Insight Report

Assignment 1 – Twitter Airline Sentiment Analysis

Data:

Let’s start with saying that the data for this assignment was not as bad as it was made out to be. Of course, with poor data preparation, it would be easy to remove much of the data, however with good data preparation the data was very easy to use and create insights with. The figure below depicts how our data is structured when it is first taken in:

Text

Description automatically generated

It’s clear that there are several categories and with a few quick lines of code to check our null values we can easily remove “tweet\_cord”, “airline\_sentiment\_gold”, and “negativereason\_gold”. The percentage of null or N/A values for the aforementioned groups were 93.04%, 99.73%, and 99.78%. We of course removed these categories as they would offer us very little insight and maybe have in fact caused some issues with our later analysis.This leaves us with a dataset looking something like this:

Graphical user interface, text

Description automatically generated

From here we were able to go ahead and create our analysis based on visualizations we worked on building.

Before jumping into the aforementioned section, let’s take a look at what our data holds in terms of its values. United Airlines had the most tweets at 3,822 and Virgin American at 504 tweets respectively. A quick word cloud helped to show us that the main reason for the negative airline sentiment was “Customer Service Issues” prior to creating any more intense graphs. In addition, a quick frequency check helps to show us that 2,910 tweets referenced “Customer Service Issues” as the leading negative reason with “Late Flight” coming in second at 1,665 tweets.

Analysis with Visualizations:

Moving into our main exploratory data analysis and our findings, things started off with a quick view at the general breakdown of sentiment between all airlines:

Chart, bar chart

Description automatically generated

Clearly, we can see that there was without a doubt a very high negative sentiment amongst customers. The next visualization created was a look at how these sentiment trends continued when viewed at each airline alone. In doing so the following graph was created:

Chart, bar chart

Description automatically generated

This visualization once again shows us that negative sentiment was the highest sentiment for all airlines but does also illustrate that Virgin America had the lowest range between all three sentiments. This may be a result of the Virgin America having the fewest number of tweets. Another visualization which achieves the same effect as above while being nicer to view is a stacked bar graph:

Chart, bar chart

Description automatically generated

This graph was selected in the presentation as it’s visually easier to understand while also being easier to view. The bold- almost harsh- colors of the first figure make it difficult to interact with.

The next area for which analysis occurred was with the timing of the data. The data was timed very oddly with tweets being collected from 02/16/2015 through 02/24/2015. Having such a period of time makes it very difficult to create any strong assumptions about the future. This will be further discussed in the last section but for now, the following visualizations were created to help understand what is going on with our data:

Chart, line chart

Description automatically generated

A screenshot of a computer

Description automatically generated with low confidence

These 2 figures alone tell a very power story in our understanding of the data. The first figure helps to show the general trend of when tweets were made while the second figure shows us the specific breakdown of negative tweets. Looking at both figures we see that the 22nd and 23rd had the most overall tweets and negative tweets. However, looking at the second figure we see an interesting trend. American airlines did not have *any* tweets prior to the last 3 days. This can highly skew our data. In other words, had the high number of negative sentiment American airline tweets been present in all of the data, the understanding could have been far worse. To help further reinforce this, the following visualization helps illustrate the proportion of negative sentiment tweets from each airline:

A picture containing shape

Description automatically generated

It can be seen that American airlines comes in a close second with US Airways for highest proportion of negative sentiment tweets.

Take Aways and Future Works:

With the above analysis discussed, there are some key take aways to be had. The first is that all airlines had strong negative sentiment from Twitter users over the brief period of time in the data. The second take away comes from looking at the most cited reason for such negativity- “Customer Service Issues’. This was the case for all airlines except Delta which cited “Late Flights”. Airlines should make an effort to ensure customers are as happy as possible even it entails providing compensation for “poor” flights. Doing so will create repeating customers who become profitable in the long run. Of course, this needs to analyzed prior to being applied. The last large take away from the data is that airlines may want to collect their own user data from flyers. Doing so will in theory provide a more well rounded understanding of how customers feel about flying with a given airline. This is important to note as it has been shown that individuals tend to voice complaints more often and more loudly as opposed to satisfaction.

With regards to future work, having a data set which spams a much larger time period would help to create a more robust and well rounded understanding of how customers felt flying. In addition, it would be a good idea to ensure that more numeric data is collected. The highly categorical data used makes predictive models beneficial to airlines difficult to create. One possible model is an NLP based predictive text model if an individual is looking to predict what a customer of a specific sentiment will say based on what has already been said.

Note that it is highly encouraged to refer to the GitHub repository for this project found [here](https://github.com/Armans3838/TwitterFlightSentiment). At the repository, all of the code used can be found in addition to the documents request for this project.

Thank you.