U1298774

Abstract

Text Analysis is an ever-evolving and improving technology to understand the feelings and thoughts of customers. This report will discuss the sentiment that Twitter users have for United, American and Delta Airlines and how they’re doing during the Covid-19 pandemic

[Covid and Air Travel Using Sentiment Analysis, LDA and LSA – a Final Project Report](#OLE_LINK2)

[Bibliography](#OLE_LINK19)

[Do We Feel Ready to Fly Again? – A Text Analysis Project Proposal](#OLE_LINK8)

Manali Harinkhere

U1350577

MKTG 6640-001 Text Analytics

7/22/21

**Do We Feel Ready to Fly Again? – A Project Proposal**

Since Covid-19 sprung into our lives, most Americans feel very uneasy about flying in an airplane. There is a tremendous amount of negative sentiment and uncertainty about catching the virus and how exposed we are as we travel. The airlines and the travel industry have suffered tremendously because of it and I would like to explore these sentiments using Twitter and sentiment analysis for American Airlines, Delta and Jet Blue. Specifically, I intend to explore how they compare between companies, the ways they are succeeding and how they are failing in the eyes of the American public.

The proposed project will entail using the Twitter Search Tweets API and pulling 50,000 tweets using various hashtags involving the aforementioned airlines and Covid-19 or Coronavirus. I am interested in this topic because I would like to know more about what the airlines are doing well and how best to successfully navigate air travel, potentially in the near future (if possible). My family loves to travel and we’re all planning trips we can’t go on until things settle down.

To effectively complete this project, I will need to study more and implement Sentiment Analysis, Latent Semantic Analysis and Latent Dirichlet Allocation. I will need to begin by cleaning the data, performing exploratory analysis and comparing the results from LSA and LDA analysis. I will definitely need to be watchful for examples of polysemy and synonymy and for things like sarcasm. Other potential problems that will need to be addressed successfully could be slang, shorthand writing styles (e.g., R U OK) and misspellings.

Once I get a better feel for the data and what it looks like, I will use Machine Learning with R by Brett Lantz as a reference in preparing and standardizing text data for analysis and will need to find and implement other resources for Sentiment Analysis.

**Covid and Air Travel Using Sentiment Analysis, LDA and LSA – A Final Project Report**

The expectation going into this project was to use LSA, LDA and Sentiment Analysis to better understand the feelings of the public as they relate to the Airline industry and the steps they’re taking to keep the public safe from the virus known as Covid-19. As I started this project, my expectations were quickly derailed as I found that pulling 50,000 tweets each about Covid, American Airlines, Delta Airlines and Jet Blue Airlines was an overestimation. The first change I had to make was to exclude any mention of Covid in order to pull in enough tweets to complete an analysis. By excluding Covid specifically, I could still get tweets about events occurring in the Airline industry due to the impact of Covid even though it was not actually mentioned by name. This was a good assumption as will become apparent later in the analysis.

Using the rtweet and twitter libraries, I pulled in 26,697 American Airline tweets, 7,272 Delta Airline tweets but only 1,771 Jet Blue tweets. Due to the large disparity between the airlines, I had to make another deviation from my original plan. I decided to try using United Airlines instead of Jet Blue and I was rewarded by pulling in 17,881 United Airline tweets. My entire dataset now consists of a total of 51,580 tweets made up of 51.5% from American Airlines (American), 14% from Delta Airlines (Delta) and 34.5% from United Airlines (United). We will come to see that the number of tweets relates directly to public sentiment.

I began my EDA by quickly combining all the individual datasets into one to review some summary statistics from the group as a whole. There were two things that stood out during my review. First, that the apps Twitter for iPhone and Twitter for iPad (both Apple products) were used 45.26% of the time which is 16% more than the next highest option, Twitter for Android. Second, I was surprised that 85% of all the tweets in this dataset were retweets. Not being a Twitter user myself, my expectation from reading about Twitter users on the news and social media is that its users put their own personal thoughts out into the world. I expected original content and personal thought to be a much bigger proportion of the content. My presumptive opinion is that these high number of retweets can perhaps give us a better indication of human nature; that most people are followers. In spite of my opinions, I felt it was important to keep the retweets because notwithstanding my reservations about the humans behind the tweets, for the most part, they still demonstrate what topics and events resonate with Twitter users and therefore have value in text and sentiment analysis.

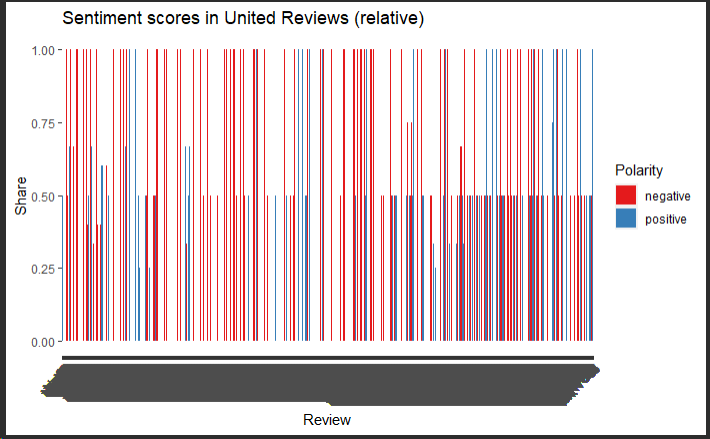
I performed Sentiment Analysis, LDA and then LSA on the data from each airline and began with United, American and then ended with Delta. While I followed that pattern for my analysis, it is better to compare and contrast the airlines as a group based on the topic of exploration being used at that moment which is how the discussion will follow. Before we move on, it is important to know that there was at least one main news story occurring about each airline when these tweets were pulled. For United, they had just announced to their employees that many should expect to be furloughed and that the company had lost billions of dollars in the 2nd quarter of 2020 due to demand reductions from Covid. For American, Senator Ted Cruz had just been “mask shamed” publicly for taking off his mask during a flight and American did not ask him to put it back on. Finally, Delta had announced that they lost $6 Billion dollars in the 2nd quarter of 2020 and the President of Mexico had just flown a Delta commercial flight to Washington D.C. to meet with President Trump.

**Sentiment Analysis**

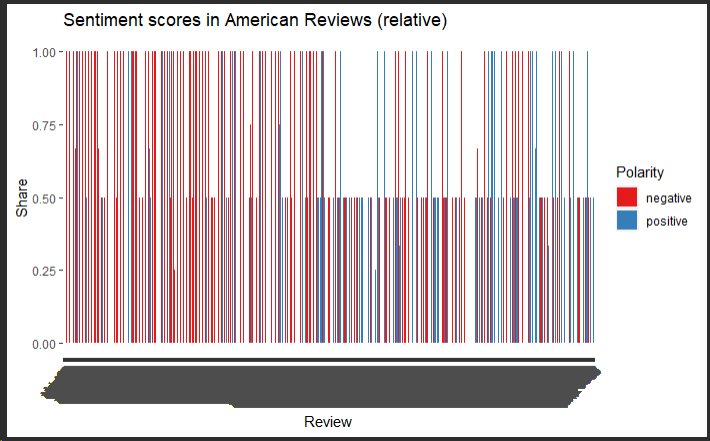
Each of the airlines averaged around 17-20 characters per tweet but what was much more interesting was the breakdown of Apple product users and Android users to make those tweets. For United, 40.5% used Apple while 26.4% used Android. For American, 51.23% used Apple while only 21.73% used Android. Here’s where it gets interesting. For Delta, 35.079% used Android and 35.051% used Apple. If we had more time and data, it would be interesting to dive more into the demographics and socioeconomics behind this phenomenon. As we will soon see, word clouds computed for United and American contained words almost entirely in English while the word cloud for Delta was mixed pretty evenly between English and Spanish. Luckily, I am fluent in Spanish so I translated the sentiment as if all of it were in English.

United tweets were retweeted 77% of the time while American tweets were retweeted a whopping 93%. Delta tweets were the lowest of the group with retweets at 73%. While these stats may be interesting for some, the really interesting part comes in the sentiment analysis.

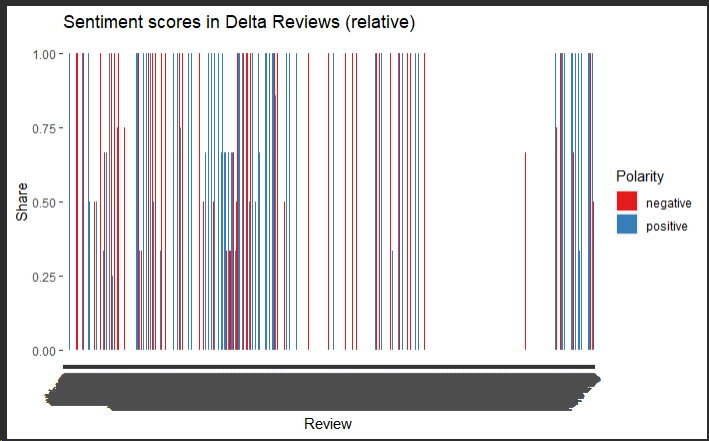
United’s word cloud was heavily influenced by the bad news about losing “billions” and the “layoff” of employees. The sentiment was especially strong because United had just “spent” billions of dollars from their cash reserves to “buyback” company stock right before Covid hit and now they’re “cut”ting “jobs”. The negative sentiment showed. United had almost 18,000 tweets and the mean Positive Sentiment was only 33.27% while the mean Negative Sentiment was 62.97%. As we dig deeper using the liwcalike() function we see that there are references to anger, fear, surprise and trust. Using ggplot, we can visualize the same negative sentiment.



American’s word cloud was dominated by outrage over the Senator Cruz mask debacle. So much so that I had to adjust my minimum word count from 150 to 1000 just to cut out a lot of the noise. Words included are “ted”, “cruz”, “wear” and “mask”. American had over 26,600 tweets and the mean Positive Sentiment was only 43% while the mean Negative Sentiment was 97.5%. As we dig deeper using the liwcalike() function we see that there are references to disgust, fear, sadness, anger and exclamation. Using ggplot, we can again visualize the same negative sentiment.



The sentiment for Delta’s word cloud was more of a mixed bag though and included a large percentage of words in Spanish. The largest focus of words included “@lopezobrador” (the Mexican Presidents last name), “vuelo” (flew), “llevara” (will arrive) and “abordo” (on board) with other words like “Covid” and “pandemic” mixed in as well. Delta had the fewest tweets at just over 7,200 and the mean Positive Sentiment was 40% while the mean Negative Sentiment was 39%. As we dig deeper using the liwcalike() function we see what was affirmed above, negative and positive sentiment are balanced. Using ggplot, we visualize the same sentiment.



**Latent Dirichlet Allocation**

The information gained from LDA was pretty limited overall but it did confirm what we found in Semantic Analysis by giving us key topics that were most prevalent in the Tweets. As we analyze top terms, if we exclude United and Airlines, the top terms are “billion”, “employee”, “bailout”, “stock”, “warn” and “layoff”. If we remove American and Airlines, their top terms are “ted”, “cruz”, “wear”, “mask”, and “investigate”.

If we exclude Delta and Airlines, their top terms are all in Spanish with "#loultimo" (#thebest), "abordo" (on board), "vuelo" (flight), "el presidente" (the president). It appears that when the President of Mexico flew a commercial flight to visit President Trump in Washington, that caused quite the stir among Mexican Twitter users. It appears that so far and even though they announced losing billions of dollars like United, Delta is keeping away from any missteps that are swinging public opinion negatively like the other airlines.

**Latent Semantic Analysis**

While Semantic Analysis and LDA provided a nice start to the data, I felt like I learned the most using LSA. Calculating Cosine Similarity and using the neighbors function really helped round out the picture of what was the most important message for each airline coming from these tweets. While the purpose of my project originally was to compare airline customers thoughts on the airlines’ response to safety issues with the ongoing pandemic, what I actually found was more of an indirect influence by the virus.

For United, the cosine similarity points more towards "warns" and "laid" (as in laid off) at 98% and "warns" "workers" at 64.8%. Comparing "United" and "virus" directly has a cosine similarity of only 31.4%. For American, the cosine similarity points toward the customers demanding that American be held accountable for lackluster safety enforcement through words like "hold", "accountable", "putting", "flyers", "lives", and "risk" at 94%. Finally, for Delta, nearly all of the words are in Spanish and have a high cosine similarity as “president” (President) and “vuello” (flight) are at 99.9% as an example. To a large extent, these results show the impact Covid is having on the world as a whole.

Since my original goal was to directly relate the airlines and Covid-19 here is a list of words that I thought would be interesting calculated with the:

United dataset

* + "Covid" - "halt", "tighter" & "services" at 64%
  + "virus" - "worst", "case", "scenario" & "outbreak" at 96.9%
  + "coronavirus" - not as interesting as the other options but is neighbors to "news" and "bbc" in the high 95% range as I'm sure it's the more frequently used term for Covid-19 by the news agencies
  + "stock" – appears people are questioning the unfortunate timing of United's recent choice to buy back stock as "stock" is highly correlated to "times", "tough" and "rainy" at nearly 99% each
  + If I use plot\_neighbors(“virus”), we get almost a clear sentence - "worst” “case” “scenario” “outbreak” “officials” “signal” “hurting” “breaking” & “airline"

American dataset

* "Covid" – result came up as NA
* "virus" - Certainly no business wants to be associated with "spreading", "virus" & "willingly" at 100%
* "coronavirus" - result came up as NA
* "american" - is associated with the name Ted Cruz and his story at 85% cosine similarity, while “mask” is 80% and “GOP” is nearly 75%. The world is hyper political
* "safety" - there is a high association with "disdain" at 99.9% along with "commitment" and "protecting" at nearly 98%
* If I use plot\_neighbors("virus") - Virus is associated with "spreading”, “virus”, and “willingly" and appears to be calling into question American Airlines willingness to allow a Platinum level customer to skirt the rules with words like "platinum" and "customer".

Delta dataset

* + "Covid" - "tested" at 97.5%, "urge" and "prevent" at 96.8% and "health" at 96.3
  + "virus" came up NA
  + "coronavirus" - most closely relates to "federal" and "strike" at nearly 98% and "deals" at 94%
  + "delta" - Back to Spanish, delta is associated with "avion" (airplane) at 84%, "llego" (arrived) and "contarnos" (count us!) at 48%
  + "safety" - For Safety, there is a moderately high association with "crew", "explaining" and "strain" in the mid 70% range
  + If I use plot\_neighbors("billion") - delta lost $6 Billion in the 2nd quarter of 2020. The highest correlated neighbors are "midst", "considerable" and "turbulence" surely speaking of the impact Covid is having on the industry

Overall, I thought this was a great project and I learned so much about the application of Semantic Analysis, LDA and LSA on tweets. I also gained a new respect for LSA. I anticipated that Semantic Analysis would give me all the information I would need but really, all three methods provided a better picture of the Tweeter’s thoughts and ideas about Covid-19 and the impact it is having on the airline industry.

**Bibliography**

Sipra, Vajiha. “Twitter Sentiment Analysis and Visualization using R.” Towards Data Science.

19 January, <https://towardsdatascience.com/twitter-sentiment-analysis-and-visualization-using-r-22e1f70f6967>

Lantz, Brett. *Machine Learning with R: Expert Techniques for Predictive Modeling*. Packt Publishing, 3rd Edition, 2019.