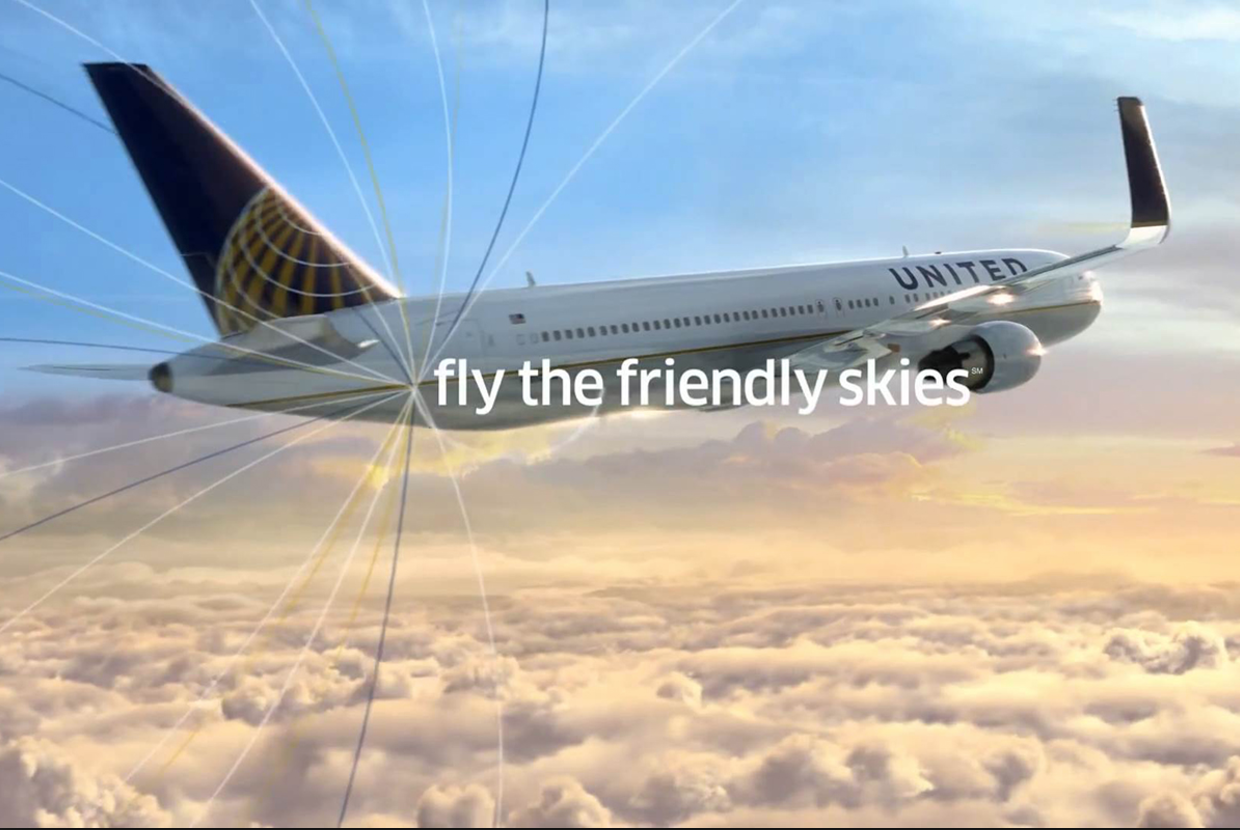
Which Airline Flies the Happiest Customers?

By: Andrew Goss, Carter Forbes and Shannon Lamberson







Business Questions:

Which airlines have the happiest/unhappiest customers? Airline travel is becoming an increasing form of travel for even the most economical family. But, air travel, as most of us have experienced, is not the same on every airline. Some airlines tote themselves as being the most friendly and flexible others as the most reliable, but saying something is much easier than delivering on that promise.

Implications of analysis:

Our analysis plans to get airline sentiment straight from the customers’ mouth (or Twitter account) and find out which airline is delivering the best (or worst) customer service. The results of the analysis will show which airline has the happiest customers and what to expect when selecting an airline to travel with.

Data collected

We approached collecting data in two ways. First, we downloaded a data set from Kaggle called Twitter U.S. Airline Sentiment and began initial analysis on airline sentiment. However, this data was collect in February of 2015, so we also collected our own data and created a program to predict airline sentiment. We'll use this data to see if the sentiment has changed over the past year. We also found airline market share from Statista.com. We'll use this to give the data perspective. Obviously, the larger the airline is and the more flights they have the higher the tweet volume should be.

Collecting our own data wasn't difficult. We used Twitter's API to stream tweets that contained a U.S. airline. We collected 5,200 tweets. We selected tweets that contained one of these key airlines: @AlaskaAir, @AmericanAir, @DeltaAssist, @FlyFrontier, @Jetblue, @SouthwestAir, @United, @VirginAmerica

The market share data from Staista.com also needed to be adjusted for analysis to match with the Kaggle data. The six airlines from Kaggle only represented 70.9% of the market share. We used a 1.41 multiplier to make these six airlines represent 100% market share.

Southwest 18.2%

Delta 17%

United 14.7

American 16.6%

US Airways 3.3%

Virgin 1.1%

Here is an overview of the Kaggle data in Stata.

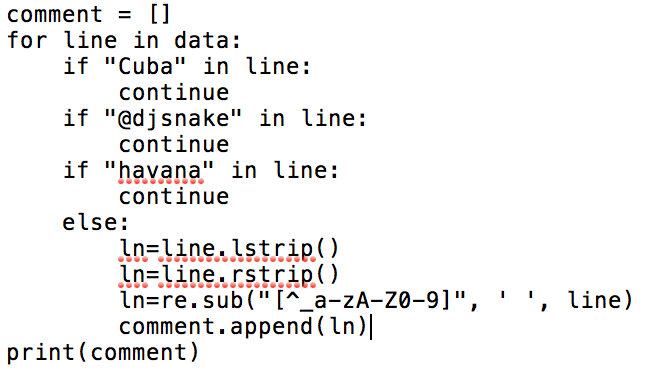


Evaluation of Models:

Our evaluation has two steps. The data we collected first needed to be cleaned, and then we needed to predict sentiment of each tweet.

Cleaning:

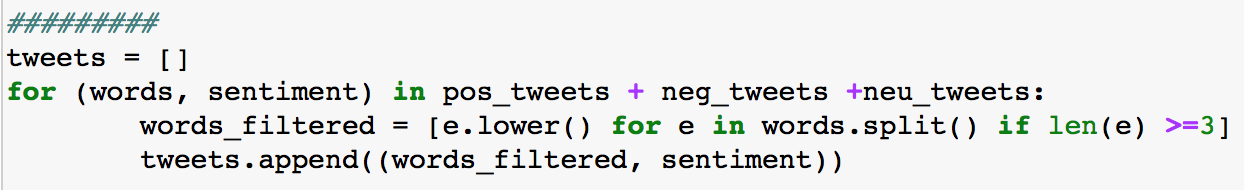
* Removing mention of the Cuba flights and mention of a new album called DJsnake. We also stripped the Tweets of extra characters and spaces.



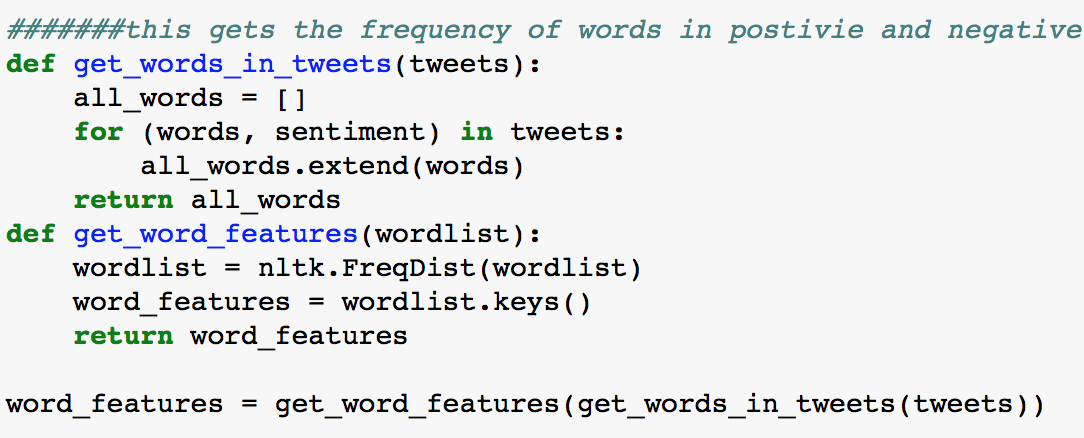
After this we created a new file to run our predictive analysis. To do this, we used Python and the natural language package NLTK. To implement the model, we first had to train it. To do this, we listed out examples of tweets and marked them as positive, negative and neutral. For example:

* pos\_tweets = [('I love this car', 'positive'), ('This view is amazing', 'positive'),]
* neg\_tweets = [('This view is horrible', 'negative'), ('I feel tired this morning', 'negative'),]
* neu\_tweets = [ ('just landed', 'neutral'), ('Happy Holidays from AmericanAir', 'neutral')]

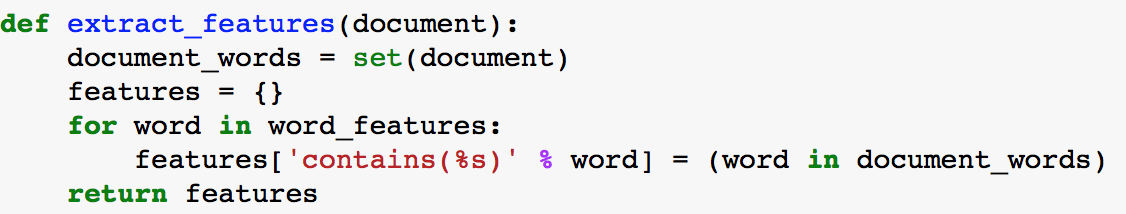
Once we created our training data, we create a single list of tuples with words and type of sentiment. We removed the words smaller than 2 characters.



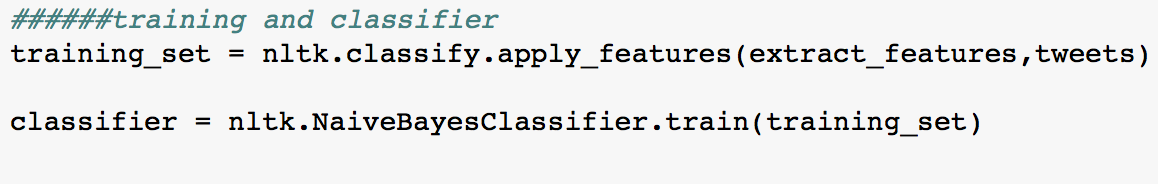
We then used the following code to get a list of distinct words ordered by frequency.



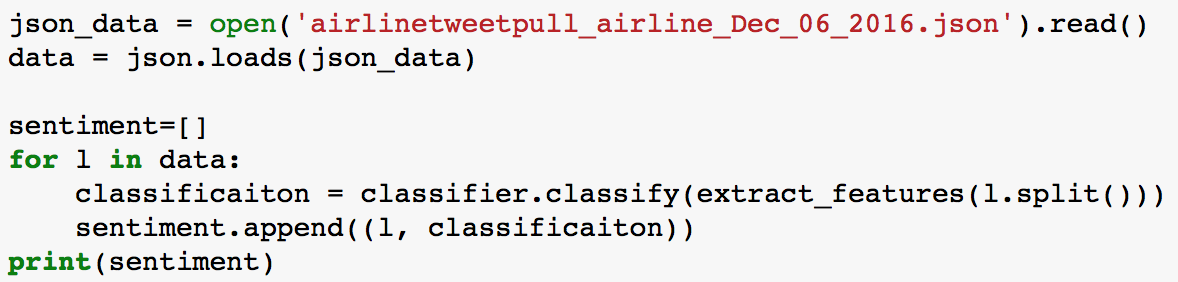
Now we create a classifier.



Next, we set the training set and classifier and apply them to the data.



Then write a new file with the tweet and prediction.



Regression Analysis in Stata

While building, and training our model, we used Stata to see if we could find a relationship between the airline and chosen sentiment.



Although each coefficient is statistically significant, the low R-square suggests that this may not be a perfect model for explaining all sentiment scores. It does however support the empirical findings from the summary tabulations and suggests that positive (3) tweets are uncommon and outliers.

We also regressed sentiment against airline, time, date, and even date bins, but did not find any significant relationships between those variables when controlling for airline choice. When we regressed against neg\_reason\_indicator, we found perfect multicollinearity so we could not draw any information with this. Below are the results from the regression that included the timebin indicator that split the day into quarters. The timebin variable was not statistically significant from zero. At the 95% level but was at the 92.2. Since the coefficient was so small economically, we did not include it in the presented analysis.



2015 Data - Descriptive Statistics

The Kaggle data set of U.S. airline sentiment from tweets included the classification of the Tweet: positive, negative or neutral. It also subcategorized the negative tweets into categories.

The tweets collected in 2015 were overwhelmingly negative.

United related tweets are the most common.

The breakdown of negative tweets is revealing. Customer service is by far the most prevalent problem for almost all airlines. Delta, unlike the others, has the biggest problem with late flights.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Reason | Customer Service | Late Flight | Uncertain | Cancelled | Lost Luggage | Bad Flight | Booking | Attendant | TOTAL |
| American | 40% | 13% | 10% | 13% | 8% | 5% | 7% | 5% | 100% |
| US Airways | 37% | 20% | 11% | 9% | 7% | 5% | 6% | 6% | 100% |
| Southwest | 34% | 13% | 14% | 14% | 9% | 8% | 5% | 3% | 100% |
| Virgin | 34% | 10% | 12% | 10% | 5% | 11% | 16% | 3% | 100% |
| United | 26% | 20% | 15% | 7% | 11% | 8% | 6% | 6% | 100% |
| Delta | 21% | 29% | 20% | 5% | 7% | 7% | 5% | 6% | 100% |

Removing the specific airline segmentation, the customer service problem rises to the top, coming in at almost double the second most common negative tweet reason: late flights.

Accumulative sentiment for the tweets show what you would expect. The polarity is in the negative, because most of the tweets are negative. The subjectivity shows that tweets are emotional.

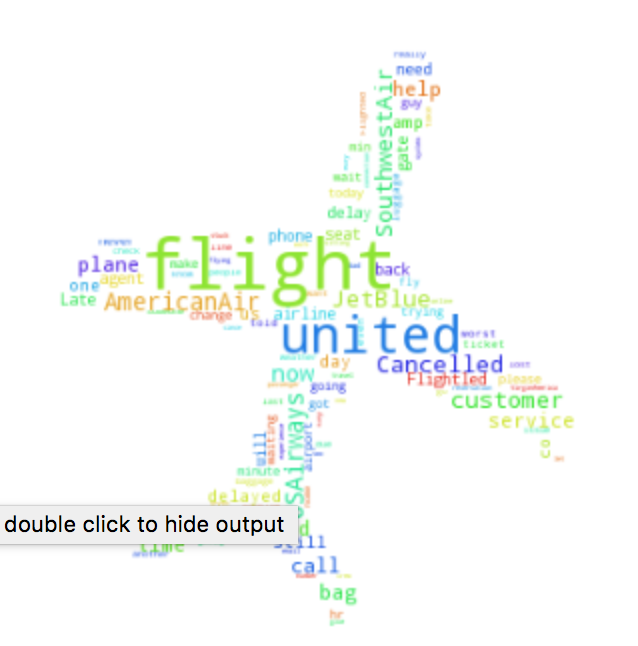
polarity=-0.14052927248717406, subjectivity=0.5788620278262808)

Showing market share along with tweet sentiment, we get a clearer picture of how each airline stands with their customers.

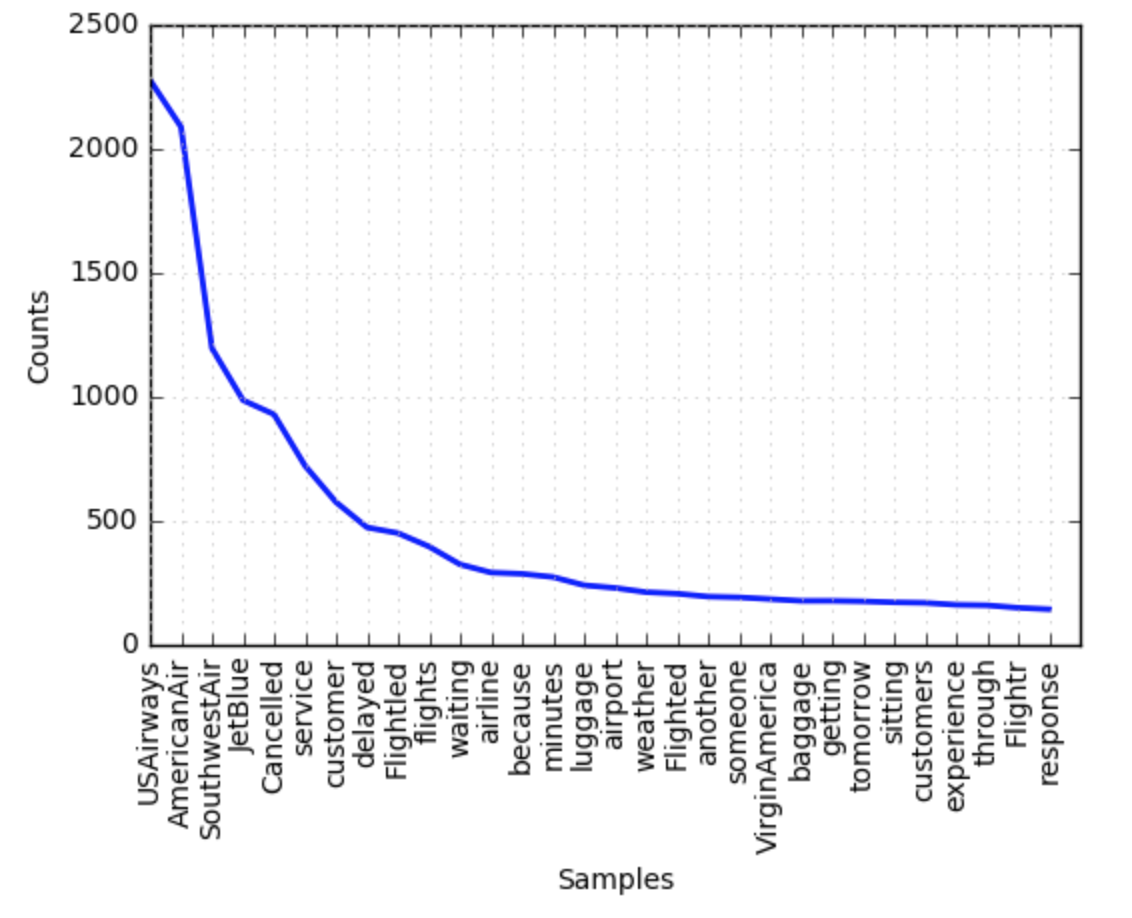
We adjusted the negative tweet percentage by market share. U.S. Airways show a small market share with an extremely high percentage of negative tweets.

|  |  |  |  |
| --- | --- | --- | --- |
| AIRLINE | NEGATIVES | ADJUSTED MARKET  SHARE | DIFFERENCE |
| UNITED | 28.69% | 20.73% | +7.95 |
| US AIR | 24.66% | 4.65% | +20.0 |
| AMERICAN | 21.35% | 23.41% | (2.05) |
| SW | 12.92% | 25.66% | (12.74) |
| DELTA | 10.41% | 23.97% | (13.57) |
| VIRGIN | 1.97% | 1.55% | +0.42 |

We created a word cloud with the highest frequency words in the 2015 data.



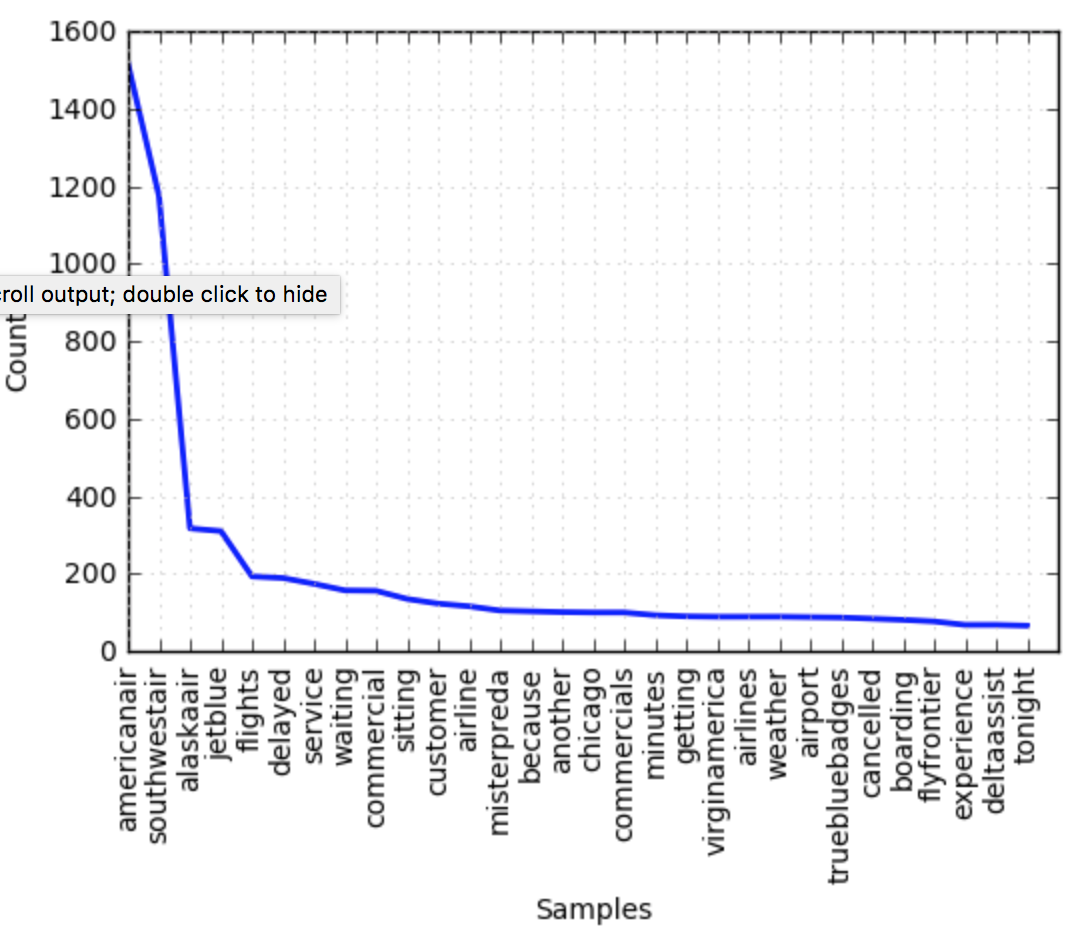
We also created a word frequency graph from the 2015 data. Not surprising ‘canceled’ is the most common word, behind airline names.



2016 Data - Descriptive Statistics

The data we pulled and predicted sentiment on showed a maintained volume of negative tweets. Both datasets, 2015 & 2016, show negative tweets coming in at around 60% of the total airline tweets. Our model either needs more training on the neutral category, or people are just more subjective than they used to be, I'm guessing the former.

The top four words from 2016, after airline names, were (1) flight, (2) delayed (3) service (4) waiting. Compared with the top four from 2015 (1) canceled, (2) service, (3) customer, (4) delayed. From this I would assume more flights were canceled in 2015.



The word cloud for the 2016 tweets is similar to the 2015 cloud.



Conclusion

We concluded that, if you want to fly among the happiest people you'll buy a ticket on Virgin Airlines. If you want to see the stress that is traveling, you'll hop on a U.S. Airways flight, but bring your patients.

This project gave our team a good understanding of how multiple data sources and platforms can be used to paint a clear story from the data.

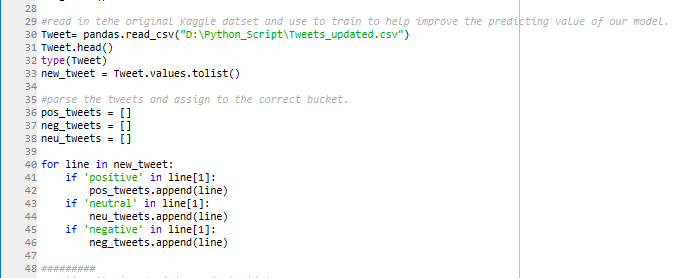
Addendum

A few of the questions remained. Could we get better results in training our model and based on these results if we took a larger number of current tweets would we see the same outcome. One thing that we have to take into consideration is the merger of US Airways with American Airlines. How would this impact the results of American?

In an attempt to answer the first question, we took the original Kaggle dataset which contained the original tweet along with the sentiment that was associated to those tweets. We narrowed the data down to simply the tweet and the sentiment (see in the example below.) The new file had approximately 14,600 tweets which had been assigned a sentiment.

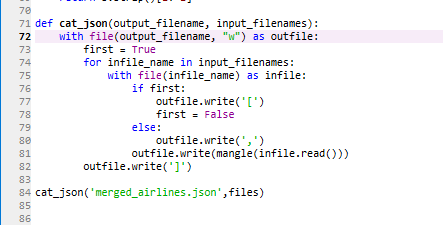
|  |  |
| --- | --- |
| **Text** | **airline\_sentiment** |
| **@VirginAmerica What @dhepburn said.** | **neutral** |
| **@VirginAmerica plus you've added commercials to the experience... tacky.** | **positive** |
| **@VirginAmerica I didn't today... Must mean I need to take another trip!** | **neutral** |
| **@VirginAmerica it's really aggressive to blast obnoxious "entertainment" in your guests' faces &amp; they have little recourse** | **negative** |
| **@VirginAmerica and it's a really big bad thing about it** | **negative** |
| **@VirginAmerica seriously would pay $30 a flight for seats that didn't have this playing. it's really the only bad thing about flying VA** | **negative** |

We then modified the code slightly to load this newly created file. We then used Pandas in order to read the file and convert the data into a list of values. Then we iterated over the list and assigned the tweets along with the associated sentiment to an appropriate variable which represented positive tweets, neutral tweets or negative tweets.



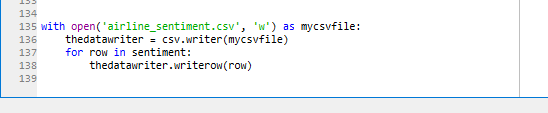
We then proceeded to collect tweets for the major airlines. Over the span of six days (December 4th - December 10th we collected over 30k unique tweets sent directly to the airlines twitter accounts.

As we had multiple distinct json files and to simplify the processing of the data. In order to accomplish this, we needed to combine the json files into a single file. We created a simple function to help with this process.



We were them able to load this data and feed it to our model which we had trained with the 14k Kaggle tweets. The outcome although slow, was effective.

Again, me made slight modification to our original code in order to use csv to generate a new csv file which contained the tweet and the sentiment assigned by the model. This would make any future analysis easier.



As we began to analysis the data a deficiency was discovered. Not all the tweet text captured listed an airline. This was because we pulled the tweets based on the twitter account and not the twitter text. As we cleaned up the data we were left with 21,927 tweets which we could analysis.



Over 80% of the tweets were Negative. The fact that the positive tweets remained at 6% lends one to believe that the model did a better job of classifying the negative tweets. Originally a greater number of these tweets would have been identified as a neutral tweet. We think this was because of the large data set that we were able to use to train the data.

When you looked at the tweet break down by airlines we see that United had the highest percentage of negative tweets at 45%. Followed by American at 32% and Southwest at 23%. The fact that Delta is coming in at just 1% leads us to believe that people don’t complain to the DeltaAssist Twitter account.



Conclusion

The conclusion of the additional work is that we definitely able to improve on the models ability to identify sentiment of a tweet by using the Kaggle dataset as a training set.

The results of this additional work lead us to believe that United, at this time, is receiving the most negative tweets and this correlates with the recent announcement that they would begin charging a fee for carry-on bags.

Again, more improvements to the process could be made to made.