Mira Wadehra

Professor Houldsworth

MGSC-410-01

October 6, 2022

**Insight Report**

**Summary:**

Variables I Kept:

1. **Airline:** Name of the Airline
2. **Airline\_sentiment:** Whether the Airline Sentiment is Positive, Negative or Neutral
3. **Airline\_sentiment\_confidence:** How confident the sentiment analysis is that the airline sentiment is correct
4. **Negativereason:** Negative reason why the tweet was posted (Ie Late Flight)
5. **Negativereason\_confidence:** How confident the sentiment analysis is that the negative reason is correct
6. **Retweet\_count:** Number of times the Tweet was retweeted
7. **Text:** The actual Tweet text

Variables I Removed:

1. **Tweet\_id:** The Twitter user’s ID
2. **Airline\_sentiment\_gold:** No Data Provided
3. **Negativereason\_gold:** No Data Provided
4. **Tweet\_coord:** Coordinates from where the Tweet was sent
5. **User\_timezone:** Timezone from where the Tweet was sent
6. **Tweet\_location:** The location from where the Tweet was sent
7. **Name:** Name of Twitter User

**What is the mean airline sentiment for each airline? Which airline has the highest negative, neutral and positive sentiments?**

| **Airline** | **Airline Sentiment** | **Mean Airline Sentiment Confidence per Airline** |
| --- | --- | --- |
| US Airways | Positive | .0.8596859 |
|  | Negative | 0.9457137 |
|  | Neutral | 0.8219223 |
| American | Positive | 0.8823024 |
|  | Negative | 0.9449547 |
|  | Neutral | 0.8259378 |
| United | Positive | 0.8560118 |
|  | Negative | 0.9333827 |
|  | Neutral | 0.8097561 |
| Southwest | Positive | 0.8861053 |
|  | Negative | 0.9205330 |
|  | Neutral | 0.8261092 |
| Virgin America | Positive | 0.8879783 |
|  | Negative | 0.9017331 |
|  | Neutral | 0.8383684 |
| Delta | Positive | 0.8671110 |
|  | Negative | 0.9022022 |
|  | Neutral | 0.8292642 |

The airlinesentiment\_confidence is how confident the sentiment analysis is that the airline sentiment that they stated (positive, negative or neutral) is true or certain. I wanted to know what the average sentiment was for each airline and which airlines had the highest negative, neutral and positive sentiments. Based on the averages of the airline sentiment confidence per airline, the airline with the most negative sentiment is US airways (mean airline sentiment confidence of .946). The airline with the most neutral sentiment is United (mean airline sentiment confidence of .810). The airline with the most positive sentiment is Virgin America (mean airline sentiment confidence of .946).

**What is the mean and median negativereason\_confidence for each airline? Which airline had the highest negativereason\_confidence?**

The negativereason\_confidence is how confident the sentiment analysis is that the negative reason they said (whether it is Late Flight or Flight Cancelled) is true or certain. I removed the blanks or N/A’s from the negativereason\_confidence when calculating the mean and median for the negativereason\_confidence. Before I removed those specific rows, there were 14640 observations and after removing them, there were 10522 observations. So the mean and median were found with 4118 less observations, so it is not a fully accurate mean and median.

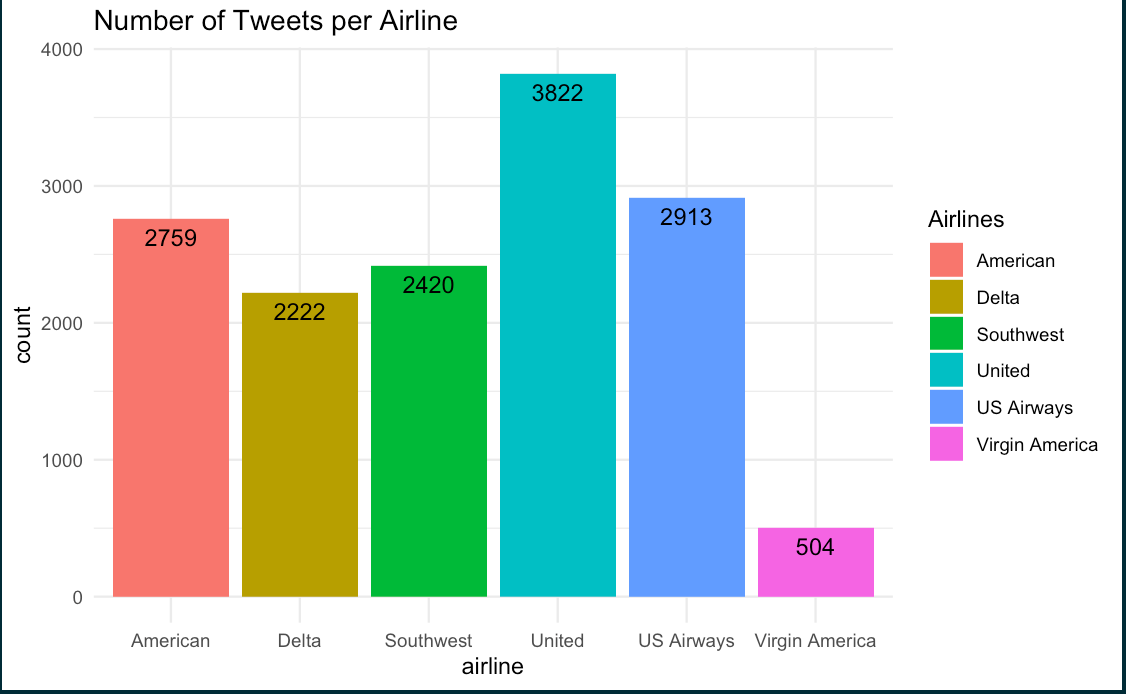
| **Airline** | **Mean Negative Reason Confidence per Airline** |
| --- | --- |
| US Airways | .687171717 |
| American | .6664389 |
| United | .6272854 |
| Southwest | .6015080 |
| Virgin America | .5730965 |
| Delta | .5569854 |

| **Airline** | **Median Negative Sentiment Confidence per Airline** |
| --- | --- |
| US Airways | .67770 |
| American | .67515 |
| United | .66805 |
| Southwest | .66520 |
| Delta | .66300 |
| Virgin America | .66170 |

Based on the averages of the negativereason\_confidence per airline, the worst airline based on their negative reason is US airways (mean of .69) and the best airline based on their negative reason is Delta (mean of .56). Using the same variable, negativereason\_confidence, I found the median. All of the airline's medians were close to each other. Based on the median of the negativereason\_confidence per airline, the worst airline (out of the six in this dataset) is US airways (median negativereason\_confidence of .678) and the best airline (out of the six in this dataset) is Virgin America (median negativereason\_confidence of .661).

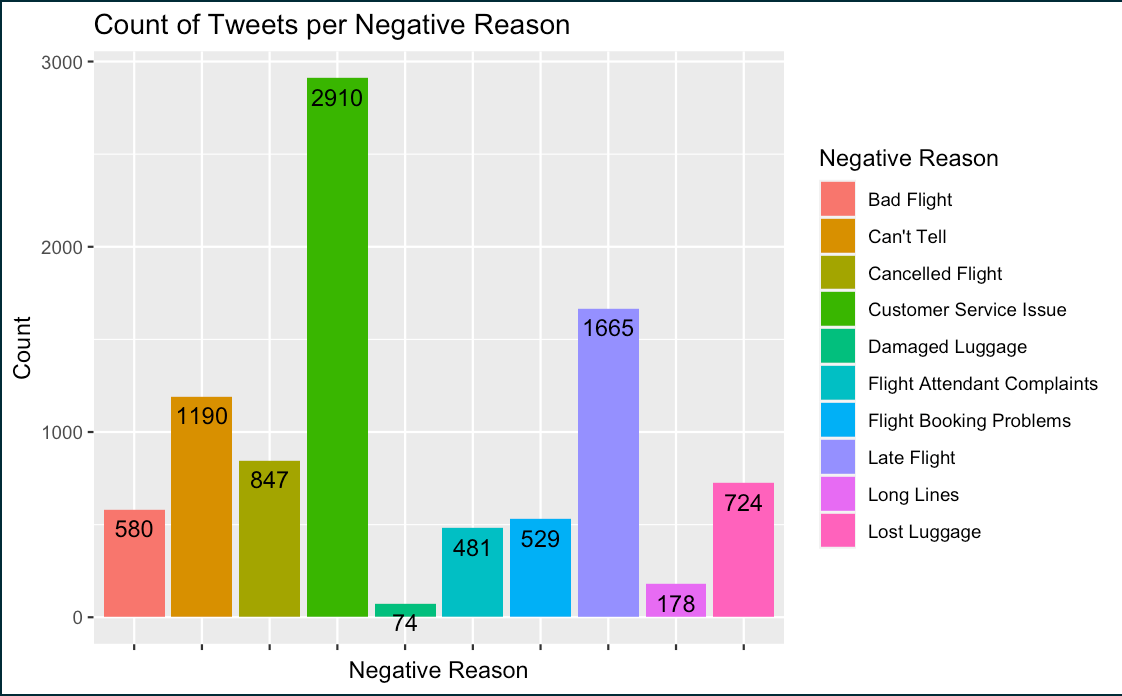
The order of the median per airline in descending order was similar to the order of means. However, Virgin America had a lower median tha Delta by .002. Delta had lower mean than Virgin America by .017. This shows that there was one Virgin Atlantic negative\_reason that was an outlier, that significantly impacted the mean.

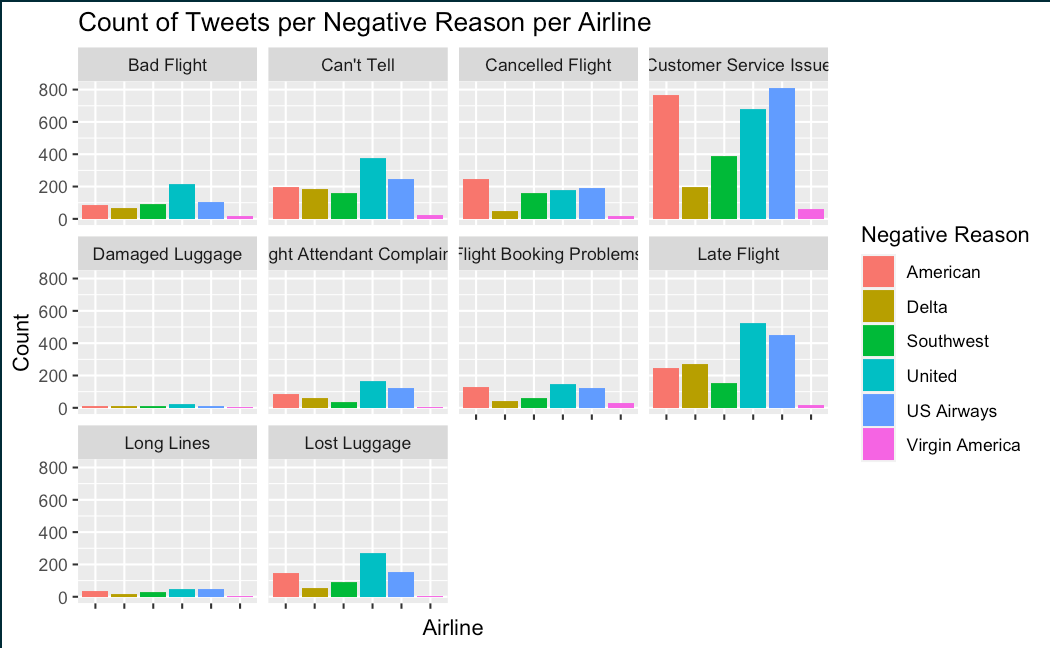
**Which airlines do consumers tweet most about?**



There were a total of 14,640 tweets (observations). In the graph above, the number of tweets is broken down by each airline. The airline with the most tweets was United (3822 tweets). The airline with the least amount of tweets was Virgin America (504 tweets). This could be for multiple reasons. One, Virgin America flies to 27 destinations in the US and 3 internationally. So they have relatively small flight times and a minimum number of routes when compared to other airlines, such as United or American airlines, which fly internationally. Southwest even has some international flight routes going to countries, including Aruba, Belize and Costa Rica. Since Virgin America does not have many flight routes, less consumers use and rate the airline. Another reason is the number of passengers per year. In 2015, Virgin America had 7 million passengers, while American Airlines had 28.256 million passengers. Since Virgin America has significantly less passengers than American Airlines, they have less tweets written about them.

**What are the main reason’s negative tweets? Which airline has the worst rating for each negative reason?**



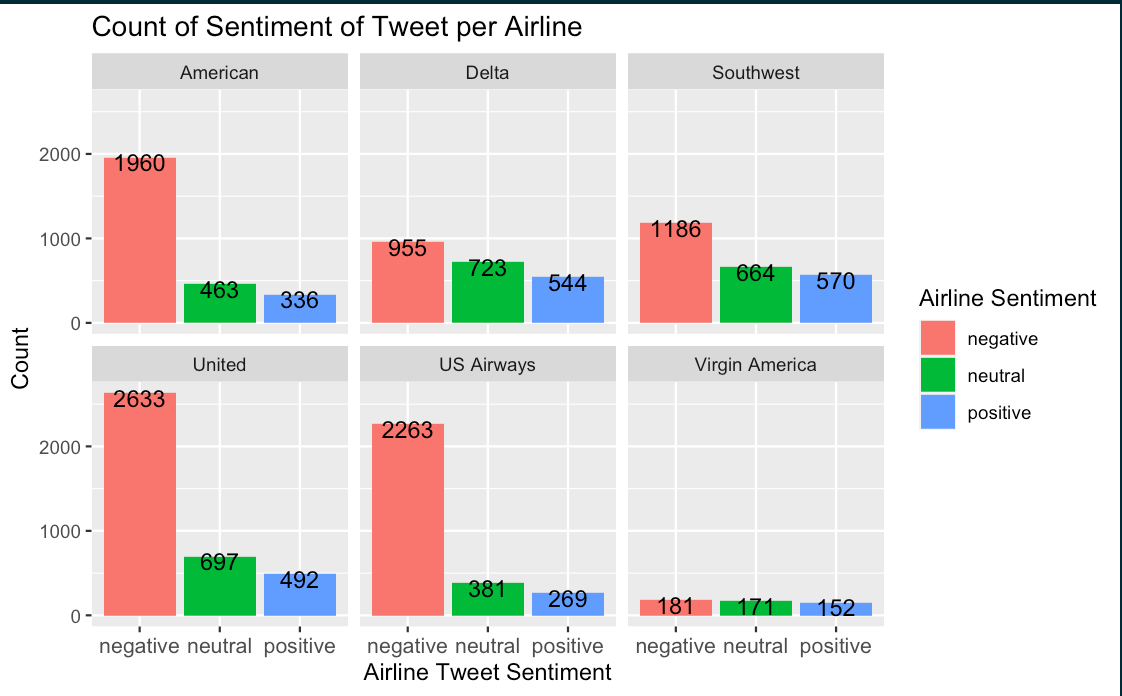


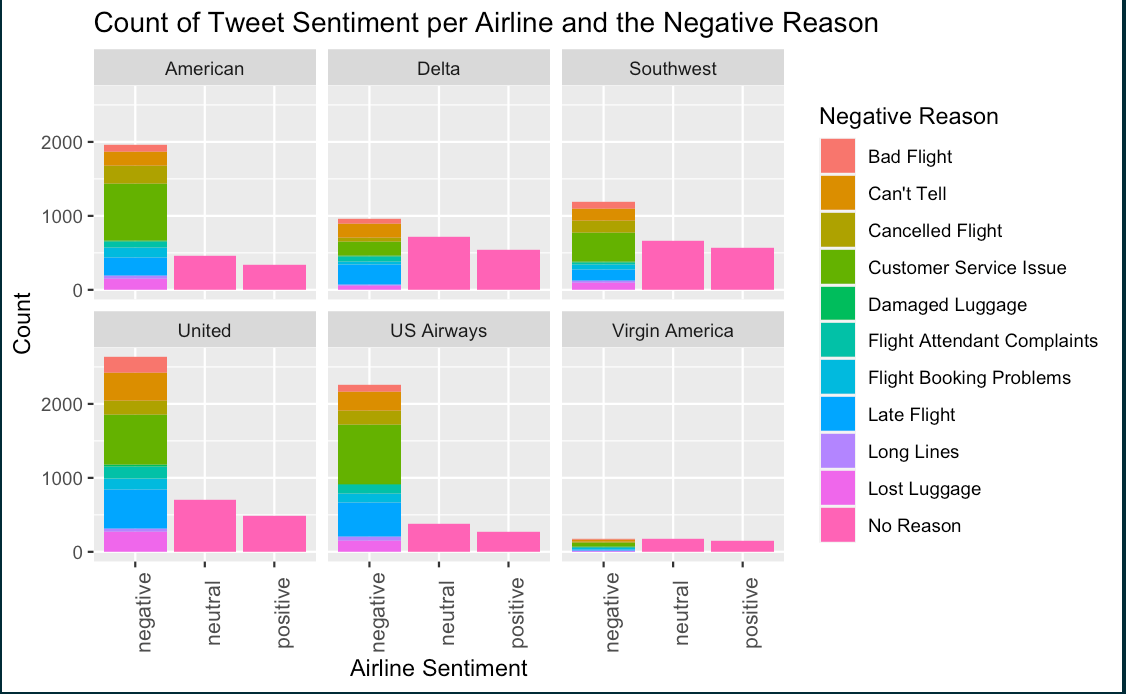
| **Negative Reason** | **Airline** | **Number of Tweets** | **% of total tweets of the specific negative reason** |
| --- | --- | --- | --- |
| Bad Flight | United | 216 | 37.2% |
| Can’t Tell | United | 379 | 31.8% |
| Cancelled Flight | American | 246 | 29% |
| Customer Service Issues | US Airways | 811 | 27.9% |
| Damaged Luggage | United | 22 | 29.7% |
| Flight Attendant Complaints | United | 168 | 34.9% |
| Flight Booking Problems | United | 144 | 27.2% |
| Late Flight | United | 525 | 31.5% |
| Long Lines | US Airways | 50 | 28.1% |
| Lost Luggage | United | 269 | 37.2% |

I wanted to find out the main reason why each airline had a tweet. I only used the airline sentiments that were negative, since I wanted to know why each airline had a negative tweet. That shrunk my observations to 9,178. The most listed negative reason was “Customer Service Issues”, 31.7% of the negative tweets. The least listed negative reason was “Damaged Luggage”, .81% of the negative tweets.

This trend was replicated when I broke it down per airline. The most frequent negative reason was “Customer Service Issues” and the least frequent negative reason was “Damaged Luggage”. The airline with the most “Customer Service Issues” was US Airways (27.9% of the “Customer Service Issues” tweets) and the least was Virgin America (2.10% of the “Customer Service Issues” tweets). The airline with the most “Damaged Luggage” was United (29.7% of the “Damaged Luggage” tweets) and the least was Virgin America (5.40% of the “Damaged Luggage” tweets).

**What is the overall positive, negative and neutral sentiment per airline? What are the top three negative reasons why for each airline?**



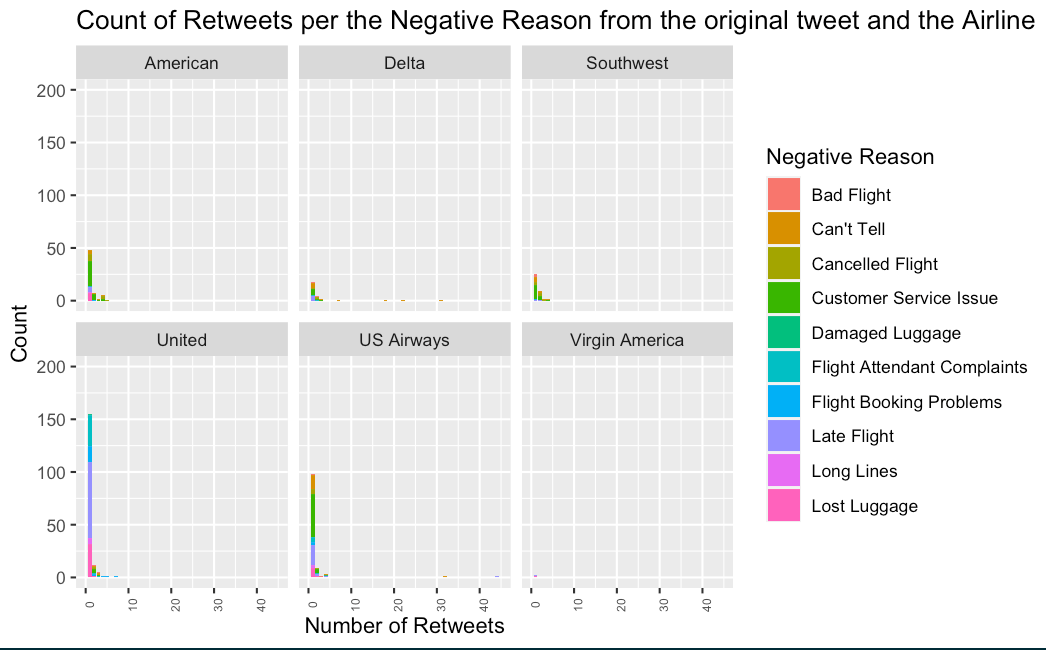


| **Airline** | **Top Three Negative Reasons** | **Number of Tweets** | **What is the % of the negative reason across each airline's negative tweets?** |
| --- | --- | --- | --- |
| American | Customer Service Issues | 786 | 40.1% |
|  | Late Flight | 249 | 12.7% |
|  | Cancelled Flight | 246 | 12.6% |
| Delta | Late Flight | 269 | 28.2% |
|  | Customer Service Issues | 199 | 20.8% |
|  | Can’t Tell | 186 | 19.5% |
| Southwest | Customer Service Issues | 391 | 33% |
|  | Cancelled Flight | 162 | 13.7% |
|  | Can’t Tell | 159 | 13.4% |
| United | Customer Service Issues | 681 | 25.9% |
|  | Late Flight | 525 | 20% |
|  | Can’t Tell | 379 | 14.4% |
| US Airways | Customer Service Issues | 811 | 35.8% |
|  | Late Flight | 453 | 20% |
|  | Can’t Tell | 246 | 10.8% |
| Virgin America | Customer Service Issues | 60 | 33.1% |
|  | Flight Booking Problems | 28 | 15.5% |
|  | Can’t Tell | 22 | 12.2% |

I wanted to know which airline had the highest and lowest positive, negative and neutral sentiment to understand which airline is “the worst” and the “the best” based on these tweets. United had the highest number of negative sentiment tweets (68.9% of all of United’s tweets) and Virgin America had the lowest number of negative sentiment tweets (35.9% of all of Virgin America's tweets). Delta had the highest number of neutral sentiment tweets (32.5% of all of Delta’s tweets) and Virgin America had the lowest number of neutral sentiment tweets (33.9% of all of Virgin America’s tweets). Southwest had the highest number of positive sentiment tweets (19.1% of all of Southwest’s tweets) and Virgin America had the lowest number of positive sentiment tweets (30.2% of all of Virgin America’s tweets). Again, Virgin Atalantic had the lowest number of tweets for each sentiment because Virgin America is represented the least in this dataset.

I wanted to know what the airline sentiment was per airline based on the negative reason. All of the tweets that had a positive or neutral sentiment did not have a reason (“No Reason”) and all the tweets that had a negative sentiment had a negative reason. Following the same trend, the most common negative reason was “Customer Service Issues” and the least common negative reason was “Damaged Luggage”. The only time “Customer Service Issues” was not the top negative reason for was Delta; their top negative reason was “Late Flight”.

**Was there a specific negative reason why the airline had a negative sentiment based on retweets?**

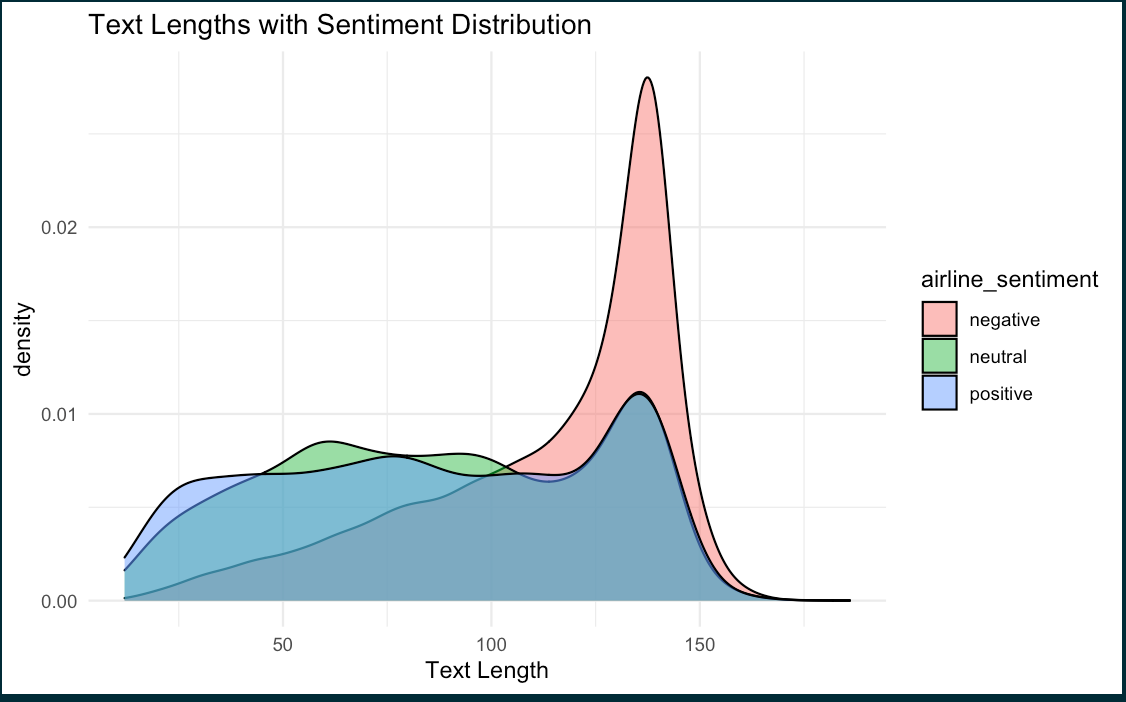


| **Top Negative Reason** | **Number of Retweets** | **% of Total Negative Retweets** |
| --- | --- | --- |
| Bad Flight | 45 | 5.3% |
| Can’t Tell | 219 | 25.6% |
| Cancelled Flight | 63 | 7.4% |
| Customer Service Issues | 221 | 25.8% |
| Damaged Luggage | 5 | .58% |
| Flight Attendant Complaints | 44 | 5.1% |
| Flight Booking Problems | 42 | 4.9% |
| Late Flight | 156 | 18.2% |
| Long Lines | 9 | 1.1% |
| Lost Luggage | 51 | 6% |

Approximately 95% of all the tweets were not retweeted. However United and US Airways had some tweets that were retweeted over 30 times. The most number of retweets was 44 retweets from one tweet about US Airways. The airline with the least amount of tweets was Virgin America.

The two most common negative reasons for retweets were “Customer Service Issues” (25.6% of all retweets) and “Can’t Tell” (25.6% of all retweets )The least common negative reason for the amount of retweets was “Damaged Luggage” ( .58% of all the retweets). All of these graphs are showing the same trend: The most common negative reason is “Customer Service Issues” and the least common negative is “Damaged Luggage”. The one outlier is United. Their top negative reason when looking at retweets, was “Late Flight”. Customers seemed to be frustrated with United having late flights, however, they chose to retweet instead of tweeting about it themselves.

**Does the length of the text show anything about the Sentiment?**



Tweet Example:

**Negative Airline Sentiment Tweet:** “@united did our bags leave Washington yet? That would be convenient seeing as we left there a day ago. #UnitedAirlines #lostluggage”

**Positive Airline Sentiment Tweet:** “@united I will, as soon as I am done with my article. Thanks! ~Tom”

From the graph above, we can see that most of the longer tweets have a negative sentiment, while most of the shorter tweets are neutral or positive. The negative airline sentiment tweet (from above) has 21 words, while the positive airline sentiment tweet has 14 words.

**References/Resources**

<https://stackoverflow.com/questions/22286419/move-a-column-to-first-position-in-a-data-frame>

<https://www.marsja.se/how-to-remove-a-column-in-r-using-dplyr-by-name-and-index/>

<https://www.statology.org/r-mean-by-group/>

<https://stackoverflow.com/questions/24172111/change-the-blank-cells-to-na>

<https://stackoverflow.com/questions/26553526/how-to-add-frequency-count-labels-to-the-bars-in-a-bar-graph-using-ggplot2>

<https://stackoverflow.com/questions/31939643/make-a-table-showing-the-10-largest-values-of-a-variable-in-r>

<https://www.geeksforgeeks.org/change-color-of-bars-in-barchart-using-ggplot2-in-r/>

<https://stackoverflow.com/questions/29278153/plotting-with-ggplot2-error-discrete-value-supplied-to-continuous-scale-on-c>

<https://www.codingprof.com/3-ways-to-replace-blanks-with-nas-in-r-examples>

<https://sparkbyexamples.com/r-programming/replace-character-in-a-string-of-r-dataframe/>

<https://r-lang.com/r-max-function-how-to-find-maximum-value-of-vector/>

<https://www.statmethods.net/graphs/bar.html>

<https://medium.com/@rohitnair_94843/analysis-of-twitter-data-using-r-part-2-word-cloud-dd423af1b2c6>

<https://towardsdatascience.com/twitter-text-analysis-in-r-ed7b81ecdb9a>

[chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf](https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf)

**data\_visualization\_ggplot\_cheatsheet.pdf**

<https://medium.com/analytics-vidhya/tdm-term-document-matrix-and-dtm-document-term-matrix-8b07c58957e2>

<https://www.programmingr.com/count-occurrences-in-column/>

<https://www.statology.org/r-remove-rows-with-na-in-one-column/>

<https://en.wikipedia.org/wiki/List_of_Virgin_America_destinations>

<https://en.wikipedia.org/wiki/List_of_Southwest_Airlines_destinations>

<https://en.wikipedia.org/wiki/List_of_US_Airways_destinations>

<https://en.wikipedia.org/wiki/Virgin_America>

<https://www.youtube.com/watch?v=-O7W___fnrE&t=88s>