US AIRLINE TWEETS SENTIMENT ANALYSIS

IS- 688: WEB MINING (FALL 2018)

SUBMITTED TO
PROF. LIN LIN

DECEMBER 18, 2018

GROUP 1

TEAM MEMBERS:

WILFRED CASTELINO
KETKI ANDHALE
ABHIMANYU RAJA
DWIJA VYAS
PRADNYA AMBRE

Table of Contents

1. INTRODUCTION 2
- **1.1. Abstract** 2
- **1.2. Problem Statement** 2
- **1.3. About the dataset** 3
- **1.4. Challenges** 3
- **1.5. Why Sentiment Analysis in R?** 4
- **1.6. Packages used in our project** 4

2. TEXT MINING 5
- **2.1. Data preprocessing** 5
- **2.2. Our Approach** 5

3. EXPLORATORY DATA ANALYSIS 9
- **3.1. Distribution of tweets** 9
- **3.2. Distribution of Negative Tweets based on date and time** 11
- **3.3. Reasons for Negative Tweets for each airline** 13
- **3.4. Word Cloud** 15

4. SENTIMENT ANALYSIS 19
- **4.1. What is Sentiment Analysis?** 19
- **4.2. Sentiment Score for all tweets** 20
- **4.3. Sentiment Score for each airline** 22

5. CONCLUSION 24

REFERENCES 25

1. INTRODUCTION

1.1. Abstract

This report summarizes the sentiment analysis of nine days of data from Twitter regarding customer sentiment for 6 US airlines: American, Delta, Southwest Airlines, United, US Airways, and Virgin America. From this data we evaluate how customer sentiment evolves over a week. We also evaluate customers tweet sentiment by airline. Additionally, we find that the likelihood of a tweet being retweeted is dependent on the sentiment and airline. Negative tweets about United, American, and US Airways are more likely to be retweeted than positive tweets. In contrast, positive tweets about Southwest, Virgin America, and Delta are more likely to be retweeted than negative tweets. We hypothesize that public opinion is driving this bias.

These are the steps we followed in our project:

1. Data preprocessing (Text Mining)
2. Analysis based on tweets distribution and airlines
3. Reasoning and assumptions
4. Word cloud
5. Calculate the Sentiment score.

1.2. Problem Statement

The goal of every airlines is to help serve their customer with the best possible service they could offer. Customer satisfaction is one of the important aspects of their goal which helps them grow their business.

However, with the rise of social media and use of technology, the communication between the customer and the airlines is more flexible and convenient for the customer. Twitter is a useful forum where people can raise their concerns and queries or share their experience that would help others who want to travel with the same airlines. Also, for instance if a new airline company want to step into the US market and want to analyze based on past experiences of customers to launch their own airline.

To get these answers, one must analyze the previous data in hand. Sentiment Analysis is one of the approaches where one use to get responses based on the sentiments of the tweets of the customer. In this project, we have done a thorough exploratory data analysis of the sample of tweets in a dataset that would help us build strategies and make decisions accordingly.

1.3. About the dataset

Our dataset was about the US Airline tweets. That is, to analyze how travelers in February 2015 expressed their feelings on Twitter. The six major US airlines we considered were Delta, Virgin America, United, Southwest, US Airways and American.

SOURCE	ABOUT	TWEETS	SENTIMENTS	AIRLINES
KAGGLE	AIRLINE TWEETS	14640	POSITIVE,NEGATIVE & NEUTRAL	6

A sentiment analysis job about the problems of each major U.S. airlines. There were about 14640 tweets. Twitter data was scraped from February of 2015 and first we classified the positive, negative, and neutral tweets, followed by categorizing negative reasons (such as "late flight" or "rude service"). It contains whether the sentiment of the tweets in this set was positive, neutral, or negative for the six US airlines.

Some of the major attributes of the dataset were:

- TWEET ID's
- TWEETS
- AIRLINE SENTIMENTS
- LOCATION
- NEGATIVE REASONS
- DATE AND TIME

1.4. Challenges

- No data on customer complaints that might be related to certain airports.
- No data on the price and category of airlines.
- No data on the time of travel
- Multiple languages, Redundant data (external links), punctuation etc.
- Abbreviations
- Hashtags
- Slang

1.5. Why Sentiment Analysis in R?

As we see that we have in hand challenges with our dataset relating to unwanted and redundant data, this makes it difficult to analyze the data. R- Studio provides in built packages that we can use, and it makes our work easier.

There are a variety of dictionaries in R that exist for evaluating the opinion or emotion in text. The lexicons are based on unigrams (or single words). These lexicons contain many English words and the words are assigned scores for positive/negative sentiment, and possibly emotions like joy, anger, sadness, and so forth. There are different packages that show the sentiments. In our project we have used the NRC sentiment from Saif Mohammad and Peter Turney.

1.6. Packages used in our project

Packages	Description
dplyr	Helps in data manipulation and to transform summarize tabular data in rows and columns
ggplot2	To create graphs using grammar of graphics
Readr	To read different types of data example csv
reshape2	To transform the data in expected structure
wordcloud	To plot a cloud of words across collective documents
syuzhet	To extract sentiment and sentiment driven plot arcs from text
lubridate	To work with date-times and time-spans
Tm	Framework for text mining applications in R

2. TEXT MINING

2.1. Data preprocessing

Real world data can be extracted from various resources to name few examples, Facebook, twitter, Flickr, blogs, boards, YouTube etc. This real-world data is raw data or unprocessed data and is simply collection of numbers, character. The raw data is can be summarized as:

1) **Incomplete:** lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
2) **Noisy:** containing errors or outliers
3) **Inconsistent:** containing discrepancies in codes or names

This data needs to be processed or cleaned and needs to be structured for further analysis and study. Data processing is a vast field and includes various tasks such as filling in missing values, smooth noisy data, identifying or removing outliers, and resolving inconsistencies, normalization and aggregation, reducing the volume but producing the same or similar analytical results, data reduction, replacing numerical attributes with nominal ones.

2.2. Our Approach

We have important libraries such as **dplyr, tm** from RStudio for text mining that we used throughout the project and primarily focused on text cleaning and stemming for our data set.

Text Cleaning mainly include:

- Removing null and missing values from the existing data in excel.
- Data Transformation (unstructured data into structured data).
- Eliminate punctuation, numbers, unnecessary spaces and stop words.

Below are the steps we performed for data preprocessing:

1) Import libraries and read csv file

```
#############################################################################
#                GROUP1: WEB  MINING - FINAL PROJECT
#############################################################################

##IMPORT LIBRARIES
library(readr)
library(tm)

##READ CSV FILE
airlineTweets <- read.csv(file.choose(), header = T)

#Display Info of the dataframe
str(airlineTweets)
negativeSubset = subset(airlineTweets, airline_sentiment == "negative")
str(negativeSubset)

```

2) Building a corpus and cleaning the text

```
#Show observations in columns
head(airlineTweets)

##Create a DataFrame of Negative Sentiments
##BUILD CORPUS AND CLEAN THE TEXT
corpus = Corpus(VectorSource(negativeSubset$text))
inspect(corpus[1:5])

corpus <- tm_map(corpus, stripWhitespace)
#strwrap(corpus[[8]])
inspect(corpus[1:5])

corpus <- tm_map(corpus, content_transformer(tolower))
inspect(corpus[1:5])
```

3) Removing stop words

```
#Remove stopwords
corpus = tm_map(corpus, removeWords, stopwords("english"))
inspect(corpus[1:5])

# Want to get rid of @virginamerica, @usairways, @united, @southwestair, @jetblue, @americanair
corpus <- tm_map(corpus, function(x) gsub('@', 'ReplacedText', x))
inspect(corpus[1:5])

corpus = tm_map(corpus, removeWords, c("ReplacedTextvirginamerica","ReplacedTextusairways","ReplacedTextunited","ReplacedTextsouthwestair'
inspect(corpus[1:5])

# We want to keep #
corpus <- tm_map(corpus, function(x) gsub('#', 'KeepHashPunct', x))
inspect(corpus[1:5])
```

4) Removing Punctuations

```
# Remove remaining punctuation
corpus <- tm_map(corpus, removePunctuation)
inspect(corpus[1:5])

```

5) Removing external links

```
# Remove all http links
corpus <- tm_map(corpus, function(x) gsub('http[[:alnum:]]*', '', x))
inspect(corpus[1:5])
```

6) Stemming the document

```
# Stem the document
corpus = tm_map(corpus, stemDocument)
cleanset = tm_map(corpus, PlainTextDocument)
inspect(cleanset[1:5])
```

After performing all these steps on the raw data, we finally got the structured data. For example,

Content (before)	Content (after)
@VirginAmerica it's really aggressive to blast obnoxious "entertainment" in your guests' faces & they have little recourse #fabulous	its really aggressive to blast obnoxious entertainment guests faces amp have little recourse fabulous

3. EXPLORATORY DATA ANALYSIS

3.1. Distribution of tweets

```
##---GROUP BY AIRLINE SENTIMENT------------##
posNegDist = airlineTweets %>% group_by(airline_sentiment) %>% dplyr::summarise(count = n())
posNegDist

##PLOT THE GRAPH
overallSentiment = as.data.frame(table(airlineTweets$airline_sentiment))
colnames(overallSentiment) = c("Sentiment","Freq")
histPlot = ggplot(overallSentiment) + aes(x=Sentiment, y=Freq, fill=Sentiment) + scale_fill_manual(values=c("#d03501", "#E69F00", "#009E7
histPlot = histPlot + geom_bar(stat="identity")
histPlot
```

In the below bar graph, the x axis has different category of sentiment and y axis has the frequency of occurrence of the tweets.

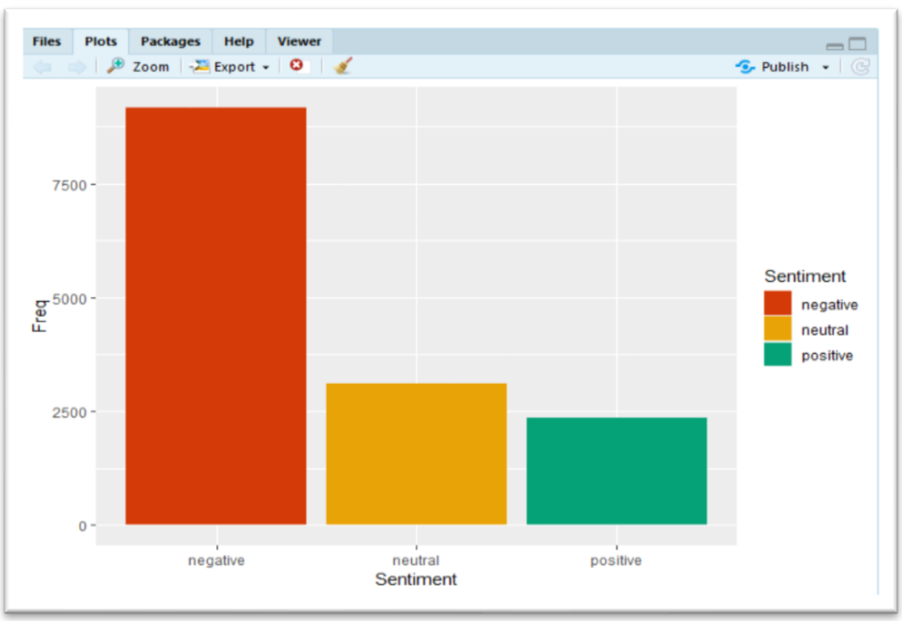


Fig 3.1.1. Distribution of Tweets based on category

We further wanted to study the frequency of each category of tweets based on airlines. So, the second graph we plotted was the distribution of tweets based on category and airlines. For this we plotted airlines names on x axis and the count of occurrences on y axis. The figure below is the representation of the graph distribution of tweets based on sentiment category and airlines.

```
##----2. DISTRIBUTION BY AIRLINE-----------------##
ggplot(airlineTweets, aes(airlineTweets$airline, fill=airline_sentiment)) + geom_bar()+ scale_fill_manual(values=c("#d03501", "#E69F00", "#009E73

airlines= airlineTweets %>% group_by(airline) %>% dplyr::summarise(count=n())
posNegByAirline <-dcast(airlineTweets, airline ~ airline_sentiment)
posNegByAirline$negPer = posNegByAirline$negative / (posNegByAirline$negative + posNegByAirline$positive + posNegByAirline$neutral)
posNegByAirline = posNegByAirline[order(-posNegByAirline$negPer),]
posNegByAirline
```

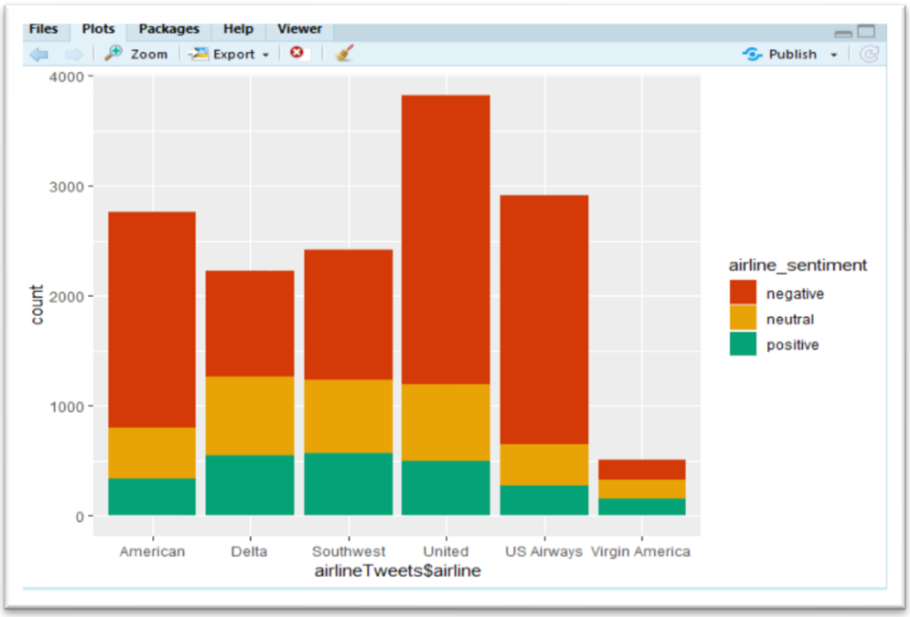


Fig 3.1.2. Distribution of tweets based on sentiment category and airlines

Observation: From the above graphs we can analyze that the negative sentiment has the highest occurrence of tweets which is followed by neutral tweets and minimum count is for positive tweets.

3.2. Distribution of Negative Tweets based on date and time

Once we evaluated that the maximum count was for negative sentiments, we plotted a line graph to understand the distribution of negative tweets based on dates. The dataset has the data for tweets from dates 16 February 2015 to 25 February 2015. For line graph, we plotted dates on x axis and count on y axis.

```
negativeTweets <- airlineTweets %>% filter(airline_sentiment=="negative")
negativeTweetsByDate <- negativeTweets %>% group_by(date) %>% dplyr::summarise(count = n())
negativeTweetsByDatePlot = ggplot() + geom_line(data=negativeTweetsByDate, aes(x=date, y=count, group = 1))+ scale_fill_manual(values=c("#d03501"
negativeTweetsByDatePlot
```

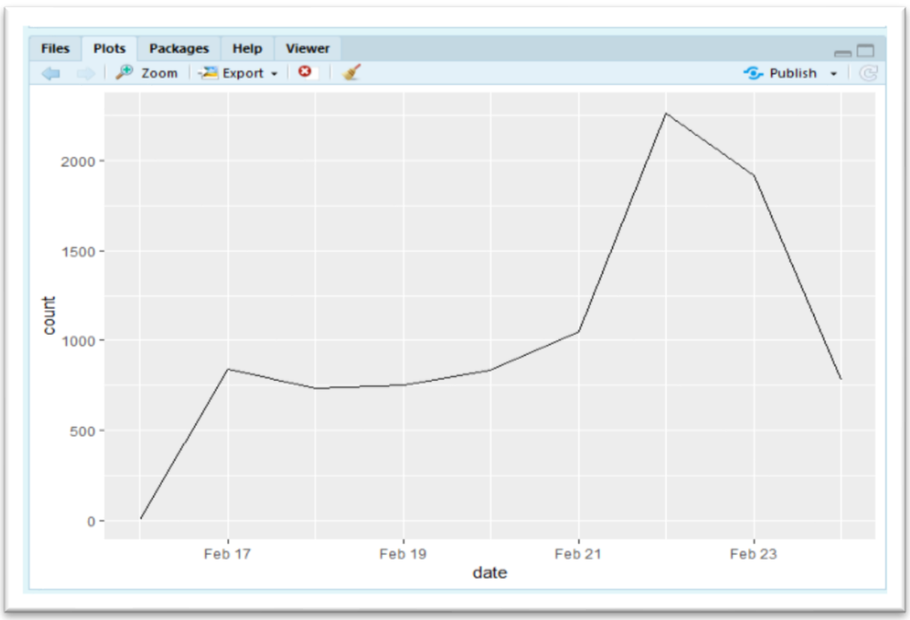


Fig 3.2.1. Distribution of negative tweets based on dates

Observation: From this graph we can evaluate that the date 22 February has the maximum count of negative tweets.

The below graph shows the count of different negative reasons that are listed over the 24-hour time. The X-axis shows the time at which the tweets were sent, and y-axis shows the count of the negative tweets.

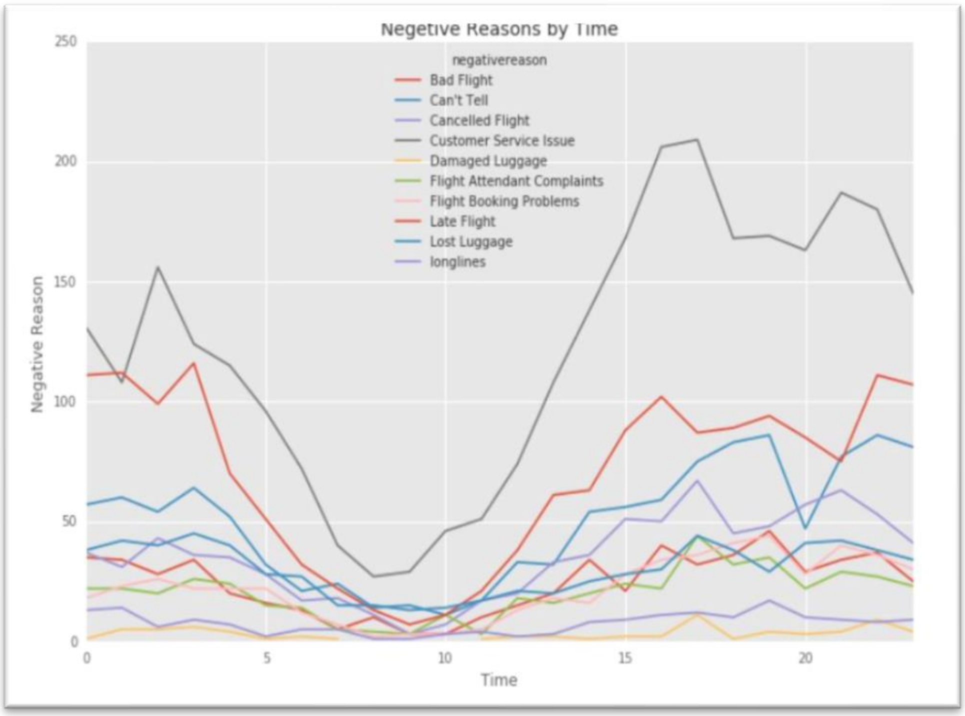


Fig 3.2.2 Negative Reasons by time

Observations:

- From the above graph, we see that the number of Customer Service Issues are more as compared to other negative reasons.
- The maximum number of customer service issue tweets are observed in the evening around 3 pm.
- After 3 pm till midnight we see a considerable number of tweets with negative reasons relating to customer service issues and bad flights.
- The least number of tweets are observed between 5 am to 10 am in the morning.
- Very few customers have damaged luggage issues which indicates that all these 6 airlines handle the luggage with care.

3.3. Reasons for Negative Tweets for each airline

After analyzing the airline dataset, we found out that there were reasons just for the negative tweets and positive and neural tweets didn't have any reasons. So, we plotted each airline against the proportion of negative reasons to figure out what resulted in the negative tweets by the customers.

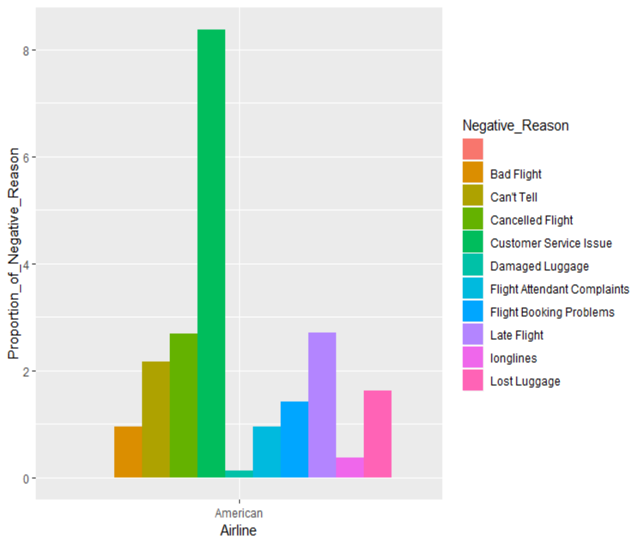


Fig 3.3.1: American Airlines

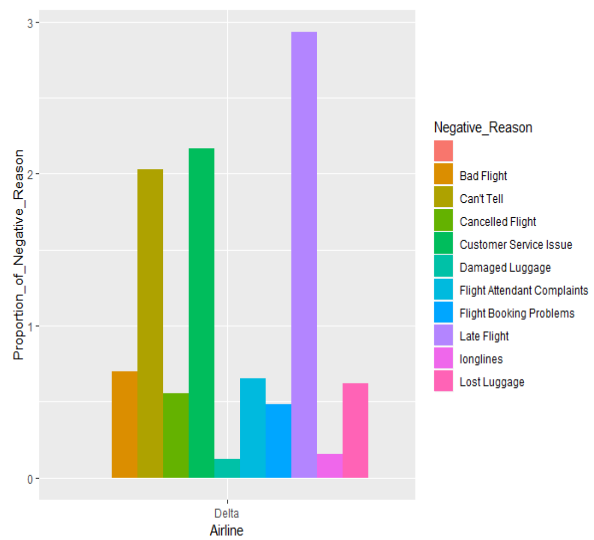


Fig 3.3.2: Delta Airlines

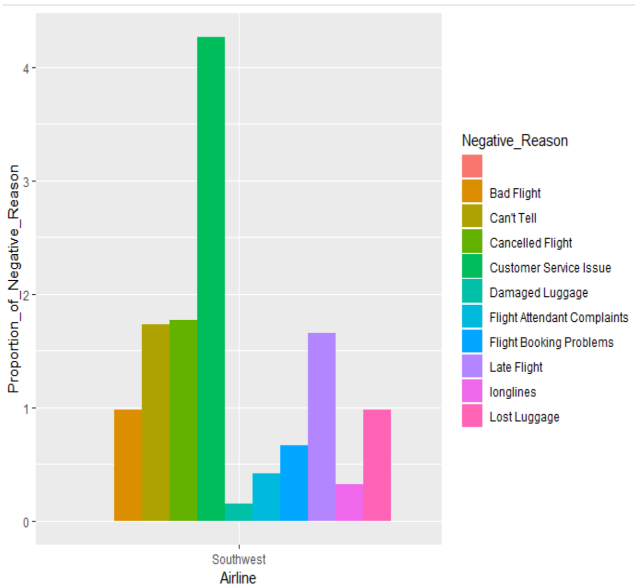


Fig 3.3.3: Southwest Airlines

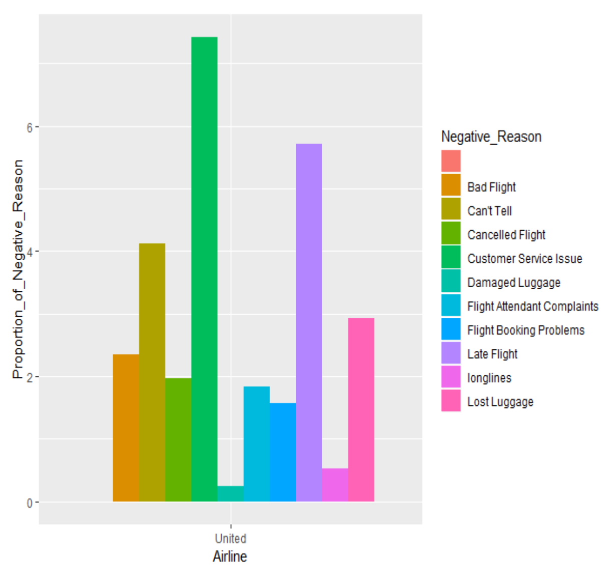


Fig 3.3.4: United Airlines

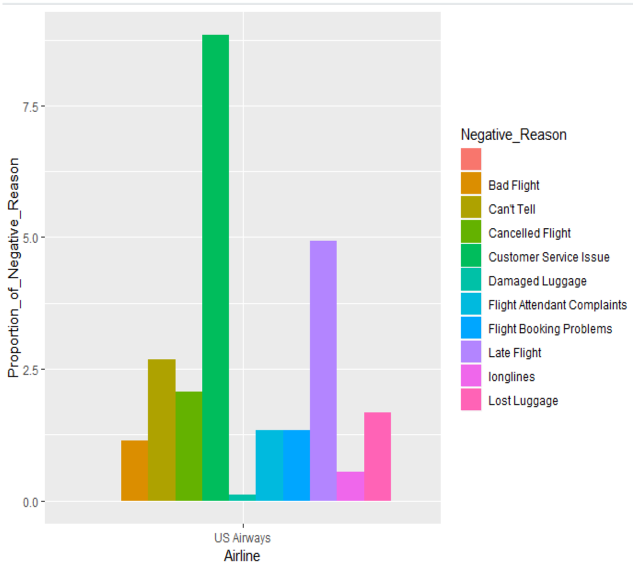


Fig 3.3.5: US Airways

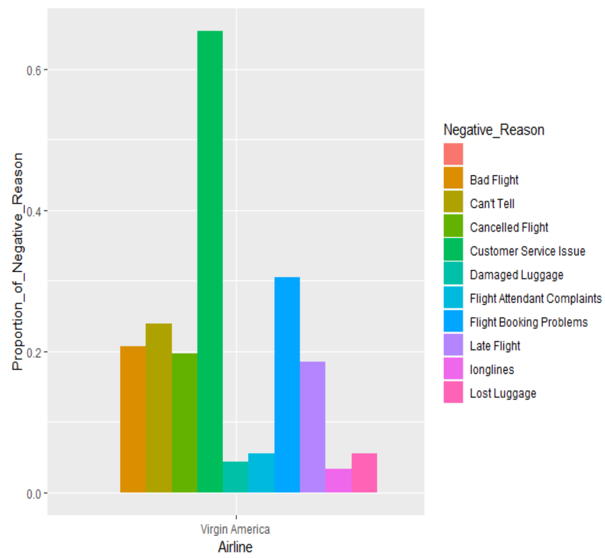


Fig 3.3.6: Virgin America

Observations:

We can see that American Airlines, Southwest Airlines, United Airlines, US Airways and Virgin America got negative reviews mostly because of their customer service issue. Only Delta Airlines received negative tweets due to their late flight scheduling issue.

However, since the data is spread out over a period of 9 days, it would be difficult to blame the airlines completely for the negative tweets received. After analyzing some of the negative reviews we could estimate that the negative tweets for poor customer service could be because of the following reasons-

- Management issues
- Planning issues
- Lack of preparation
- Catering issues
- Lack of training
- Incentives/salary issues

Similarly, the reasons for the delay in flights for Delta Airways could be due to-

- Inclement weather conditions
- Mechanical delays
- Wait for connecting flight passengers, bags
- Loading bags and cargo
- Airport congestion
- Catering
- Aircraft preparation

3.4. Word Cloud

Text mining methods allow us to highlight the most frequently used keywords in a paragraph of texts. One can create a **word cloud**, also referred as *text cloud* or *tag cloud*, which is a visual representation of text data.

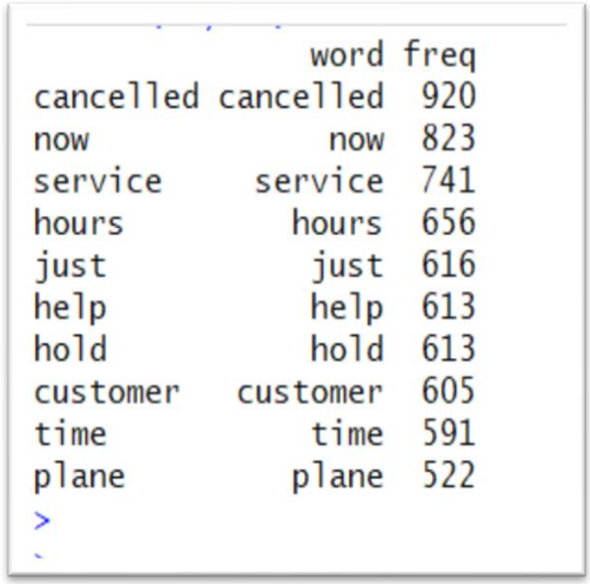
```
ıegativeTweets <- airlineTweets %>% filter(airline_sentiment=="negative")
locs2 <- Corpus(VectorSource(negativeTweets$text))
locs2 <- tm_map(docs2, content_transformer(tolower))
locs2 <- tm_map(docs2, removeNumbers)
locs2 <- tm_map(docs2, removeWords, stopwords("english"))
locs2 <- tm_map(docs2, removeWords, c("usairways" ,"united", "flight" , "americanair" ,
                                      "jetblue" , "southwestair","get","can"))
locs2 <- tm_map(docs2, removePunctuation)
locs2 <- tm_map(docs2, stripWhitespace)
```

The text mining package (*tm*) and the word cloud generator package (*word cloud*) are available in R for helping us to analyze texts and to quickly visualize the keywords as a word cloud.

```
dtm2 <- TermDocumentMatrix(docs2)
m2 <- as.matrix(dtm2)
v2 <- sort(rowSums(m2),decreasing=TRUE)
d2 <- data.frame(word = names(v2),freq=v2)
head(d2, 10)
```

```
vordcloud(words = d2$word, freq = d2$freq, min.freq = 1,
          max.words=200, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```

Stop words, punctuation and whitespaces are removed because they do not contain important significance. Since we are trying to analyze all tweets irrespective of airlines, we decided to remove the names of all airlines and other words that are of less significance.



```
               word freq
cancelled cancelled  920
now             now  823
service     service  741
hours         hours  656
just           just  616
help           help  613
hold           hold  613
customer   customer  605
time           time  591
plane         plane  522
>
```

Fig 3.4.1. Frequency of Negative words

Finding the Sentiment

The word “cancelled” is most commonly used in tweets to describe the negative experience. Other commonly used words to describe negative experience are “service”, “help” etc. These words display dissatisfaction, frustration and hate toward the airlines. And it also correlates to the results of negative tweet analysis where most negative comments are due to poor customer service. However, in cases such as Delta it is due to late flights.

Word cloud for all tweets

The **tm_map()** function is used to remove unnecessary white space, to convert the text to lower case, to remove common stopwords like ‘the’, “we”.

The information value of ‘stopwords’ is near zero since they are so common in a language. Removing this kind of words is useful before further analyses.

```
docs <- Corpus(VectorSource(airlineTweets$text))
docs <- tm_map(docs, content_transformer(tolower))
docs <- tm_map(docs, removeNumbers)
docs <- tm_map(docs, removeWords, stopwords("english"))
docs <- tm_map(docs, removeWords, c("usairways" ,"united", "flight" ,
                                    "americanair" , "jetblue" , "southwestair","get","can"))
docs <- tm_map(docs, removePunctuation)
docs <- tm_map(docs, stripWhitespace)
```

```
dtm2 <- TermDocumentMatrix(docs2)
m2 <- as.matrix(dtm2)
v2 <- sort(rowSums(m2),decreasing=TRUE)
d2 <- data.frame(word = names(v2),freq=v2)
head(d2, 10)
```

Word cloud library is used to show the frequency of words

```
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
          max.words=200, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```

	word	freq
thanks	thanks	1073
cancelled	cancelled	1056
now	now	1023
just	just	966
service	service	955
help	help	856
time	time	783
will	will	763
customer	customer	748
hours	hours	681

Fig 3.4.2 Frequency of all words

Observations:

The word "thanks" occurs more frequently which is the only positive word used to describe the experience.

Finding the sentiment of positive tweets

We found that many of the positive tweets directed toward airlines were sarcastic. For example, "RT @Samantha_billy: Thanks Delta! I had to cancel the meeting because my flight got delayed and no reason was provided upon request. I am never getting on your plane again". Another example "RT @ZainMobeen: Thanks for ruining my day United. Your customer care representative was disrespectful and incompetent. I was on the phone for 2 hours and still couldn't get my problem solved #BoycottUnited.

4. SENTIMENT ANALYSIS

4.1. What is Sentiment Analysis?

Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral**.** A sentiment analysis system for text analysis combines natural language processing (NLP) and machine learning techniques to assign weighted sentiment scores to the entities, topics, themes and categories within a sentence or phrase.

Sentiment analysis is done based on lexicons. A lexicon in simpler terms is a vocabulary, say the English lexicon. A lexicon is a selection of words with the two polarities that can be used as a metric in sentiment analysis.

In this project, we have made use of the syuzhet package. While there are several packages for sentiment analysis on CRAN, the syuzhet package is great to learn with because it is a combination of the most common lexicons like nrc. We also make use of ggplot2 to further visualize our results from the sentiment analysis.

We have used **get_nrc_sentiment** function to get the Emotions and Valence from NRC Dictionary. To perform the Sentiment Analysis, we used the NRC SENTIMENT which implements the Saif Muhammad's NRC Emotion lexicon. The NRC emotion lexicon is a list of words and their associations with eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) generated automatically from tweets with emotion-word hashtags such as #happy and #anger and two sentiments negative and positive.

The output of **get_nrc_sentiment** is a data frame where each row represents a sentence from the original file. The columns include one for each emotion type as well as a positive or negative valence. The ten columns are as follows: "anger", "anticipation", "disgust", "fear", "joy", "sadness", "surprise", "trust", "negative", "positive."

4.2. Sentiment Score for all tweets

Below is the sentiment score for all 8 emotions plus 2 sentiments which we have generated using the **get_nrc_sentiment** function for all the 6 airlines.

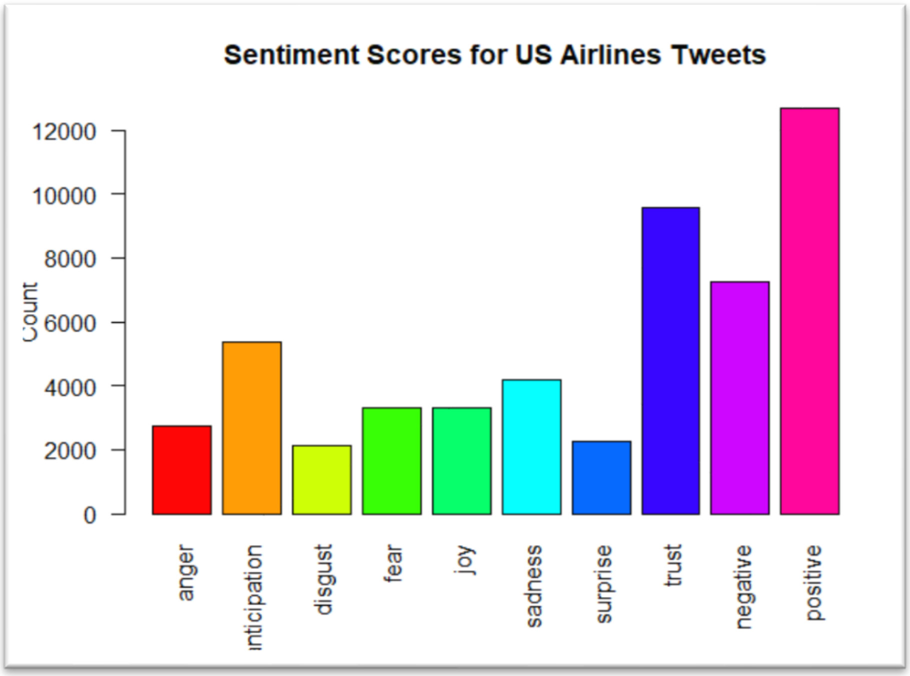


Fig 4.2.1 Sentiment Score for all the tweets

OBSERVATION:

- Out of the 8 emotions, we see that the cumulative count of trust tweets is maximum for all the 6 airlines.
- Since, trust counts for maximum counts that leads to a positive sentiment towards all the 6 airlines.
- Almost 80% of the total positive tweets have emotions showing trust.
- The ratio of negative to positive sentiment is approx. 1:2

Below is the count for all the tweets. We are considering the highlighted tweet 4 for our analysis.

```
Console   Terminal ×
C:/Users/wilfr/Downloads/IS 688 Sentiment Analysis/Web_Mining-Sentiment_Analysis/R Scripts/
> head(s)
  anger anticipation disgust fear joy sadness surprise trust negative positive
1     0            0       0    0   0       0        0     0        0        0
2     0            0       0    0   0       0        0     0        0        0
3     0            0       0    0   0       0        1     0        0        0
4     3            1       1    2   1       1        2     1        3        1
5     1            0       1    1   0       1        0     0        1        0
6     1            1       1    2   1       1        0     1        1        2
```

The tweet 4 contains 3 words that are showing anger, 1 word that shows anticipation, 1 word that shows disgust and so on. Let us consider, tweet 4 for our analysis.

Our aim is to find the 3 words in tweet 4 that shows Anger.

The below 3 words when we pass to the get_nrc_sentiment, we get the below results

A. Aggressive

```
C:/Users/wilfr/Downloads/IS 688 Sentiment Analysis/Web_Mining-Sentiment_Analysis/R Scripts/
 tweets[4]
1] "@VirginAmerica it's really aggressive to blast obnoxious \"entertainment\" in your guests' faces &amp; they have little
ecourse #fabulous"
 get_nrc_sentiment('aggressive')
  anger anticipation disgust fear joy sadness surprise trust negative positive
      1            0       0    1   0       0        0     0        1        0
```

B. Obnoxious

```
Console   Terminal ×
C:/Users/wilfr/Downloads/IS 688 Sentiment Analysis/Web_Mining-Sentiment_Analysis/R Scripts/
 [ reached getOption("max.print") -- omitted 13640 rows ]
> get_nrc_sentiment('obnoxious')
  anger anticipation disgust fear joy sadness surprise trust negative positive
1     1            0       1    0   0       1        0     0        1        0
> |
```

C. Blast

```
Console   Terminal ×
C:/Users/wilfr/Downloads/IS 688 Sentiment Analysis/Web_Mining-Sentiment_Analysis/R Scripts/
[1] "@VirginAmerica it's really aggressive to blast obnoxious \"entertainment\" in your guests' faces &amp;
they have little recourse #fabulous"
> get_nrc_sentiment('blast')
  anger anticipation disgust fear joy sadness surprise trust negative positive
1     1            0       0    1   0       0        1     0        1        0
>
```

4.3. Sentiment Score for each airline

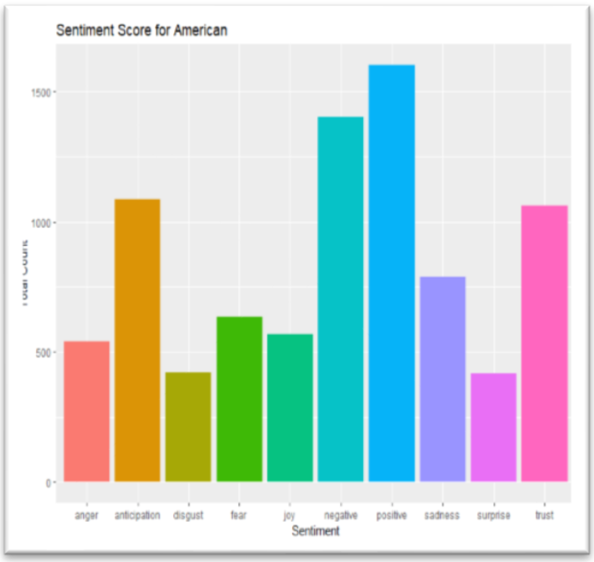


Fig 4.3.1 American Airlines

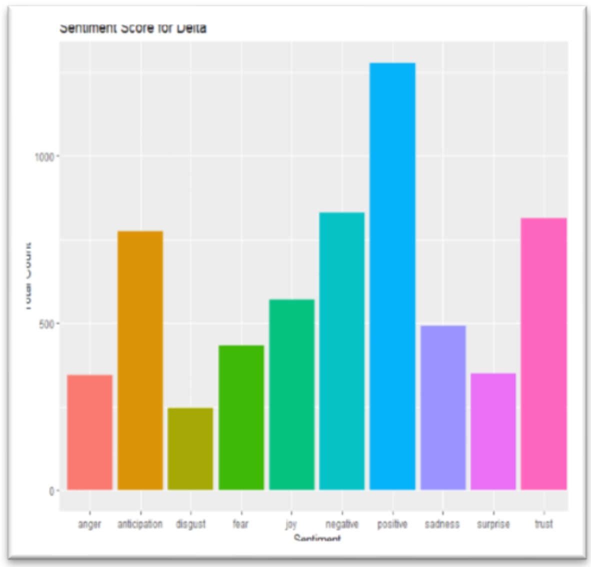
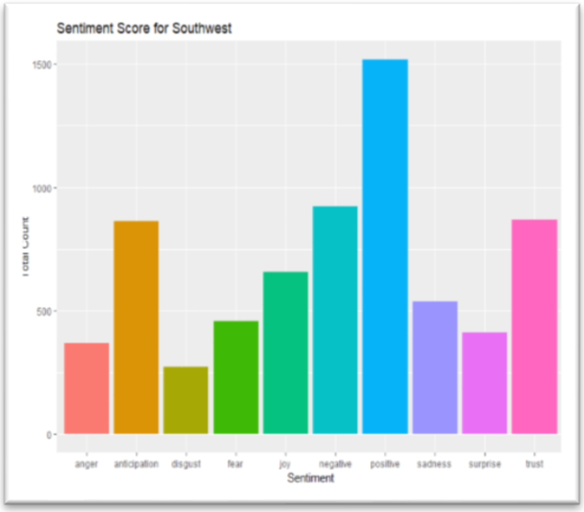


Fig 4.3.2 Delta Airlines

OBSERVATIONS:

- For American Airlines, the negative to positive sentiment ratio is almost equal.
- For Delta Airlines, the negative to positive sentiment ratio is almost half.



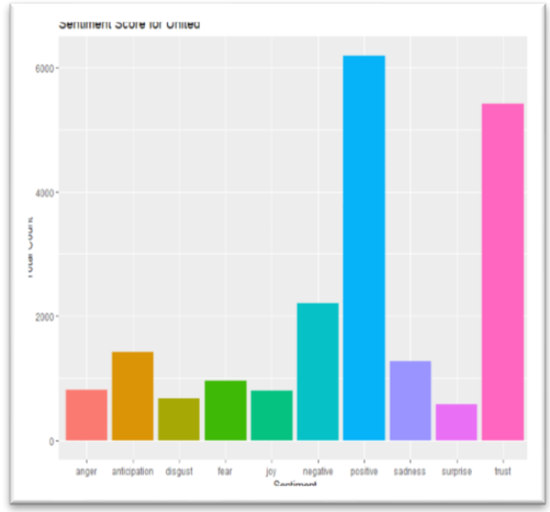
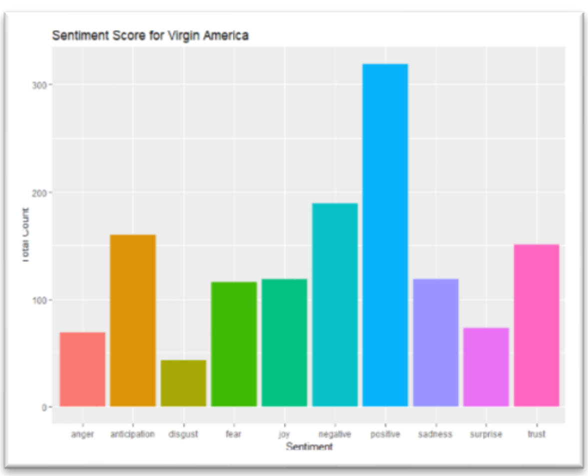


Fig 4.3.3 Southwest

Fig 4.3.4 United Airlines

OBSERVATIONS:

For Southwest, the negative to positive ratio is almost half.
For United Airlines, almost 80% of the total positive tweets have emotions showing trust



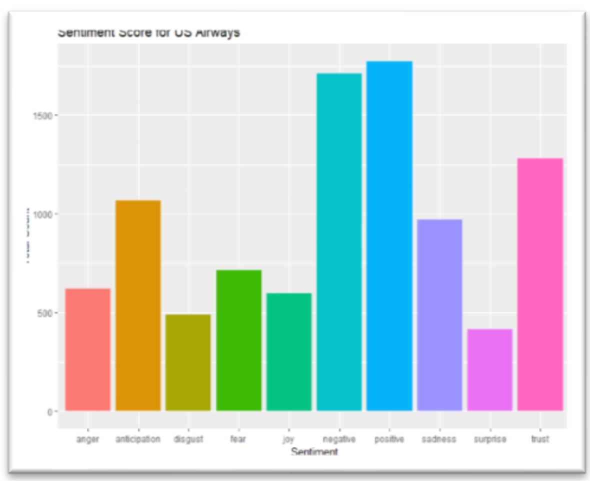


Fig 4.3.5 Virgin America

Fig 4.3.6 US Airways

OBSERVATIONS:

- For US Airways, the negative to positive sentiment ratio is almost equal.
- For Virgin America, the negative to positive sentiment ratio is almost half.

5. CONCLUSION

Based on our analysis, below are the findings that we observed:

- Most tweets have negative sentiment (>60%).
- Most tweets are targeted towards United airlines, followed by American and US Airways.
- Virgin receives very few tweets.
- Main reasons for negative sentiment are Customer Service Issues and Late Flights.
- Negative sentiment tweets towards Delta are based mostly on late flights and not so much on Customer Service Issues as for the rest of the airlines.
- Most tweets are not retweeted.
- Most tweets come from the States.
- Most of the tweets targeted towards American, United and US Airways contain negative sentiment.
- Tweets targeted towards Delta, Virgin and Southwest containing roughly same proportion of negative, neutral and positive sentiment.

On performing the exploratory analysis based on Distribution, Time, Reasons for negative sentiment and Sentiment Score for all tweets going forward we can predict the future events for different predictive models.

REFERENCES

1. https://www.figure-eight.com/data-for-everyone/
2. https://rpubs.com/williamsurles/316682
3. http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know
4. https://www.lexalytics.com/technology/sentiment-analysis
5. http://cnorval.github.io/2015/12/24/Sentiment_Analysis/
6. https://www.rdocumentation.org/packages/syuzhet/versions/1.0.4
7. https://www.rdocumentation.org/packages/syuzhet/versions/1.0.4/topics/get_nrc_sentiment
8. https://rstudio-pubs-static.s3.amazonaws.com/302066_fe1dd2a635fa41198b18c87a64f5620c.html
9. https://rstudio-pubs-static.s3.amazonaws.com/66739_c4422a1761bd4ee0b0bb8821d7780e12.html
10. https://www.kaggle.com/
11. https://www.rdocumentation.org/packages/ggplot2/versions/3.1.0
12. https://www.rdocumentation.org/packages/dplyr/versions/0.7.8