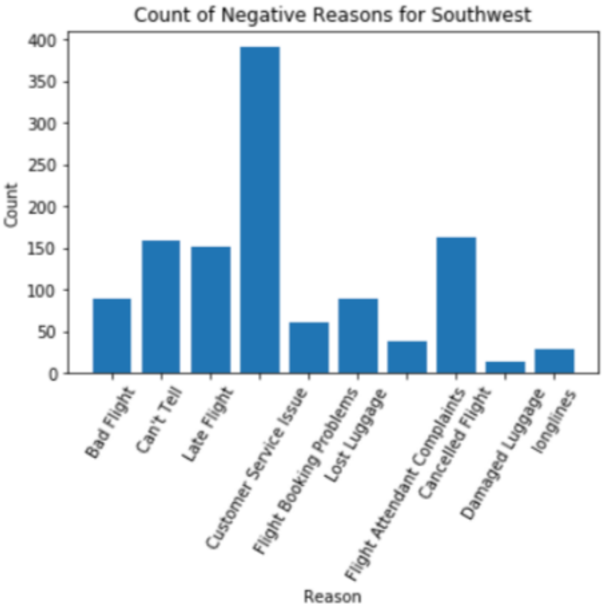
Figure 9.

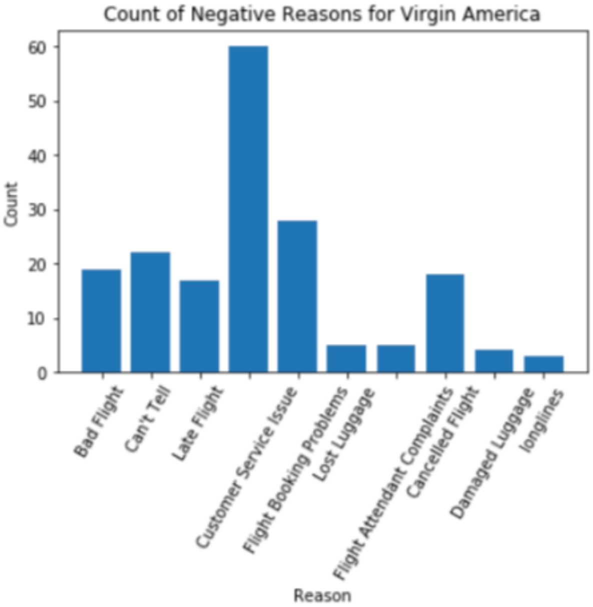
Individual histograms of the negative reasons found per U.S. airline.






Count of Negative Reasons for American
800
700
600
500
400
300
200
100
0
Count
Bad Flight
Can't Tell
Late Flight
Customer Service Issue
Flight Booking Problems
Lost Luggage
Flight Attendant Complaints
Cancelled Flight
Damaged Luggage
longlines
Reason

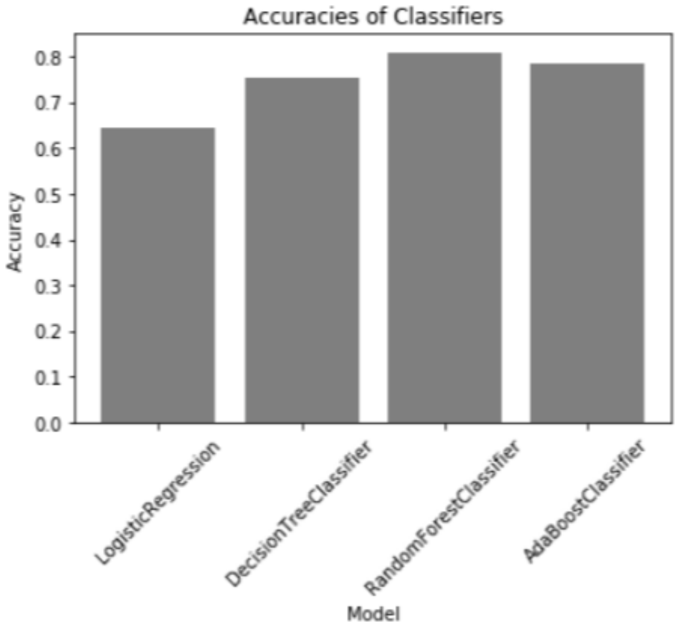




The airlines were also looked at individually to be able to identify the specific issues that each U.S. airline was having to be able to make improvements that are representative of the issues found by the company and not as a whole. The most common issue that was found between the negative sentiment tweets in the airlines was found to be the customer service issue factor. A total of four different classifiers were built and executed on the training and test dataset to predict the sentiment of the tweets. The types of classifiers built for this analysis were a logistic regression classifier, a decision tree classifier, a random forest classifier, and an AdaBoost classifier. The accuracies of the classifiers were converted into a visual for easier comparison. The accuracy of all four of the classifiers was not found to be below 65%.

Figure 10.

The accuracy of the classifiers built was converted into a visual for easier interpretation.



Conclusion

The results yielded during this project managed to provide a lot of useful information that can be utilized among the U.S. airlines. This analysis found that all of the airlines in this analysis are having the most issues and negative sentiments when looking at the customer service issues factor. The feedback that the customers are providing via this sentiment analysis serves as a clear indication that the use of social media tweets can provide an accurate and representative picture of the type of experience that the customer is receiving. This analysis and its results provide the U.S. airlines with visuals and results that can be used to make a difference and improve the overall experience that the customer is having. As found during this analysis, customers tweeted that all of the airlines had the most issues with customer service issues, canceled flights, and late flights. The use of classifiers can also be implemented into the customer's data, in this case, tweets, to evaluate the sentiment behind it. The Random Forest classifier was the most accurate, followed by the AdaBoost classifier, the Decision Tree classifier, and the Logistic Regression classifier. The random forest classifier yielded 81% accuracy, while the logistic regression classifier yielded the lowest accuracy of the models of 65%.

Acknowledgments

I would like to thank all of the individuals that take the time to express their voice via social media platforms such as Twitter. These individuals provide data that can be utilized to increase business needs and provide the business with representative data. I would also like to thank the social media platform, Twitter, given that they managed to provide the public with an accessible API where data can be retrieved for further exploration.

References

Crowdflower. (n.d.). About. Retrieved from https://www.welcome.ai/crowdflower

Kaggle. (2019). Twitter US airline sentiment. Retrieved from https://www.kaggle.com/crowdflower/twitter-airline-sentiment

Lim. F. (2019). Twitter U.S. airline sentiment analysis using keras and rnn's. Retrieved from https://medium.com/@francesca_lim/twitter-u-s-airline-sentiment-analysis-using-keras-and-rnns-1956f42294ef

Wikipedia Wikipedia. (2020). Sentiment Analysis. Retrieved from https://en.wikipedia.org/wiki/Sentiment_analysis