Using Airline Tweet Sentiment Analysis as a Managerial Business Advantage CSI 5810

Adam DeJans

December 9, 2017

Date Presented: December 7, 2017
Instructor: Professor Sethi

The following report is arranged as five parts. The first section gives the objective and some background information on the problem at hand. In the second section, the data collection process is introduced. In the third section preprocessing and cleansing of the data in order to make sense of it is discussed. The fourth section elaborates on the sentiment analysis and the packages used to do this analysis. Some important features of the data are discussed within this section and interesting visualizations of the dataset are shown as a result of the sentiment analysis. In the fifth section, the conclusion along with an interesting result. The final section also includes some concerns of improvement.

1 Objective and Background Information

Twitter is one of the largest social media platforms that exist in the U.S. today. The great thing about twitter, unlike other social media platforms, is almost all of the user profiles are public. This means that there is a lot of free public data out there that can be used as a business advantage. By exploiting the fact that tweets are public businesses can see trends about users location, what certain users like, and what theyre likely to buy amongst other things and use this information to make marketing decisions. The data can also be used to gain insight on the public relations of a business, which is the direction we will focus on.

The objective of this project is to do a sentiment analysis on twitter data to provide a business case, or a general overall feeling, regarding customers satisfaction amongst major U.S. airliners. If done correctly data can be used by an airline company to determine what customers like and dislike the most and perhaps gear training and resources appropriately to give rise to greater customer satisfaction and lower customer dissatisfaction. Due to this approach we

will rank a set of chosen major U.S. airliners but also keep in mind the business aspect of using the data to make decisions.

We will consider four major U.S. airliners: American, Delta, Southwest, and United airlines.

2 Data Collection

The tweets were collected using the tweepy API. Tweepy is an open-sourced and enables the Python programming language to communicate with Twitter platform and use its API. The data was collected over the course of the week from 11/28/2017 to 12/07/2017 by querying the following associated hashtags per airline:

- American
  - #americanairlines
  - #americanair
- Delta
  - #deltaairlines
  - #deltaair
- Southwest
  - #southwestairlines
  - #wouthwestair
- United
  - #unitedairlines
  - #unitedair

TOTAL	American	Delta	Southwest	United
2577	1584	178	375	440

3 Preprocessing the Data

When the tweets come in they are not in an easy format for any type of analysis. To overcome this there was a series of steps taken to preprocess the data: tokenization of the tweets, use of regular expressions, and elimination of stop words.

- Tokenization:
  - Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens, perhaps at the same time throwing away certain characters, such as punctuation.
  - This was done with the Natural Language Tool Kit (NLTK) package in python.

- Regular Expressions:
  - These were first used to move all words to lower case so that words with different capitalization were not double counted. After this all URLs and additional spaces were eliminated. Finally hashtags had the hash removed from the word (e.g., "#hashtag" was converted to "hashtag").
  - This was done with the Regular Expressions (re) package in python.
- Regular Expressions:
  - Words such as "the", "is", and "a" were removed as these do not add any value.
  - This was done by importing "STOPWORDS" from the *wordcloud* package on python.

We can see a glimpse of United Airlines *pandas* dataframe in the following figure which stores the original tweet along with the *normalized* tweet.

	Tweet	normalized_tweet
0	RT @Cshells33Wells: SIGN PETITION➡ Request to #UnitedAirlines for free air travel for dogs rescued from meat trade in #SouthKorea✓	[well, sign, petition, request, unitedairlines, free, air, travel, dog, rescued, meat, trade, southkorea]
1	1 downside of @united 's increased in ontime performance is less chances to use these. Sad to throw them out…	[united, increased, ontime, performance, le, chance, use, sad, throw]
2	@kategebo I had a great time with your magical @united team at #UAFantasyFlights. It was my second year helping out…	[great, time, magical, united, team, uafantasyflights, second, year, helping]
3	New York to Oslo, Norway for only $318 roundtrip with United. #UnitedAirlines #NewYork	[oslo, norway, roundtrip, united, unitedairlines, newyork]
4	A view of our private beach 🍁🍁🍁🌴🌴🌴#privatebeach #caribbeanbeach #surf #TnT #tobago #trinidad\n #everywhere…	[private, beach, privatebeach, caribbeanbeach, surf, tnt, tobago, trinidad, everywhere]

Figure 1: An example of tweets from United Airlines dataframe being normalized.

4 Sentiment Analysis

The sentiment analysis of the tweets was done by importing the "TextBlob" package. Using this package were able to determine the polarity of each tweet with the call *analysis.sentiment.polarity()*. Here polarity is defined as the score of tweet within the range $[-1.0, 1.0]$, which says that a tweet has a higher positive sentiment the higher the score greater than 0, a high negative sentiment if the score is closer to -1.0 and less than 0, and the tweet is mostly neutral if the score is at 0.

The next step was to classify the tweets into the three natural categories: positive, negative, and neutral. To adjust the error for tweets that are mostly neutral but may be given a slight positive/negative score by ± 0.05, the following bins were used:

Negative	Neutral	Positive
-1 to -0.1	-0.1 to 0.1	0.1 to 1.0

The figure displays the results of this binning process.

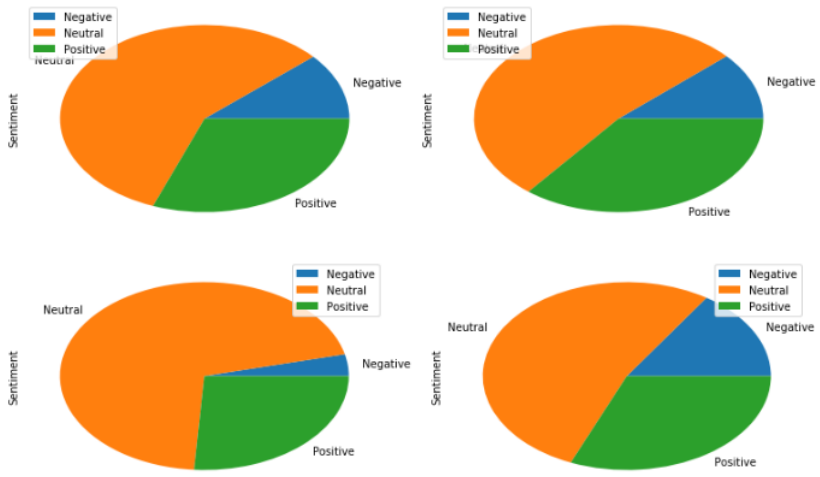

Figure 2: American, Delta, Southwest, and United in respective order (left to right)

Discarding the neutral tweets we see in the following figure the number of positive and negative tweets per airline.

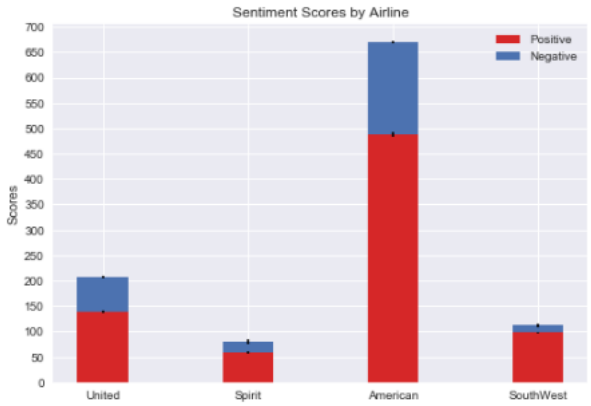

Figure 3: Barchart displaying the number of tweets binned positive and negative across all airlines.

While considering the number of positive versus negative tweets is one way to compare the airlines, we take note that this may not capture the full picture.

Consider the case that the magnitude of the sum of the polarity of negative tweets with American Airlines is much lower than that of United Airlines. In this case it may be to the advantage of American Airlines. For example, imagine that you are comparing two people on a background investigation and person A has three civil infractions and person B has a felony. It is clear that although person A has three negative ticks against him, he is likely to be a much better candidate than person B who only has one tick against him but a very severe tick.

Taking this type of thought into consideration we look at the cumulative sum of polarity scores over the course of the week per airline.

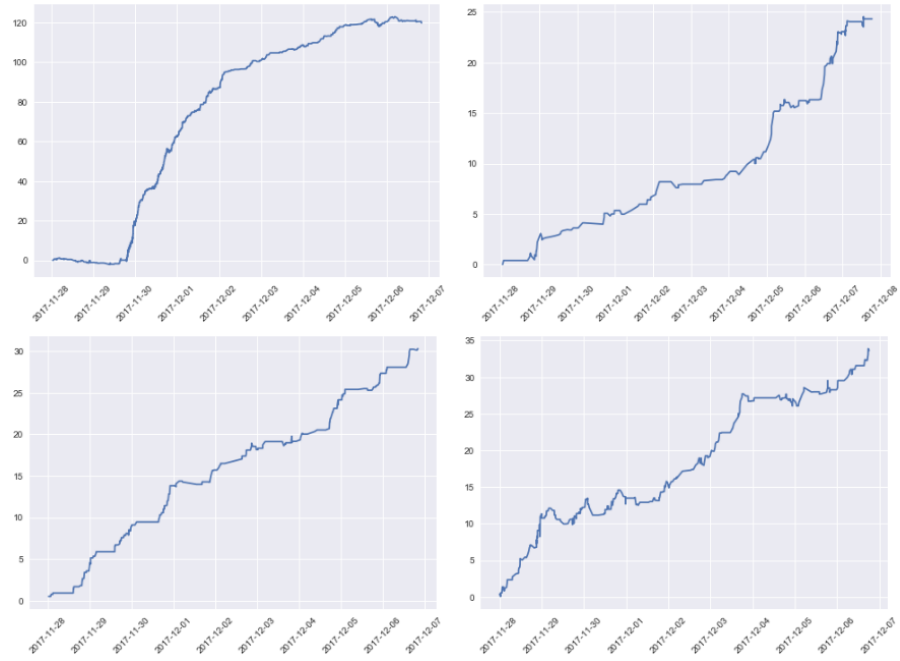

Figure 4: American, Delta, Southwest, and United in respective order (left to right)

We see that the trend amongst all airlines is closely similar with perhaps American airlines taking a different shape.[1]

Another interesting aspect to delve into is on what days were the most negative tweets tweeted on. One interesting aspect is that a large percentage of negative tweets occur during the beginning of the work week. We also see that this corresponds with the cumulative sums flatting out on the beginning of the workweek days in the previous figures.

[1]We note that American airlines also had by far the most tweets.

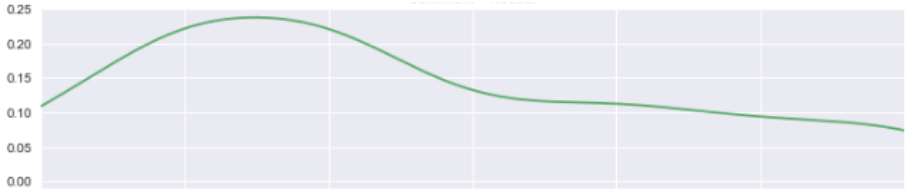

Figure 5: Percentage of negative tweets across days of the week. Notice 0 is Monday and 6 is Sunday. This chart was made with *datetime* package on python.

While this data may seem strange it is good to consider that there are many companies that send employees for travel at the beginning of the work-week. This information can be used to perhaps notify managerial heads to schedule more people to assure better customer service during these highly negative tweet days.

Next we look at the top negative words for each of the airlines via word-clouds:

Figure 6: American, Delta, Southwest, and United neative word clouds.

Perhaps the data encoded in these clouds can be useful keywords that could be

used as indicators to the company as areas to work on to boost public relations. We see keywords such as:

- Slow
- Mistake / Careless
- Precheck
- Reseat
- Refusing/Refused
- Bags
- Sorry
- Staff
- Rude

This could tell a managerial staff that maybe employees need to be trained on how to handle certain situations so that they do not come off as rude or careless. From this list we may also want to look into improving processes involving bags as this word is frequent in negative tweets.

Another thing that we can do is explore what positive things people are saying about airliners. Perhaps using these top keywords an airliner could reinforce these to retain customers. Some top words we see are:

- Free
- New
- Many
- Pilot(s)
- Customer
- Service

These top words can be a way to see what airline customers really appreciate. As for this analysis it seems that the customers like when they're given something (probably anything) free. We also can see that there was great customer service and there was probably not a lack of some desired resource as we can see that the word *many* occurs frequently amongst positive tweets.

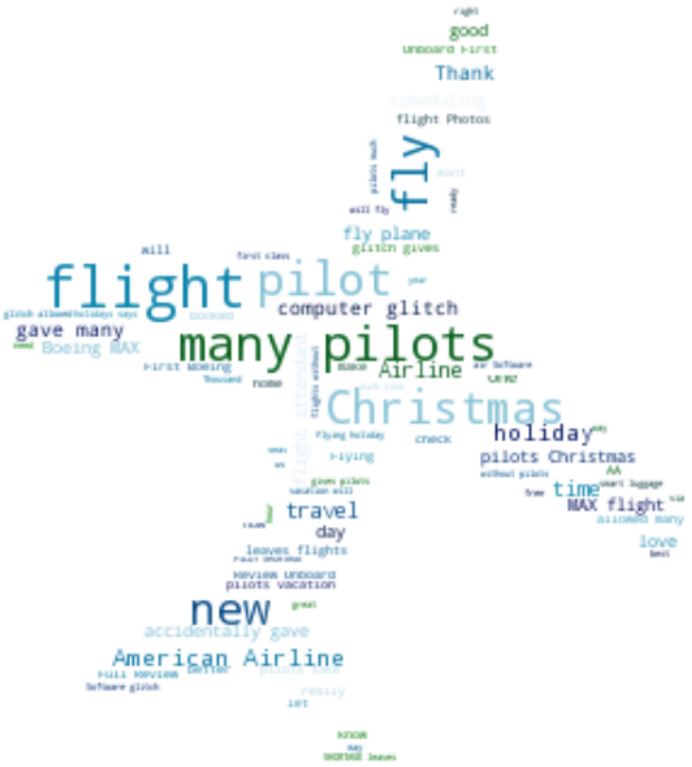

Figure 7: Positive word cloud in th shape of an airplane amongst airliners.

5 Conclusions and Closing Remarks

Theres no clear distinction on the top reasons why airlines receive negative feelings as this analysis was done using top negative words. We can however see that customer service seems to be very important as this occurred in both positive and negative tweets, perhaps being the deciding factor? Almost all airlines suffered from the same top negative words however.

We see that being timely with great and friendly customer service is likely the key to not having negative tweets and perhaps gaining positive publicity (as discussed above). Possibly with a bigger dataset, more conclusive results would arise, but one-week of data does not seem very sufficient. We thus cannot conclude to whether twitter is a sufficient way to rank airlines due to our relatively small number of tweets (and bias) but it may be a good indicator of what airlines need to work on.

With a weeks worth of data however were able to see that overall Delta and Southwest have the best ratio of positive to negative tweets, followed by United and then American airlines. One very interesting result comes from this, and that is that the ordering obtained by the sentiment analysis on the week of data collected follows precisely the ranking of these four major U.S. airlines that is given on the CBS news review.[2]

6 Learning Outcomes, and Difficulties

6.1 Learning Outcomes

From an informational point of view, prior to this project I never considered the idea of Twitter mining nor did I consider the wide array of information and approaches would could take on mining Twitter data. I also never fathomed the valuable data that Twitter can provide due to the fact that most accounts are public. Seeing the presentations of my peers brought light to many ideas and possibilities of using Twitter as a source of data.

From a technical standpoint a lot was learned. I never knew that there was an API like Tweepy and I never used any of the natural language processing tool-kits. This project exposed me to many packages Ive never used before. I found the idea of tokenization and the use of regular expressions to be especially interesting in the project. I also learned creative visualization techniques such as making a word cloud in the form of an airplane.

6.2 Difficulties Recognized

One important realization of how much work is yet to come in sentiment analysis and natural language processing is the recognition of sarcasm. For example there was a tweet that said "My favorite airline of all time #SouthwestAirlines." Almost all humans would recognize this as sarcasm and see that there is a tone of disappointment in the tweet by the use of the ellipsis; however this is very difficult for a machine (at least with the out-of-the-box packages) to recognize.

Another aspect that I found difficult, mostly due to my inexperience, was being able to live-stream capture tweets. This was mostly a learning experience as I've never done anything like it but after seeing how it was done it is not very difficult. One last difficulty I learned was that it's hard to collect old tweets, for example the Tweepy API only lets you go back in time for up to a week.

[2] *https://www.cbsnews.com/media/these-are-the-best-and-worst-u-s-airlines/14/*