Text Mining Airline Tweets

Group 2: Leesann Sutherland, Maryan Mahamed, Stanislav Taov

16/02/2020

Introduction

Companies, for a long time, have relied of consumer feedback to help improve the quality of their product or service, and today consumers are presented with an abundance of avenues to provide that feedback, from yelp, to google +, to twitter and other social media platforms. This has led to the overwhelming accumulation of data, both structured and unstructured, across all industries globally. While structured data has provided the essential statistical insights to help companies improve their decision-making process, most data collected is unstructured in the form of text, video and audio. The challenge of unlocking valuable insights from these unstructured has proven to be pivotal to the success and continued growth of businesses as thy strive to meet consumer demands.

In this project, we explore the use of Text Mining as a means of extracting valuable, quantifiable business insights from consumer tweets. The data used, obtained from Kaggle, focuses on tweets scarped from Twitter in February 2015 concerning six major U.S. airlines. Through various data exploration techniques, we reveal some insights from the results of the previous sentiment analysis that led to the dataset used. It will be shown that the previous study sought to analyse negative sentiments, but do not explicitly reveal anything about the positives.

Our analysis seeks to dig deeper, analysing the tweets from scratch to not only determine the sentiments, but also the major topics for each. We use the Latent Dirichlet Allocation natural language topic modeller to extract meaning from the tweets, provide insights into the consumer experience, and discover what the companies are doing right and wrong.

The insights will be presented in a shiny app that will allow these companies to better gauge the sentiments of their customer-base, discover the areas of their serve that need improvement and assist them in their decision making process.

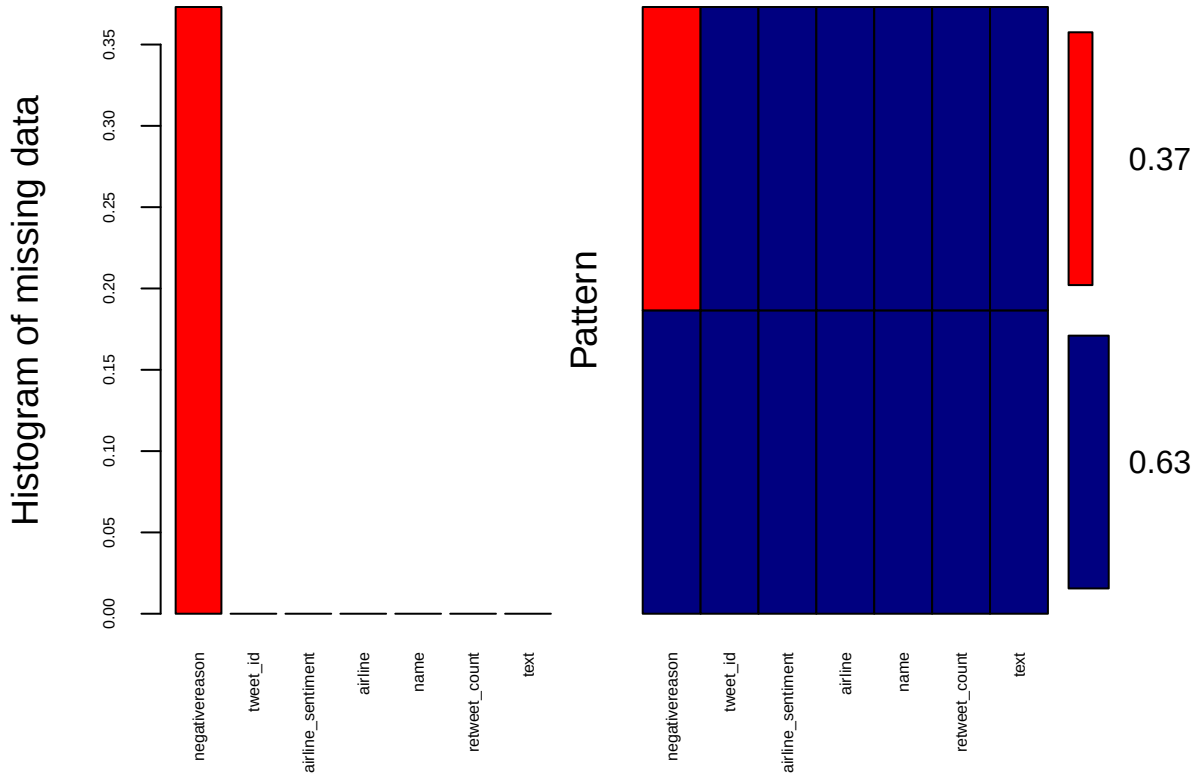
Data Preparation and Cleaning

The data used contained 15 variables with 14640 observations. The variables included the tweet_id number, airline_sentiment - a three level factor of Negative, Neutral and Positive, airline_seniment_confidence indicating the level of confidence in the sentiment categorization, negativereason - the dominant reason for the negative comment, negative_reason_confidence, airline, airline_sentiment_gold, name - the Twitter user, negativereason_gold, retweet_count, text - the actual tweet, tweet_coord, tweet_created, tweet_location, and user_timezone. The details can be seen below.

```
## Observations: 14,640
## Variables: 15
## $ tweet_id                     <dbl> 5.703061e+17, 5.703011e+17, 5.703011e+...
## $ airline_sentiment            <fct> neutral, positive, neutral, negative, ...
## $ airline_sentiment_confidence <dbl> 1.0000, 0.3486, 0.6837, 1.0000, 1.0000...
```

```
## $ negativereason               <fct> NA, NA, NA, Bad Flight, Can't Tell, Ca...
## $ negativereason_confidence    <dbl> NA, 0.0000, NA, 0.7033, 1.0000, 0.6842...
## $ airline                      <fct> Virgin America, Virgin America, Virgin...
## $ airline_sentiment_gold       <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ name                         <fct> cairdin, jnardino, yvonnalynn, jnardin...
## $ negativereason_gold          <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ retweet_count                <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ text                         <fct> "@VirginAmerica What @dhepburn said.",...
## $ tweet_coord                  <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ tweet_created                <fct> 2015-02-24 11:35:52 -0800, 2015-02-24 ...
## $ tweet_location               <fct> NA, NA, "Lets Play", NA, NA, NA, "San ...
## $ user_timezone                <fct> Eastern Time (US & Canada), Pacific Ti...
```

For the purposes of exploration goals, we removed the confidence columns, airline_sentiment_gold, negativereason_gold, user_ timezone, tweet_created, and location related columns, as they will not be used for our further analysis, and explored the 7 remaining columns for missing values. Of the remaining columns, it was discovered that there were 5462 missing values, all located in the negativereason column.




```
##
##  Variables sorted by number of missings:
##           Variable     Count
##     negativereason 0.3730874
##           tweet_id 0.0000000
##  airline_sentiment 0.0000000
##            airline 0.0000000
##               name 0.0000000
```

```
##      retweet_count 0.0000000
##               text 0.0000000
```

To discover why there were NA values present in the "negative reasons" column, the airline sentiment, negative reason and airline columns were isolated, and filtered for NA values. It was discovered that NA values only existed in the negative reasons column where the tweet sentiments were labeled as either positive or neutral, amounting to 5462 NA's. The NA values in this column were treated by replacing "NA" with an empty character string.

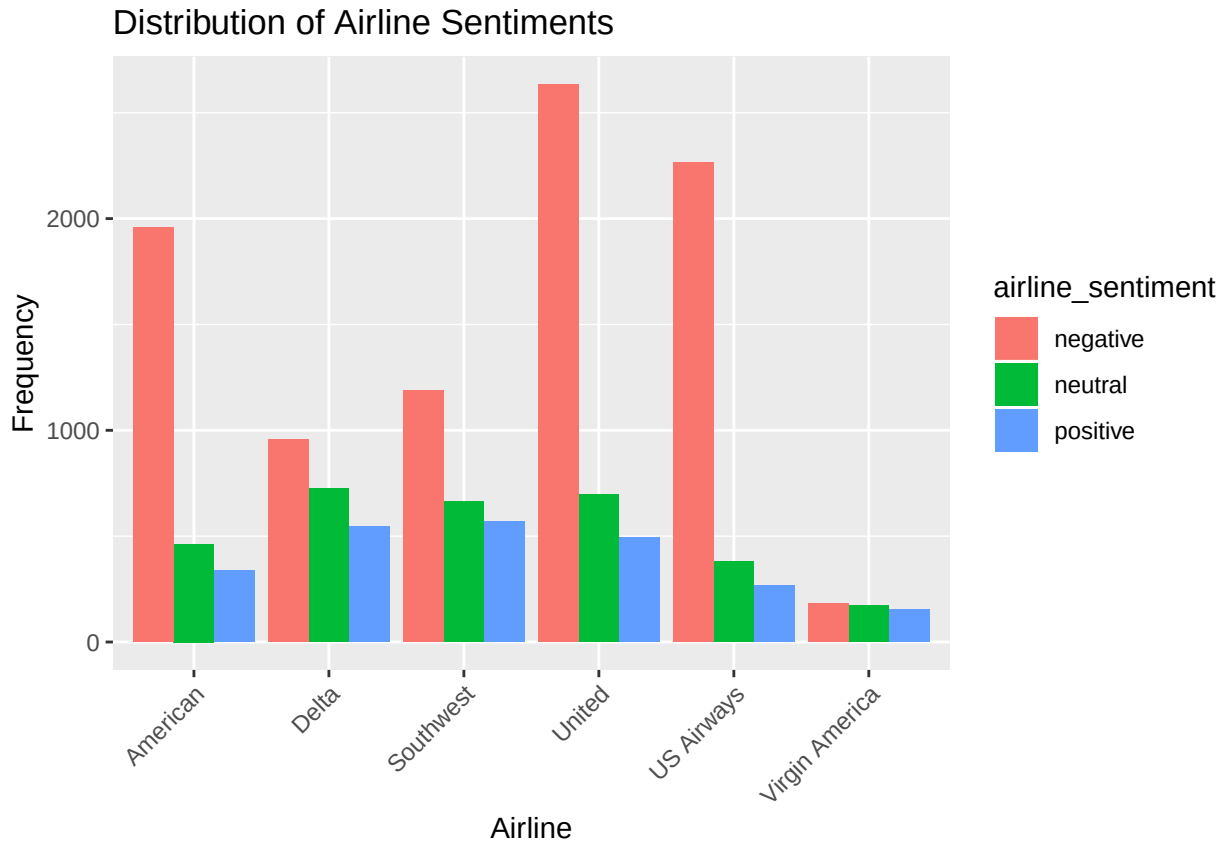
```
##  airline_sentiment                negativereason           airline
##  negative:   0     Bad Flight            :   0   American      : 799
##  neutral :3099     Can't Tell            :   0   Delta         :1267
##  positive:2363     Cancelled Flight      :   0   Southwest     :1234
##                    Customer Service Issue:   0   United        :1189
##                    Damaged Luggage       :   0   US Airways    : 650
##                    (Other)               :   0   Virgin America: 323
##                    NA's                  :5462
```

The details of the final dataframe used for data exploration are shown below.

```
## Observations: 14,640
## Variables: 7
## $ tweet_id          <dbl> 5.703061e+17, 5.703011e+17, 5.703011e+17, 5.70301...
## $ airline_sentiment <fct> neutral, positive, neutral, negative, negative, n...
## $ negativereason    <fct> , , , Bad Flight, Can't Tell, Can't Tell, , , , ,...
## $ airline           <fct> Virgin America, Virgin America, Virgin America, V...
## $ name              <fct> cairdin, jnardino, yvonnalynn, jnardino, jnardino...
## $ retweet_count     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ text              <fct> "@VirginAmerica What @dhepburn said.", "@VirginAm...
```

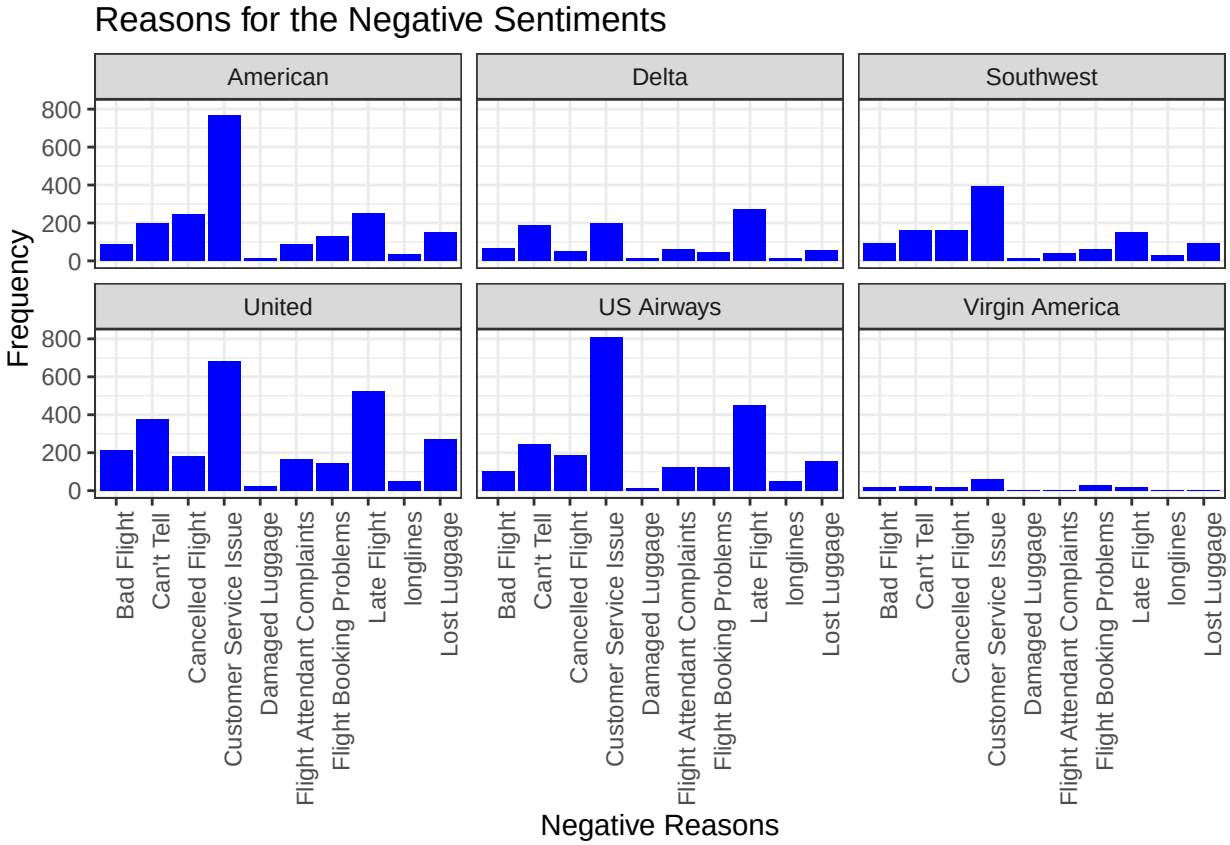
Data Exploration

In order to get an idea of the customer experience between the six airline companies, we explored the distribution of airline sentiments for each company. The graph below shows that negative sentiments far outweighs the positive or neutral sentiments in the cases of American Airlines, United and US Airways, while Virgin America performs the best. However, none of the airlines appear to be performing satisfactorily, which begs the question, what are they doing wrong and what are they doing right?



Negative Sentiments

There were no columns representing the positive or neutral reasons, further enforced by the presence of NA values found during the cleaning process, showing that previous work sought merely to explore the negative sentiments. We, therefore, first explored the reasons for the negative sentiments as previously determined. The data breaks down negative reasons into core categories as shown in the graph below. Apart from Delta, where the main complaint concerned flight tardiness, each of the other airline companies underperformed in customer service, with the worst airlines being American, United and US Airways.



These categories for negative sentiments can be considered quite general, therefore, to get a better idea of what words were used to express dissatifaction a word cloud was created. To do this, tokens were unnested from the text (tweets) column, and stop words were removed. A count of the negative words was calculated, the results of which are shown below.

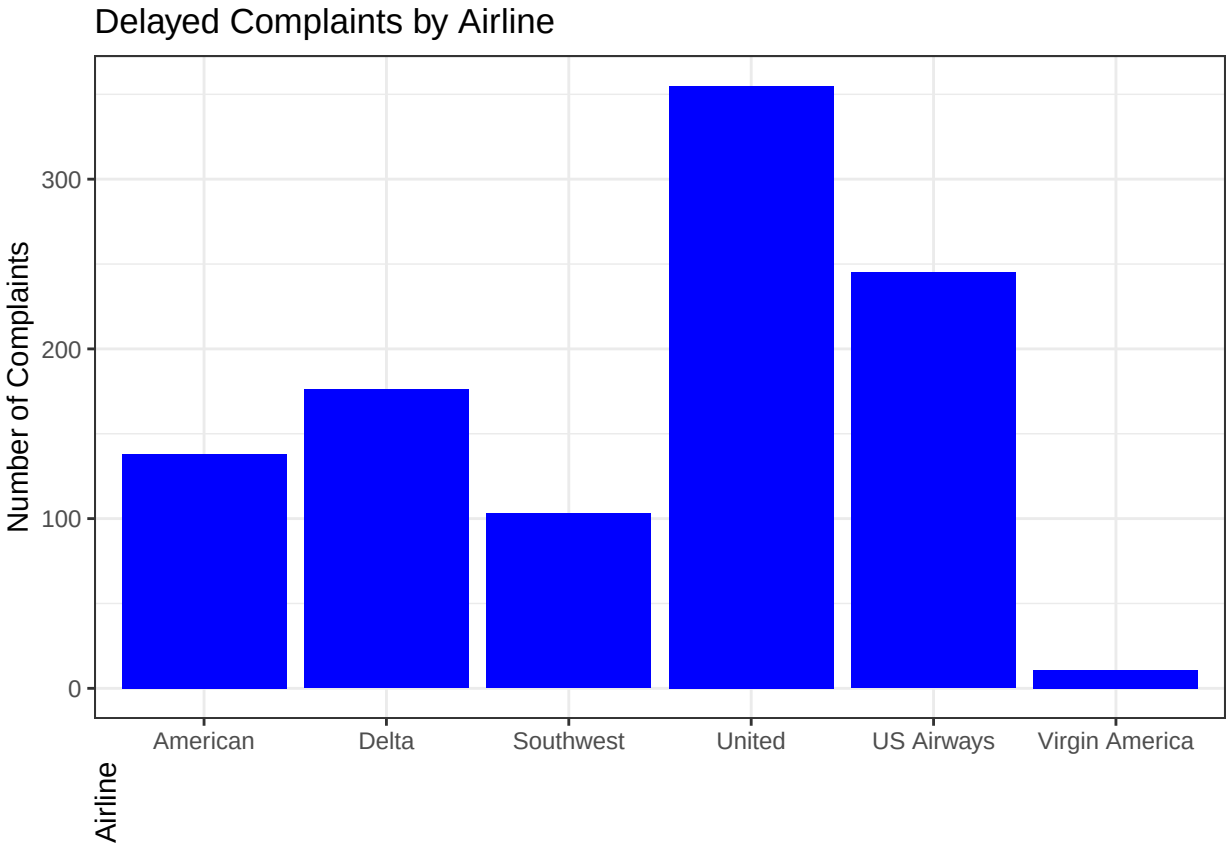
```
## Joining, by = "word"
```

```
## # A tibble: 15,093 x 2
##    word             n
##    <chr>        <int>
##  1 united        4153
##  2 flight        3922
##  3 usairways     3051
##  4 americanair   2962
##  5 southwestair  2457
##  6 jetblue       2364
##  7 t.co          1211
```

```
##  8 http          1153
##  9 cancelled     1064
## 10 service        963
## # ... with 15,083 more rows

## Joining, by = "word"
## Joining, by = "word"
```

The most frequently occuring words clearly refer to delayed flights. To see which airlines had the worst reputation with respect to tardiness, all words related to "delay" were converting to "delayed", and the tweets filtered for only those containing "delayed". These tweets were then groups by airline, and their frequency plotted, shown below. We see that United and US Airways have the worst reputation for tardiness.

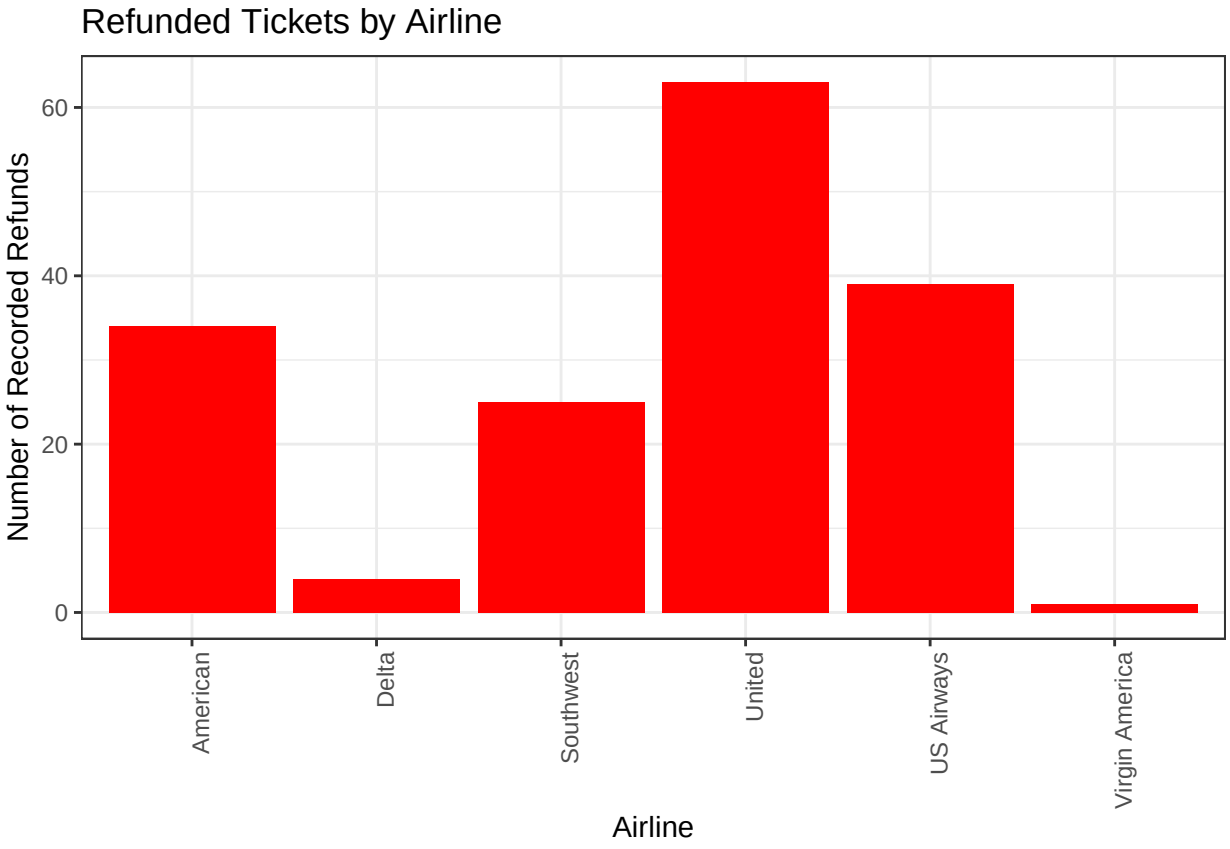


Positive Sentiments

Despite not having a column categorizing the reasons for positive sentiments, we also created a word cloud to show the most frequently occuring word in the positive sentiments. The results, shown below, reveal that the second most common reason for positive sentiments are in regards to refunds, which can only have been associated with an initial negaative experience.

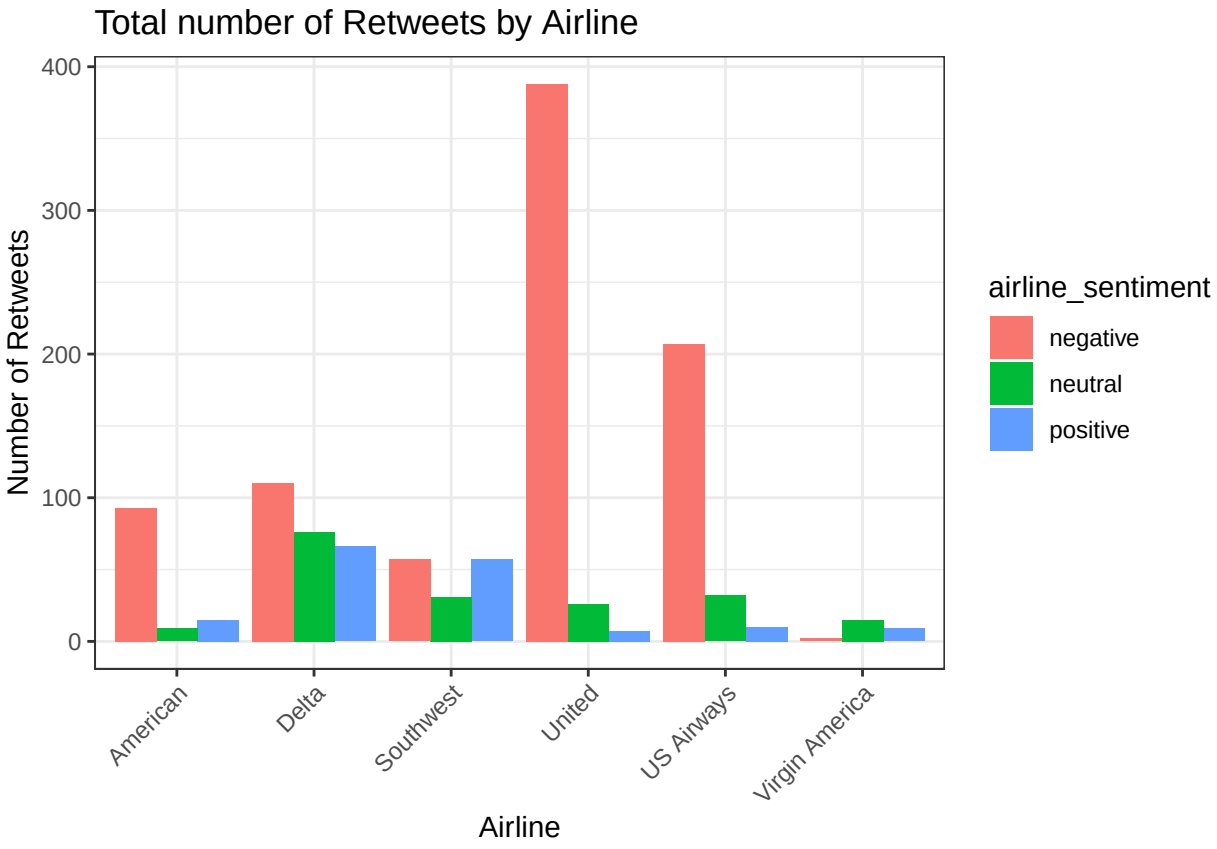
```
## Joining, by = "word"
## Joining, by = "word"
```

As was done with the "delayed", we wanted to see which airlines are most associated with the refunds. To do this, all words rooted with "refund" were converted to "refund", The tweets then filtered for those associated with refunds, grouped by airline and plotted below. Once again, United and US airlines, which are associated with most negative sentiments ultimately are also issuing the most refunds.

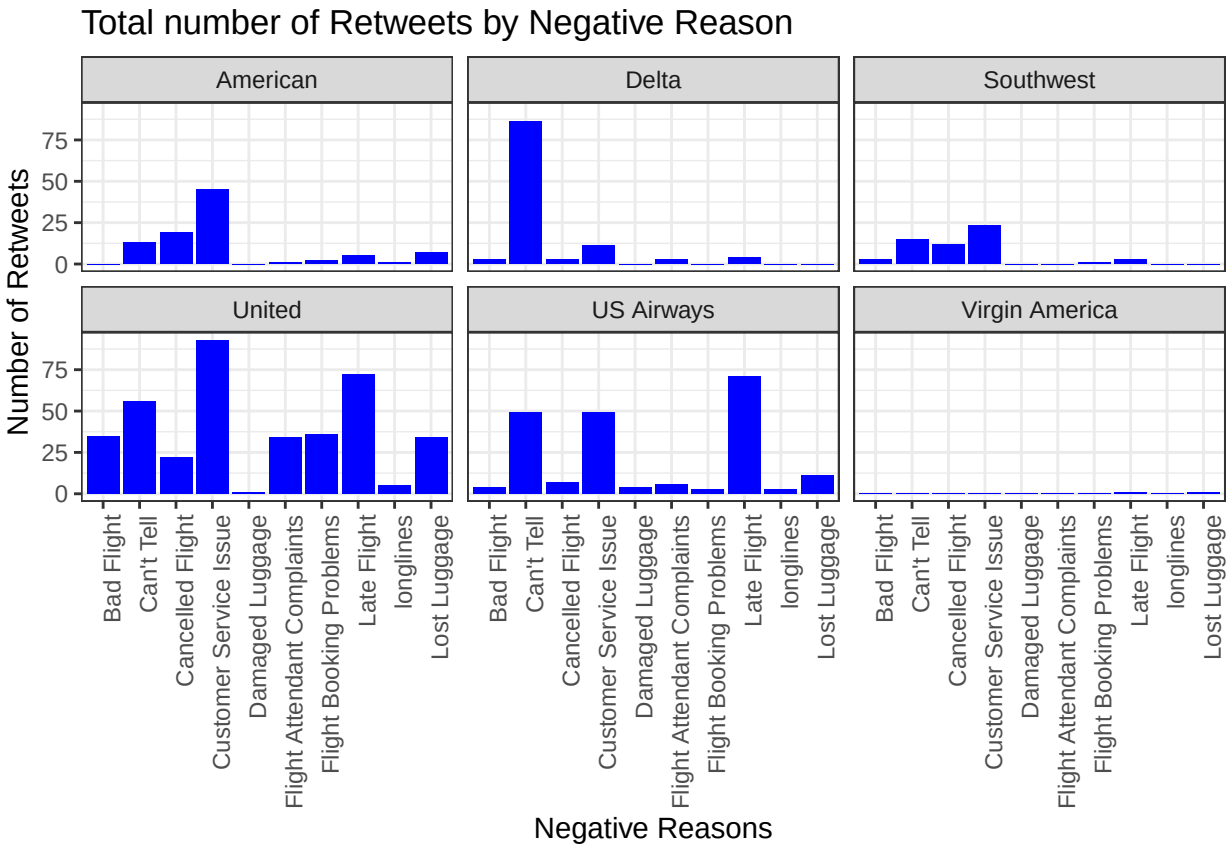


Retweet Revelations

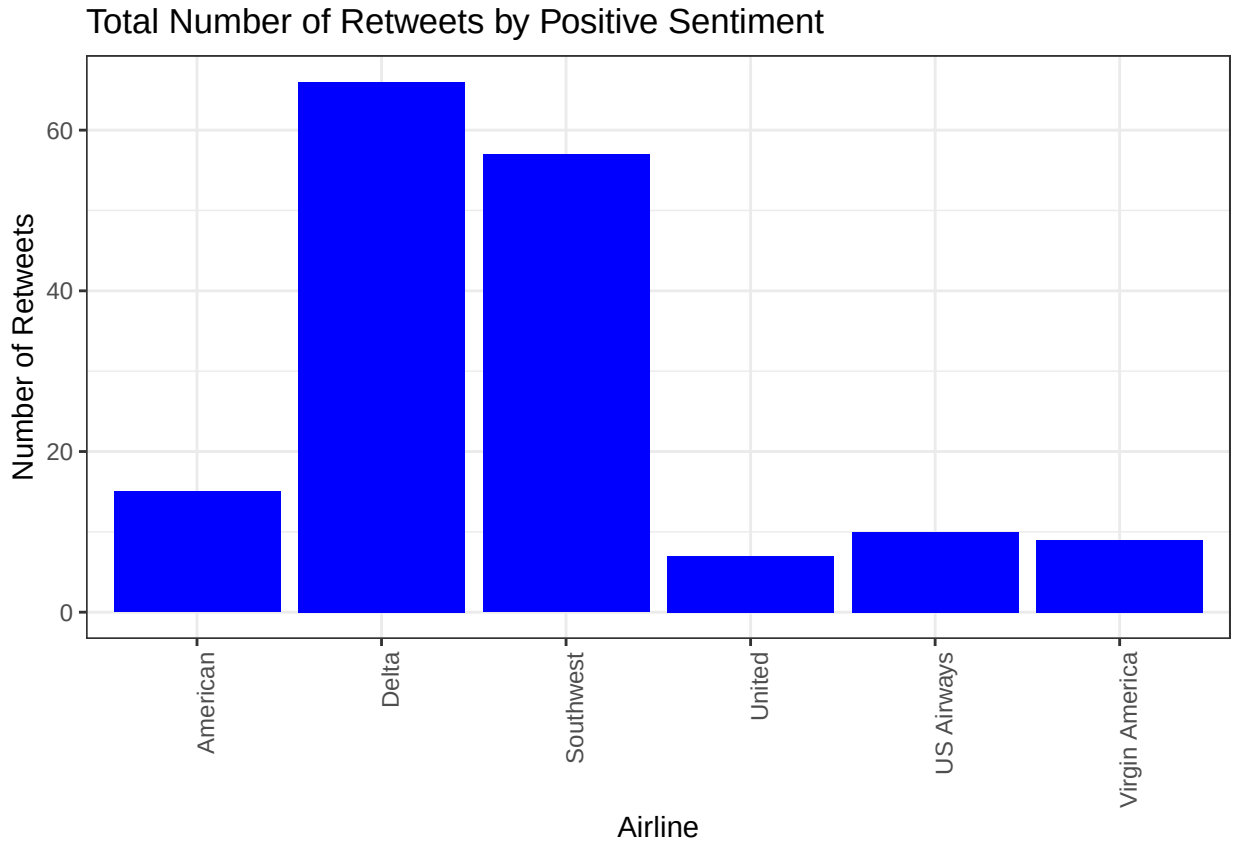
Counting the number of retweets can help to uncover more information about customer sentiments. May people do not comment on their experience, good or bad, but may express their sentiments by retweeting comments that reflect them. Below, we plotted retweet counts with respect to airline, negative reasons, and positive sentments. It can be suggested from the graph below that negative sentiments for United and US Airways are shared greatly in the wider Twitter community, several times that of the neutral and positive sentiments combined. This negativity is shared to a lesser extent by customers of American Airlines. While Delta has high number of negative sentiment retweets, the wider community overall appear to share neutral or positive sentiments. South Western and Virgin American on the other hand, appear to have a much more positive image amongst the wider community.



Taking a deeper look at the retweets within the negative sentiments, we can see where the wider Twitter community share their negative experiences. As previously seen, most negative tweets were related to customer service, and the graph below also reflects similar senitments, along with tardiness. It can be concluded that United has a poor reputation across all negative categories, which the other airlines can narrow down specific areas for improvement.



Exploring the number of retweets for positive sentiments can also expose which airlines resonate most positively with the wider Twitterverse. The graph below shows that the wider community share most positive sentiments with Delta and Southwest. When compared with number of negative retweets, this might suggest that these airlines do a far better job in meeting customer expectations when addressing issues.



Model Development and Validation

To contruct our model, we first tokenized all the tweets based on words, and extracted "tweet_id" column along with the (tokenized) "words" column. We then cleaned our list of tokenized words by removing digits, punctuation, stop words, and any remaining blank spaces. The data was then indexed and grouped by tweet_id. For each unique id, the individual keywords were spread across the dataframe by their value of importance. Some tweets hold only two or three keywords in a row, while others contain several more. Topic modelling using LDA requires the document to contain the key text in one row per id, after which the model will conditionally assign topics and extract terms which are correlated to those topics. Our tokenized words were then united, with all punctuation and digits removed in preparation for the model to extract key insights.

```
clean_tokens <- data
clean_tokens$word <- gsubfn('[[:digit:]]+', '', clean_tokens$word)
clean_tokens$word <- gsubfn('[[:punct:]]+', '', clean_tokens$word)
data("stop_words")

clean_tokens <- clean_tokens %>%
  filter(!(nchar(word) == 1))%>%
  anti_join(stop_words)
```

```
## Joining, by = "word"
```

```
data_tokens <- clean_tokens %>%
  filter(!(word==""))

data_tokens <- data_tokens %>%
  mutate(ind = row_number())

data_tokens <- data_tokens %>%
  group_by(tweet_id) %>%
  mutate(ind = row_number()) %>%
  tidyr::spread(key = ind, value = word)

data_tokens[is.na(data_tokens)] <- ""

data_tokens <- tidyr::unite(data_tokens, word,-tweet_id,sep =" " )
data_tokens$word <- trimws(data_tokens$word)
```

Using the now clean tokenized words, a Document Term Matrix of dgCMatrix-class was created, naming the documents based on tweet_id and using an ngram window up to 2. Upon checking the frequency of the 83209 individual terms, terms that occurred only once or appeared in more than half of all the documents were removed, as shown below.

```
#Document Term Matrix

dtm <- CreateDtm(data_tokens$word,
                 doc_names = data_tokens$tweet_id,
                 ngram_window = c(1, 2),
                 stopword_vec = c(stopwords::stopwords(language = "en",
                                                       source = "smart")))

#Term Frequency
```

```
tf <- TermDocFreq(dtm = dtm)

original_tf <- tf %>%
  select(term, term_freq, doc_freq)
rownames(original_tf) <- 1:nrow(original_tf)

#remove any words that appear <2 or in >half the doc

vocab <- tf$term[tf$term_freq >1 & tf$doc_freq < nrow(dtm)/2]

dtm = dtm
```

Latent Dirichlet Allocation (LDA) Model using the LDA Function

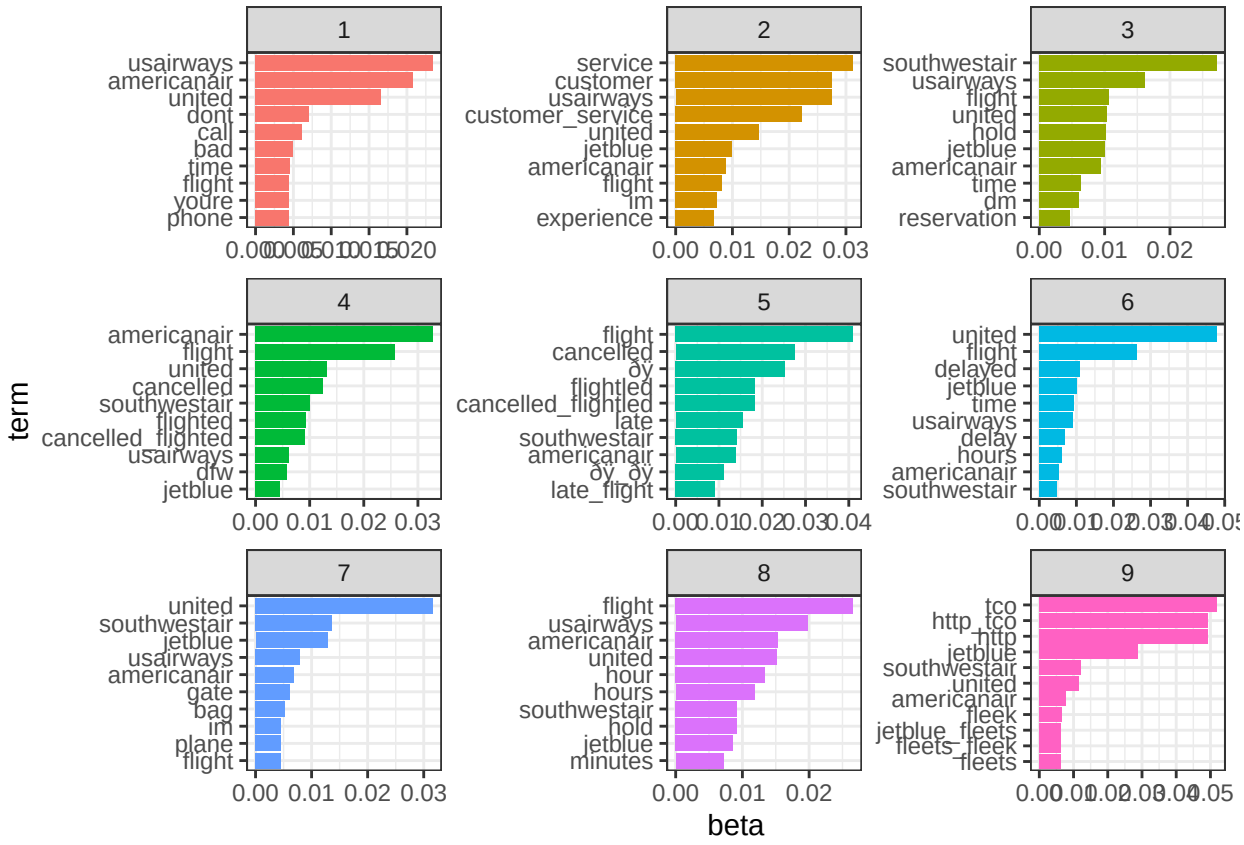
In order to find meaning within the tweets, we applied the Latent Dirichlet Allocation (LDA) model to determine 9 possible topics within each tweet document (k = 9). In the first run we used the LDA Function, which employs the VEM algorithm. This created a LDA_VEM topic model with 9 topics. The topic probabilities were extracted from the LDA into matrix format and organized in descending order of probability. The top 10 terms per topic are plotted below, where "beta" is the probablity.

```
#use Latent Dirichlet Allocation

data_lda <- LDA(dtm, k=9, control = list(seed = 1234))
#data_lda

#check per-topic per-word probabilites
data_topics <- tidy(data_lda, matrix = 'beta')
#data_topics

data_top_terms <- data_topics %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
```



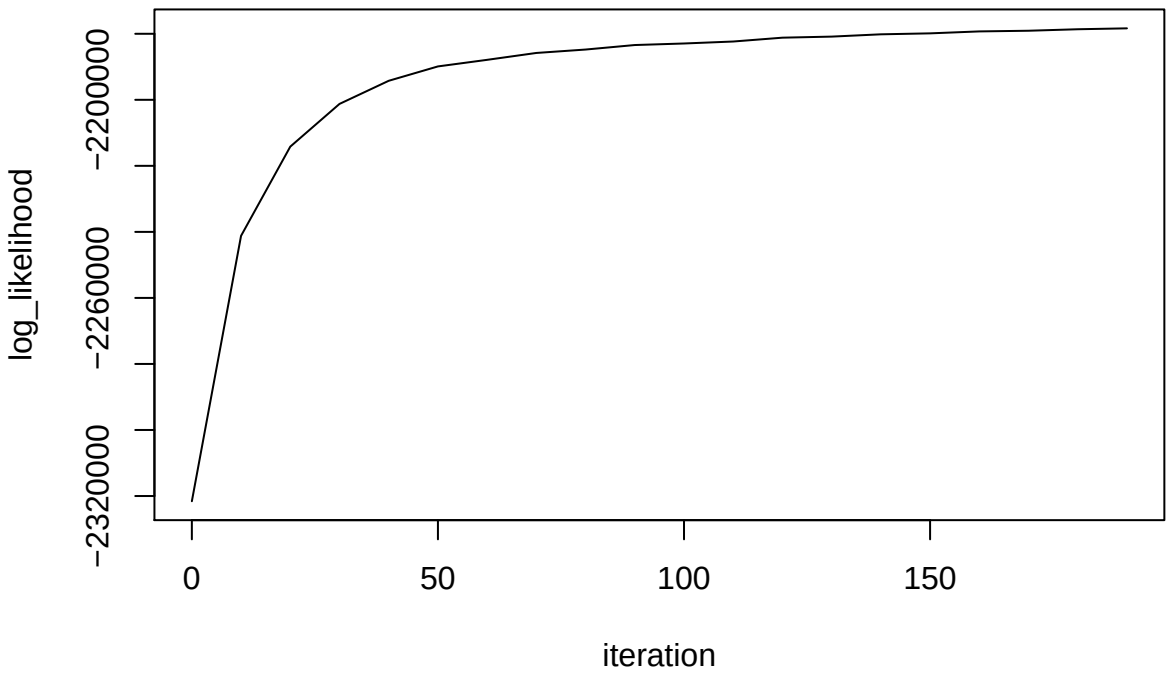
Latent Dirichlet Allocation (LDA) Model using the FitLdaModel Function

For the second run, the FitLdaModel Function was used to build the model. Both methods work well, but have different strengths: the previous method, which employs the VEM algorithm, requires less work and doesn't have as many calculations, while the method FitLdaModel function, which employs Gibb's Sampling, allows for the creation of topic labels easily using theta values. In the model below, alpha and beta control the distribution of topics and terms, where a high alpha and beta correspond to more terms and topics. To limit the number of topics and terms we use low values.

```
model <- FitLdaModel(dtm = dtm, k = 9, iterations = 200,
                     burnin = 180, alpha = 0.1, beta = 0.05,
                     optimize_alpha = TRUE, calc_likelihood = TRUE,
                     calc_coherence = TRUE, calc_r2 = TRUE, cpus = 2)
```

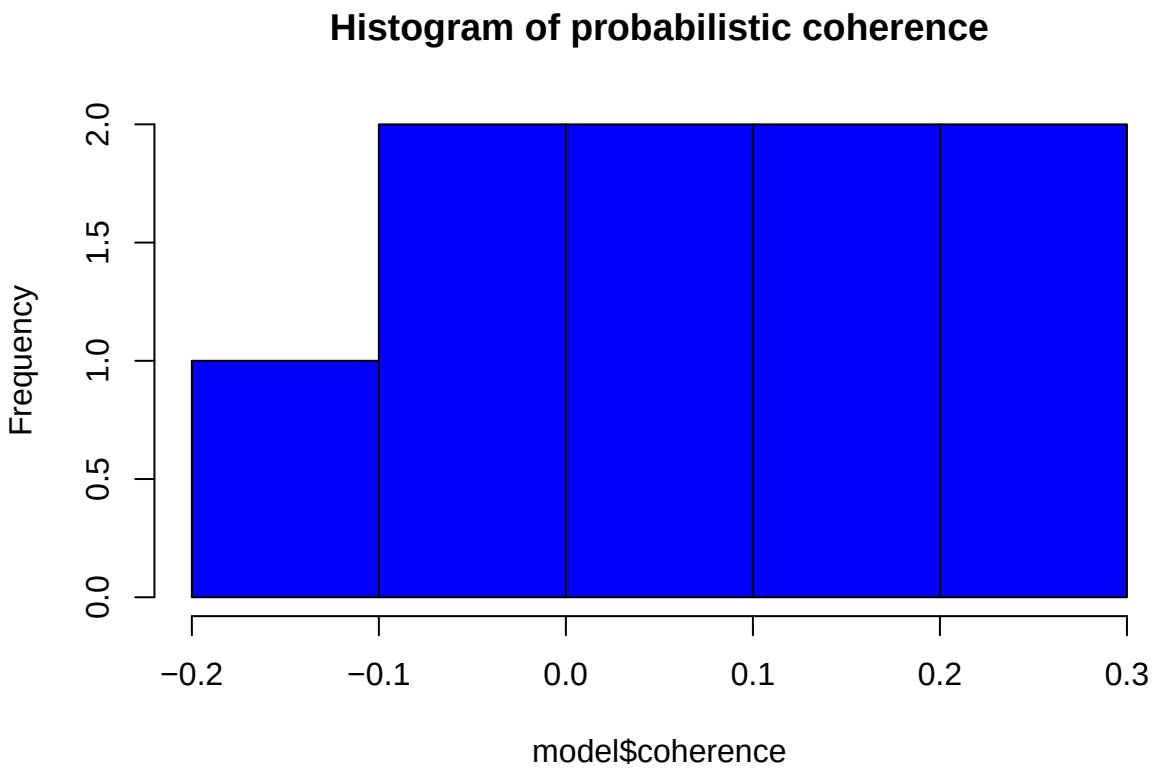
The model calculated R-squared of 0.02827366 signifying a low proportion of variability in terms. This suggests that the terms typically revolve around the same topics, which in the case of airline services is expected. We plot the log likelihood with respect to the number of iterations below. It can be seen that the log likelihood begins to stabilize after around 100 iterations.

```
## [1] 0.02734762
```



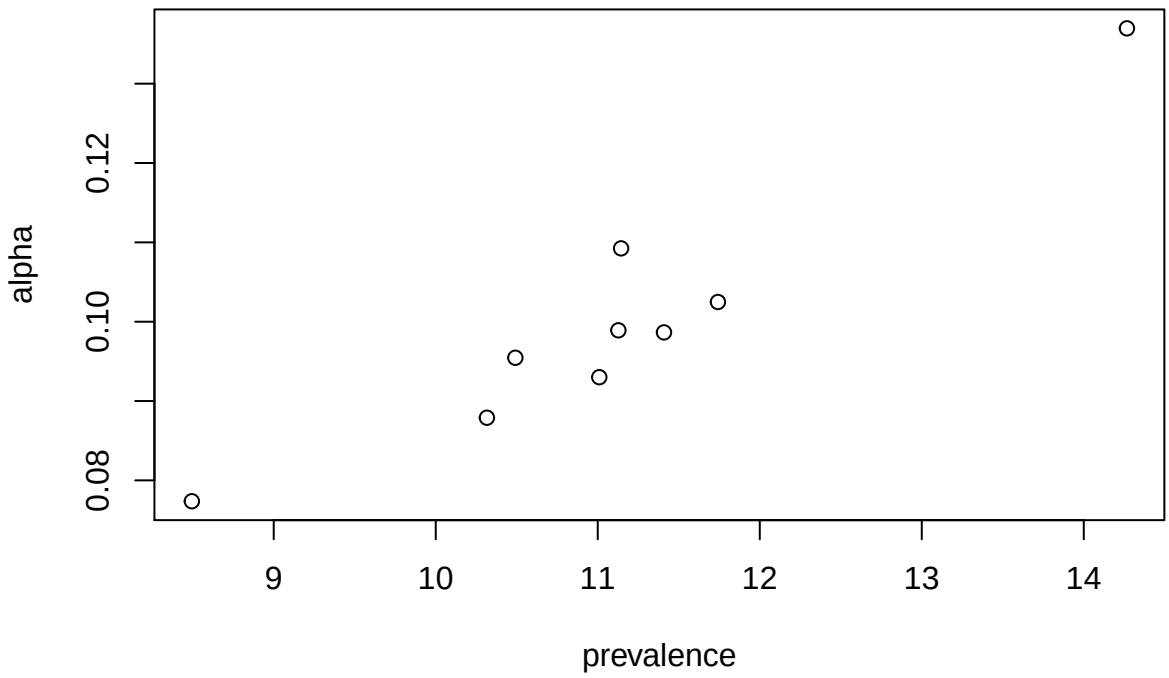
Below is a statistics summary and histogram of the term coherence.

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -0.104454 -0.005407  0.043026  0.077634  0.185140  0.269819
```



From the model, we then extracted the top 10 terms per topic and plotted the prevelence of each topic.

```
##      [,1]         [,2]           [,3]          [,4]                  [,5]
## t_1 "usairways"  "americanair"  "hold"        "call"                "phone"
## t_2 "flight"     "united"       "usairways"   "delayed"             "plane"
## t_3 "united"     "bag"          "lost"        "baggage"             "luggage"
## t_4 "jetblue"    "southwestair" "ðÿ"          "tco"                 "http"
## t_5 "flight"     "cancelled"    "americanair" "cancelled_flightled" "flightled"
## t_6 "united"     "dont"         "americanair" "usairways"          "im"
##      [,6]           [,7]           [,8]          [,9]
## t_1 "hours"        "ive"          "wait"        "hour"
## t_2 "late"         "gate"         "jetblue"     "delay"
## t_3 "usairways"    "bags"         "americanair" "seat"
## t_4 "http_tco"     "ðÿ_ðÿ"        "love"        "virginamerica"
## t_5 "southwestair" "flights"      "usairways"   "tomorrow"
## t_6 "jetblue"      "southwestair" "fly"         "home"
##      [,10]
## t_1 "southwestair"
## t_2 "hours"
## t_3 "check"
## t_4 "fleek"
## t_5 "flight_cancelled"
## t_6 "email"
```

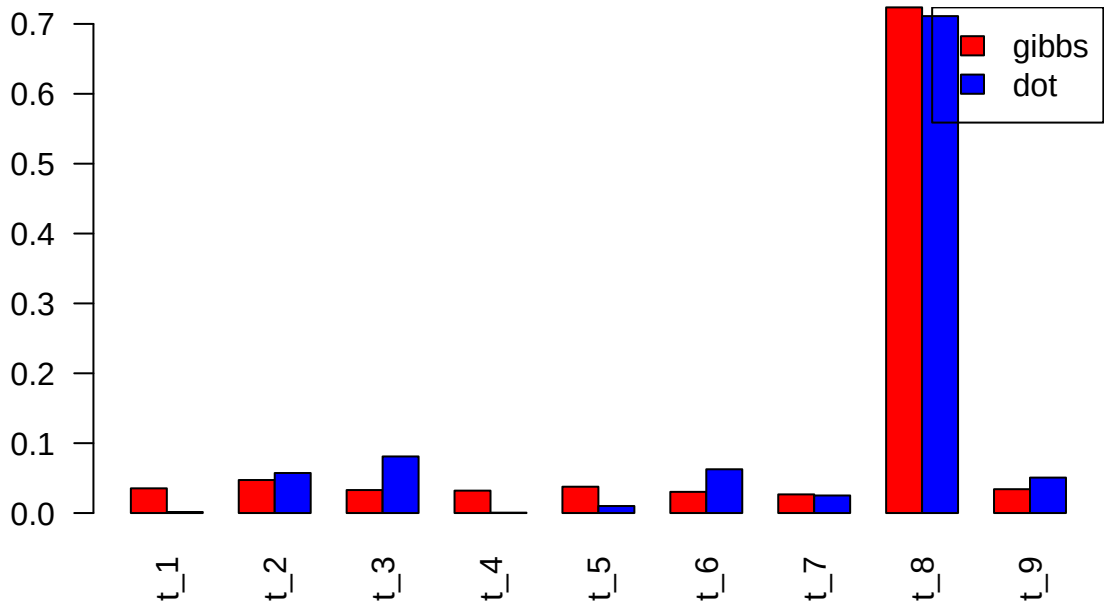


From here we labeled our model topics using theta values greater than 0.05, and combined the topics, coherence score, prevelance and top terms into one dataframe, displayed below in html format.

```
## # A tibble: 9 x 5
##   topic label_1       coherence prevalence top_terms
##   <chr> <chr>             <dbl>      <dbl> <chr>
## 1 t_1   hold_hours        0.017       11.7  usairways, americanair, hold, call, ~
## 2 t_2   late_flight      -0.005       14.3  flight, united, usairways, delayed, ~
## 3 t_3   baggage_claim     0.043       10.5  united, bag, lost, baggage, luggage,~
## 4 t_4   http_tco          0.106       11.0  jetblue, southwestair, ðÿ, tco, http~
## 5 t_5   cancelled_fl~     0.233       11.1  flight, cancelled, americanair, canc~
## 6 t_6   united_im        -0.045       10.3  united, dont, americanair, usairways~
## 7 t_7   http_tco          0.27         8.49 tco, http_tco, http, jetblue, united~
## 8 t_8   flight_booki~    -0.104       11.4  united, southwestair, flight, americ~
## 9 t_9   customer_ser~     0.185       11.1  service, customer, united, customer_~
```

Model Validation

Once both models were created, we compared the two methods ability to accurately predict topics. Noting that Gibb's method uses conditional probability to decide topics, while the dot method calculates dot-product probability of topics, the plot below shows which method works best per topic. Overall, while each method have their own pros and cons, and Gibb's method is slower, we ultimately chose the Gibb's prediction model.



Implementation in Shiny

We created an app to display the results of our analysis in a more easily interprettable way for users. The app, named the "Twitter Analytical Dashboard for Airlines", can be found at https://stan-t.shinyapps.io/shiny/, and its functionality is described below.

Users are prompted to select an airline whose tweets they wish to analyse. Upon pressing the "Go" button, all tweet related to the selected airline are run through the analytical model described previously. Once processed 4 plots are diplayed in the main panel: a word cluster, a word frequency bar graph, our topic model, and a word cluster dendrogram. Using the sliders, users may control the minimum frequency of the words they wish to see in the word cloud, as well as the maximum number of words to display, therefore determining how populated the word cloud appears, while the word frequency bar graph displays the entire list of relevant words extracted from the tweets. By looking at these words, airlines can get a sense of what they customers are talking about the most, though it does not reveal the senitments around those words.

In order to gauge sentiments, users can look at the Topic Model graphs and the Cluster Dendrogram, which give the user a sense of which word combinations or patterns often occur together. The number of clusters in the appearing in the cluster dendrogram can be controlled by the third slider. Note that adjusting the

sliders automatically updated the word cloud and cluster plots. The "Go" button only controls the input data to be analyzed, therefore only one airline may be analyzed at a time.

Limitations

Twitter data inherently comes with its own limitations. While text mining social media allows companies to gain a better understanding of client sentiments, Twitter's limited character allowance leaves little room for much to be gained beyond the client creating a thread. Even in those instances, threads are more likely to contain negative sentiments regarding customer experience. As the tweets we mined were from the era of Twitter's 140 character limit, it is possible that more recent data would allow companies to gain better insights. Of course, with a character limit, and with the focus of Twitter being a platform to allow customers to express their thoughts in limited words, the insights gained tend to be either positive or negative, with little room for nuance. In such a limited setting, customers would more likely use the platform to air their complaints, leading to an increasingly higher number of negative reviews from such platforms. We cannot discredit Twitter data because of the limited characters; when we see words revolving around customer service such as in topics 4 and 5, we can see key terms that could show an area airlines need improvement on, or they could find it's an area they are doing well in.

In our dataset, we had 15 variables; tweet_id, airline_sentiment, airline_sentiment_confidence, negativereason, negativereason_confidence, airline, airline_sentiment_gold, name, negativereason_gold, retweet_count, text, tweet_coord, tweet_created, tweet_location, user_timezone. Twitter users are allowed, for privacy concerns and other reasons, to decide whether Twitter is allowed to use their location or coordinates. In addition, Twitter allows them to specify their locations in their own words. A user could input their location as "Somewhere over the rainbow", for example. With this option available to users, companies seeking to understand whether there are trends in airline sentiment based on location will find it very difficult to do so. In cases such as those, perhaps airline reviews on other sites would provide better insights.

Conclusion: Insights/How Insights can be used to improve business

Our goal for this project was to conduct text analysis using natural language processing on a series of tweets pertaining to six major US airlines, that could be used to generate insights that these companies could potentially use to improve the quality of their services. We implemented two methods to build a Latent Dirichlet Allocation (LDA) model, used to predict possilbe topics within each tweet document, for comparison. Our secondary model, using the textmineR FitLdaModel function, allowed us to get a closer look at topic labels and pull together a table with labelled topics and terms. We ultimately found similar results as previously studies showed, that customers main grievances were related to poor customer service and flight tardiness (or cancellation). We found that due to the nature of the platform, where people can express themselves in real time, and the character limitations at the time the data was scraped, customer were more likely to express dissatisfaction as opposed to positive experiences, therefore making Twitter an unreliable source to obtain an accurate representation of an airline's overall performance. It is, however, an excellent means by which issues can be handled in real time, making it a means by which to garner some insights into how areas such as customer service can be improved. For instance, if numerous customers complain that airline representatives were rude, then an airline may address this issue by retraining staff in communication and conflict resolution. That being said, if an airline wishes to obtain an accurate overview of performance, they would benefit more greatly to rely on other platforms specifically designed to gauge customer satisfaction.