Twitter US Airline Sentiment

Abstract:

Micro-blogs are a challenging new source of information for data mining techniques. Twitter is a micro-blogging service built to discover what is happening at any moment in time, anywhere in the world. Twitter messages are short, and generated constantly, and well suited for knowledge discovery using data stream mining. By utilizing the Twitter data, Sentiment analysis was performed to analyze how travelers express there feeling on Twitter and also the reasons associated with these sentiments. The main concentration of the sentiment analysis is on negative or bad sentiments with respect to the flights used.

Introduction:

In the past few years, there has been a huge growth in the use of microblogging platforms such as Twitter. Spurred by that growth, companies and media organizations are increasingly seeking ways to mine Twitter for information about what people think and feel about their products and services.

Anyone who travels regularly recognizes that airlines struggle to deliver a consistent, positive customer experience. Through extensive interview and survey work, the American Customer Satisfaction Index quantifies this impression. Meanwhile, the immediacy and accessibility of Twitter provides a real-time glimpse into consumer's frustration.

Sentiment extraction systems usually require an extensive set of manually supplied sentiment words or a handcrafted sentiment-specific dataset. With the recent popularity of article tagging, some social media types like blogs allow users to add sentiment tags to articles. This allows to use blogs as a large user-labeled dataset for sentiment learning and identification.

Here, a twitter data Extracted from February 2015 is used to get a general idea as to analyze how travelers generally express their feeling based on their travel experience. The idea is to classify the tweets based on positive, negative and neutral tweets and further try to analyze the reason for negative tweets. This analyzation is done both on overall airlines and individual airlines used by travelers and classifying the negative tweets and finding the frequency of these tweets.

Apart from Sentiment Analysis, Text Analysis of Sentiments would also show similar results. Clustering methods can also be used to analyze the association of words. Also Word clouds can be used to show the sentiments.

Code with Documentation:

Additional packages needed

- If necessary install these packages.

```
install.packages("ggplot2"); install.packages("gridExtra");
install.packages("maps");
```

```
require(ggplot2)

## Loading required package: ggplot2

require(gridExtra)

## Loading required package: gridExtra

## Warning: package 'gridExtra' was built under R version 3.2.5

require(maps)

## Loading required package: maps

## Warning: package 'maps' was built under R version 3.2.5

##
##  # maps v3.1: updated 'world': all lakes moved to separate new #
##  # 'lakes' database. Type '?world' or 'news(package="maps")'.  #
```

loading the data

```
getwd()

## [1] "C:/Users/Neha/Desktop"

setwd("C:/Users/Neha/Desktop")
data = read.csv('Tweets.csv')
dim(data)

## [1] 14640    15

str(data)

## 'data.frame':    14640 obs. of  15 variables:
##  $ tweet_id                    : num  5.7e+17 5.7e+17 5.7e+17 5.7e+17 5.7e
+17 ...
##  $ airline_sentiment           : Factor w/ 3 levels "negative","neutral",.
.: 2 3 2 1 1 1 3 2 3 3 ...
##  $ airline_sentiment_confidence: num  1 0.349 0.684 1 1 ...
##  $ negativereason              : Factor w/ 11 levels "","Bad Flight",..: 1
1 1 2 3 3 1 1 1 1 ...
##  $ negativereason_confidence   : num  NA 0 NA 0.703 1 ...
##  $ airline                     : Factor w/ 6 levels "American","Delta",..:
6 6 6 6 6 6 6 6 6 6 ...
```

```
##  $ airline_sentiment_gold      : Factor w/ 4 levels "","negative",..: 1 1
1 1 1 1 1 1 1 1 ...
##  $ name                        : Factor w/ 7701 levels "___the___","__betr
ayal",..: 1073 3477 7666 3477 3477 3477 1392 5658 1874 7665 ...
##  $ negativereason_gold         : Factor w/ 14 levels "","Bad Flight",..: 1
1 1 1 1 1 1 1 1 ...
##  $ retweet_count               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ text                        : Factor w/ 14427 levels "\"LOL you guys ar
e so on it\" - me, had this been 4 months ago...â<U+0080><U+009C>@JetBlue: Ou
r fleet's on fleek. http://t.co/LYcARlTFHlâ<U+0080>",..: 14005 13912 13790 1
3844 13648 13926 14038 13917 14004 13846 ...
##  $ tweet_coord                 : Factor w/ 833 levels "","[-33.87144962, 1
51.20821275]",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ tweet_created               : Factor w/ 14247 levels "2015-02-16 23:36:
05 -0800",..: 14212 14170 14169 14168 14166 14165 14164 14160 14158 14106 ...
##  $ tweet_location              : Factor w/ 3082 levels "","'Greatness has
no limits'",..: 1 1 1465 1 1 1 2407 1529 2389 1529 ...
##  $ user_timezone               : Factor w/ 86 levels "","Abu Dhabi",..: 32
64 29 64 64 64 64 64 64 32 ...
```

Exploratory data analysis: chiecking the columns containing no data (NAs)

```
# find the cells cintaining "" , " " or NAs
data <- as.data.frame(apply(data, 2, function(x) gsub("^$|^$", NA, x)))

# Checking for cols containing NAs and their total number
checkdata <- apply(data, 2, function(x) sum(is.na(x)))
as.data.frame(checkdata )
```

```
##                               checkdata
## tweet_id                              0
## airline_sentiment                     0
## airline_sentiment_confidence          0
## negativereason                     5462
## negativereason_confidence          4118
## airline                               0
## airline_sentiment_gold            14600
## name                                  0
## negativereason_gold               14608
## retweet_count                         0
## text                                  0
## tweet_coord                       13621
## tweet_created                         0
## tweet_location                     4733
## user_timezone                      4820
```

I would try to explore Negative sentiment and tweet location.

Sentiment Analysis:

Trying to get the proportion of tweets with each sentiment

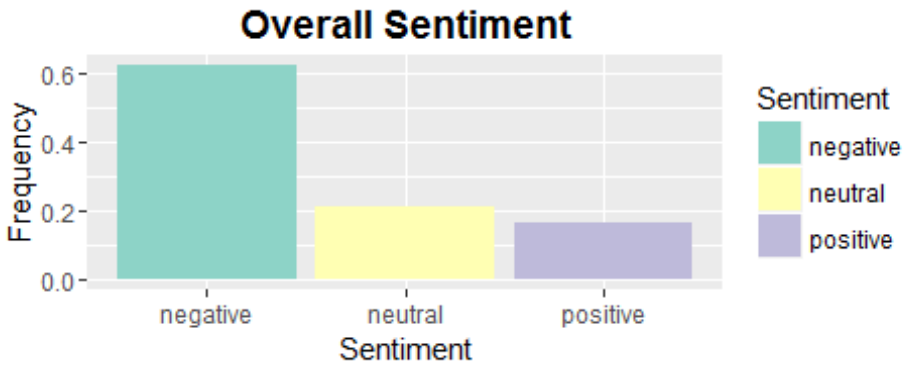
```
prop.table(table(data$airline_sentiment))
```

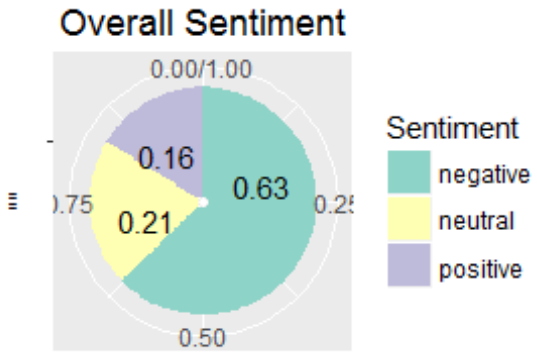
```
##
##  negative   neutral  positive
## 0.6269126 0.2116803 0.1614071
```

```
# generate a dataframe for plotting in ggplot2
SmallData <- as.data.frame(prop.table(table(data$airline_sentiment)))
colnames(SmallData) <- c('Sentiment', 'Frequency')
SmallData
```

```
##   Sentiment Frequency
## 1  negative 0.6269126
## 2   neutral 0.2116803
## 3  positive 0.1614071
```

```
# create blank theme for pie chart


gbar <- ggplot(SmallData, aes(x = Sentiment, y = Frequency, fill = Sentiment)
)+  scale_fill_brewer(palette="Set3")
gpie = ggplot(SmallData, aes(x = "", y = Frequency, fill = Sentiment)) + scal
e_fill_brewer(palette="Set3")

plot1 <- gbar + geom_bar(stat = 'identity') + ggtitle("Overall Sentiment") +
theme(plot.title = element_text(size = 14, face = 'bold', vjust = 1), axis.ti
tle.y = element_text(vjust = 2), axis.title.x = element_text(vjust = -1))

plot2 <- gpie + geom_bar(stat = 'identity') + coord_polar("y", start = 0)  +
theme(axis.title.x = element_blank()) + geom_text(aes(y = Frequency/3 + c(0,
cumsum(Frequency)[-length(Frequency)]),label = round(Frequency,2)), size = 4)
+ ggtitle('Overall Sentiment')

grid.arrange(plot1, plot2, ncol = 1, nrow = 2)
```



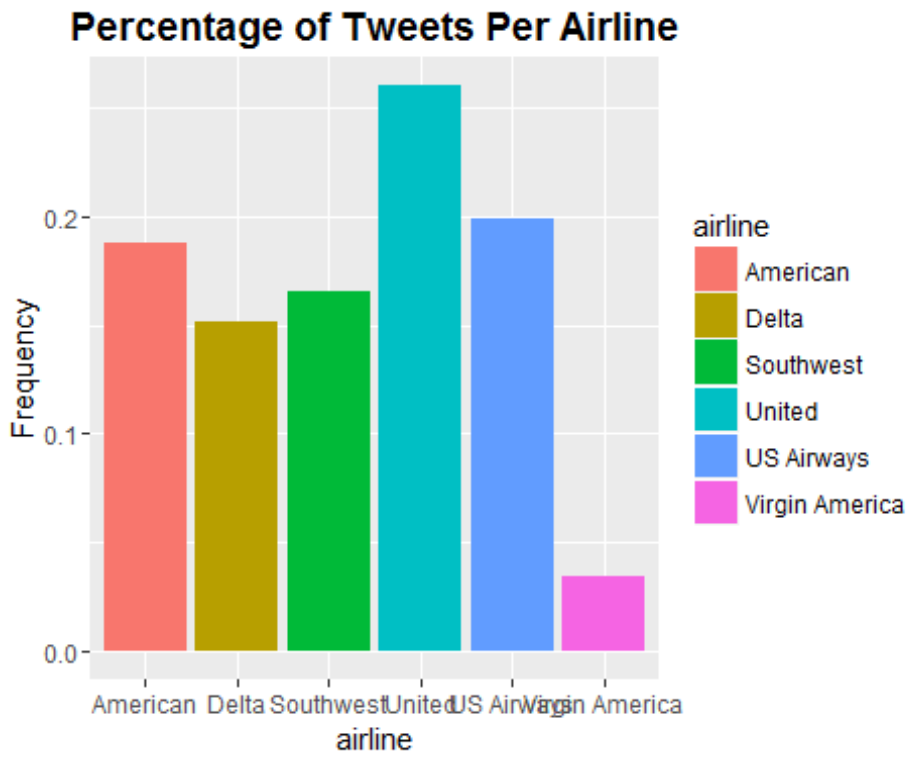


Percentage of tweets per airline

```
# get the proportion of tweets per airline and convert it into a dataframe
TweetDataFrame <- as.data.frame(prop.table(table(data$airline)))
colnames(TweetDataFrame) <- c('airline' , 'Frequency')
TweetDataFrame
```

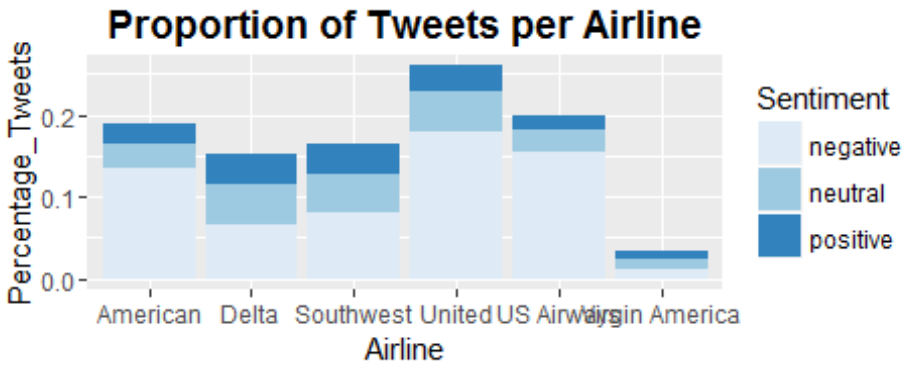
```
##          airline  Frequency
## 1       American 0.18845628
## 2          Delta 0.15177596
## 3      Southwest 0.16530055
## 4         United 0.26106557
## 5     US Airways 0.19897541
## 6 Virgin America 0.03442623
```

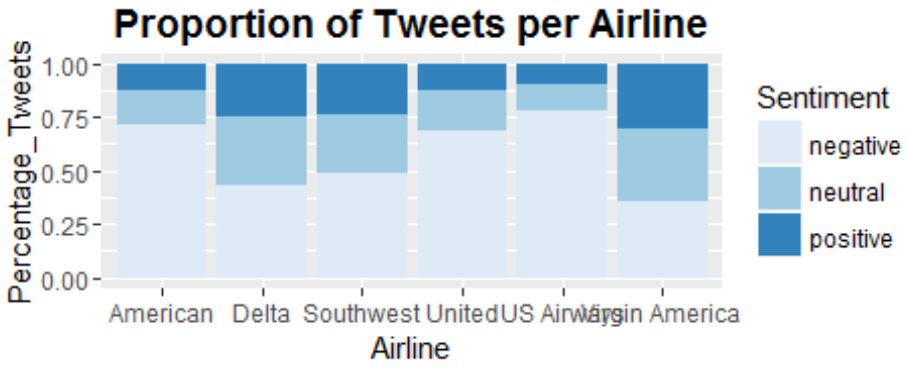
```
# Plotting a bar graph to get the percentage of tweets per Airline

gbar <- ggplot(TweetDataFrame, aes(x = airline, y = Frequency, fill = airline
))

gbar + geom_bar(stat = 'identity') + scale_colour_gradientn(colours=rainbow(4
)) + ggtitle('Percentage of Tweets Per Airline') + theme(plot.title = element
_text(size = 14, face = 'bold', vjust = 1))
```



Proportion of Negative sentiments per airline

```
# get the proportion of tweets per airline and convert it into a dataframe

TweetDataFrame <- as.data.frame(prop.table(table(data$airline_sentiment, data
$airline )))
colnames(TweetDataFrame) <- c('Sentiment', 'Airline', 'Percentage_Tweets')

# PLotting the graph

gbar <- ggplot(TweetDataFrame, aes(x = Airline, y = Percentage_Tweets, fill =
Sentiment))

# Plotting the graph to show tweet sentiment per

plot <- gbar + geom_bar(stat = 'identity') +   ggtitle('Proportion of Tweets
per Airline') + scale_fill_brewer() + theme (plot.title = element_text(size =
14, face = 'bold', vjust = 1))

plot2 <- gbar + geom_bar(stat = 'identity', position = 'fill') +   ggtitle('P
roportion of Tweets per Airline') + scale_fill_brewer() + theme (plot.title =
element_text(size = 14, face = 'bold', vjust = 1))

grid.arrange(plot, plot2)
```





Finding the general Reasons for Negative Sentiment

```
# Creating a data frame to get the reasons for negative tweets

TweetDataFrame <- as.data.frame(prop.table(table(data$negativereason)))
colnames(TweetDataFrame) <- c("reason", "Frequency")
TweetDataFrame
```

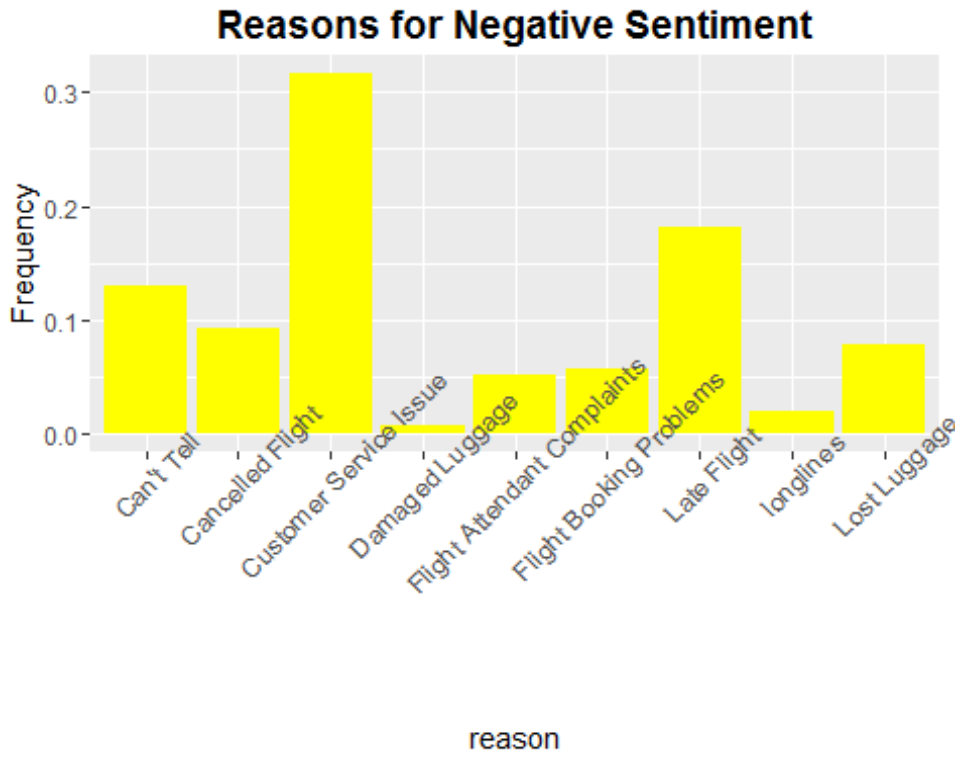
```
##                          reason   Frequency
## 1                   Bad Flight 0.063194596
## 2                   Can't Tell 0.129657878
## 3               Cancelled Flight 0.092285901
## 4         Customer Service Issue 0.317062541
## 5                 Damaged Luggage 0.008062759
## 6  Flight Attendant Complaints 0.052407932
## 7       Flight Booking Problems 0.057637830
## 8                    Late Flight 0.181412072
## 9                      longlines 0.019394204
## 10                 Lost Luggage 0.078884289
```

```
# Removing the first row
TweetDataFrame = TweetDataFrame[-1,]
TweetDataFrame
```

```
##                          reason   Frequency
## 2                   Can't Tell 0.129657878
```

```
## 3             Cancelled Flight 0.092285901
## 4       Customer Service Issue 0.317062541
## 5              Damaged Luggage 0.008062759
## 6  Flight Attendant Complaints 0.052407932
## 7      Flight Booking Problems 0.057637830
## 8                  Late Flight 0.181412072
## 9                    longlines 0.019394204
## 10               Lost Luggage 0.078884289
```

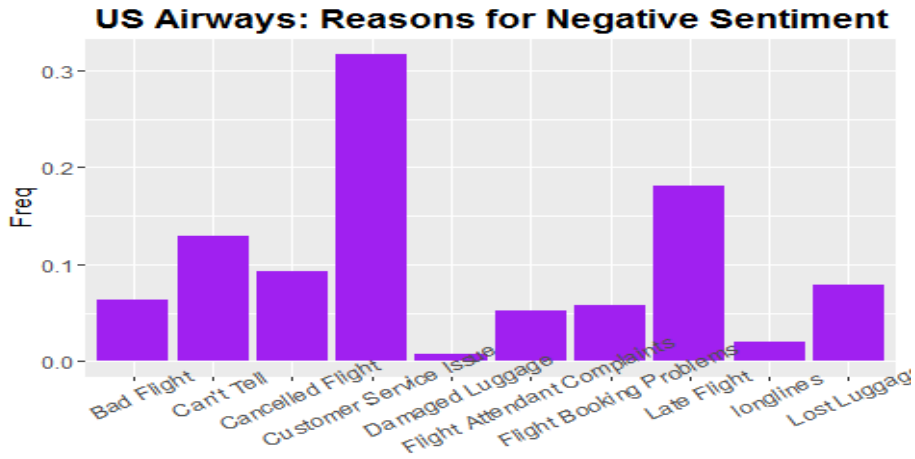
```
NegReaplot <- ggplot(TweetDataFrame, aes(x = reason, y = Frequency)) + geom_b
ar(stat = 'identity', fill = 7) + ggtitle('Reasons for Negative Sentiment') +
theme(plot.title = element_text(size = 14, face = 'bold', vjust = 1), axis.ti
tle.x = element_text(vjust = -0.1), axis.text.x = element_text(angle = 45, si
ze = 10, vjust = 1))
NegReaplot
```



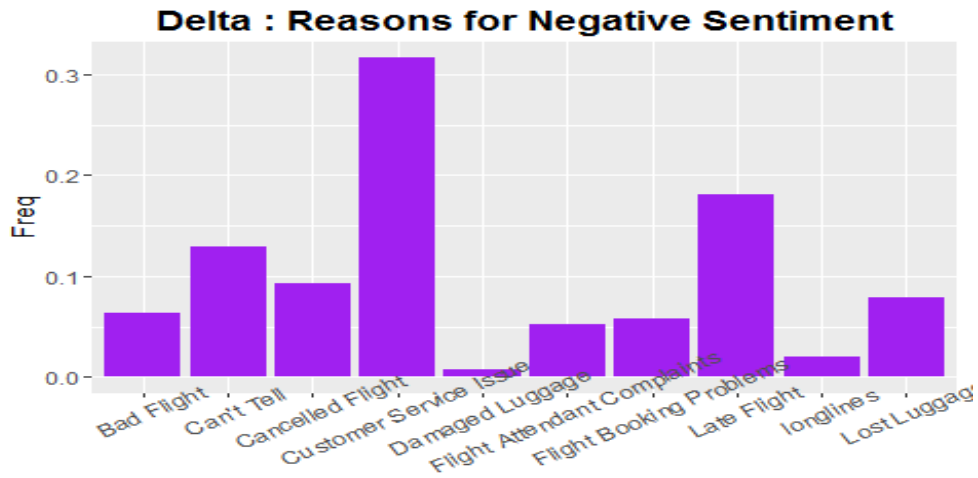
Reasons For Negative Sentiment per airline

```
# First subset the data airline wise then plot to show the reason for negativ
e sentiment

# Subset the data for American Airlines
AmericanAirline <- subset(data, airline = 'American')

# Get the plot
AAplot <- ggplot(as.data.frame(prop.table(table(AmericanAirline$negativereaso
```

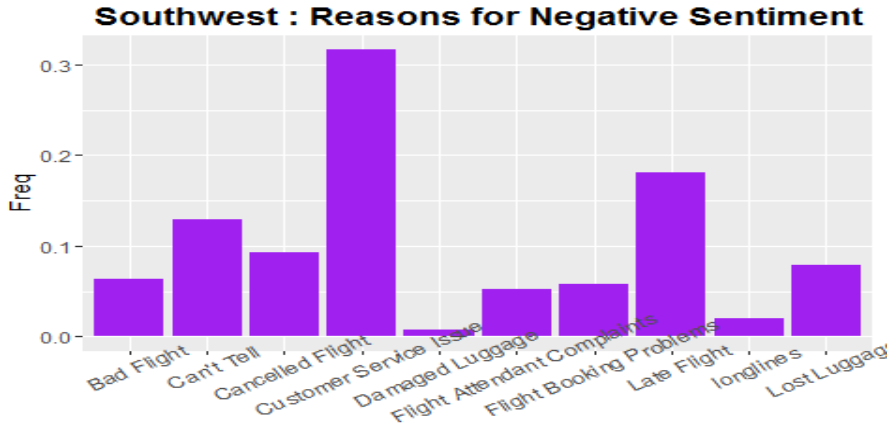
```
n))), aes(x = Var1, y = Freq)) + geom_bar(stat = 'identity', fill = 'Purple')
+ ggtitle('American Airlines: Reasons for Negative Sentiment') + theme(plot.t
itle = element_text(size = 14, face = 'bold', vjust = 1), axis.title.x = elem
ent_blank(), axis.text.x = element_text(angle = 30, size = 10, vjust = 1))

# Subset the data for US Airways
USAirways <- subset(data, airline = 'US Airways')

# Generate the plot
USAplot <- ggplot(as.data.frame(prop.table(table(USAirways$negativereason))),
aes(x = Var1, y = Freq)) + geom_bar(stat = 'identity', fill = 'Purple') + ggt
itle('US Airways: Reasons for Negative Sentiment') + theme(plot.title = eleme
nt_text(size = 14, face = 'bold', vjust = 1), axis.title.x = element_blank(),
axis.text.x = element_text(angle = 30, size = 10, vjust = 1))
USAplot
```




```
# Subset the data for Delta
Delta <- subset(data, airline = 'Delta')

# Generate the plot
Delplot <- ggplot(as.data.frame(prop.table(table(Delta$negativereason))), aes
(x = Var1, y = Freq)) + geom_bar(stat = 'identity', fill = 'Purple') + ggtitl
e('Delta : Reasons for Negative Sentiment') + theme(plot.title = element_text
(size = 14, face = 'bold', vjust = 1), axis.title.x = element_blank(), axis.t
ext.x = element_text(angle = 30, size = 10, vjust = 1))
Delplot
```

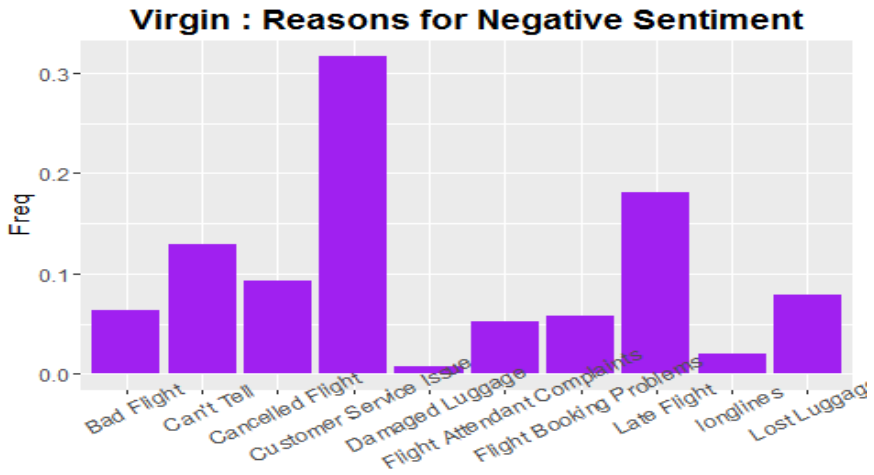



```r
# Subset the data for Southwest
Southwest <- subset(data, airline = 'Southwest')

# Generate the plot
SWplot <- ggplot(as.data.frame(prop.table(table(Southwest$negativereason))),
aes(x = Var1, y = Freq)) + geom_bar(stat = 'identity', fill = 'Purple') + ggt
itle('Southwest : Reasons for Negative Sentiment') + theme(plot.title = eleme
nt_text(size = 14, face = 'bold', vjust = 1), axis.title.x = element_blank(),
axis.text.x = element_text(angle = 30, size = 10, vjust = 1))
SWplot
```

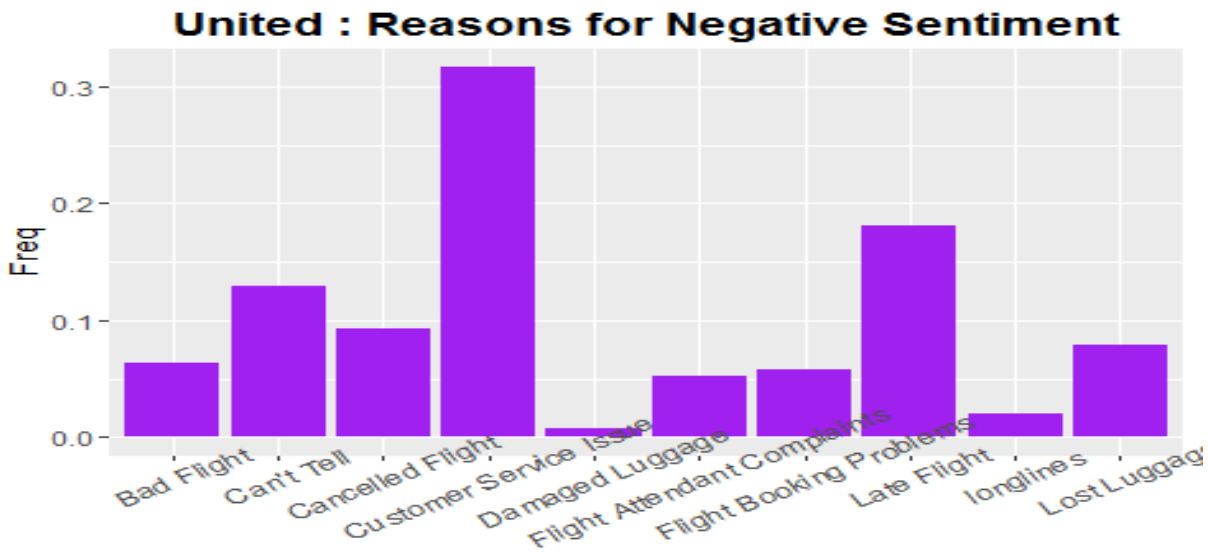



```r
# Subset the data for Virgin
Virgin <- subset(data, airline = 'Virgin')

# Generate the plot
Virplot <- ggplot(as.data.frame(prop.table(table(Virgin$negativereason))), ae
s(x = Var1, y = Freq)) + geom_bar(stat = 'identity', fill = 'Purple') + ggtit
le('Virgin : Reasons for Negative Sentiment') + theme(plot.title = element_te
xt(size = 14, face = 'bold', vjust = 1), axis.title.x = element_blank(), axis
```

```
.text.x = element_text(angle = 30, size = 10, vjust = 1))
Virplot
```




```
# Subset the data for United
United <- subset(data, airline = 'United')

# Generate the plot
Unplot <- ggplot(as.data.frame(prop.table(table(United$negativereason))), aes
(x = Var1, y = Freq)) + geom_bar(stat = 'identity', fill = 'Purple') + ggtitl
e('United : Reasons for Negative Sentiment') + theme(plot.title = element_tex
t(size = 14, face = 'bold', vjust = 1), axis.title.x = element_blank(), axis.
text.x = element_text(angle = 30, size = 10, vjust = 1))
Unplot
```




```{r}

# Get locations of the tweet
Location <- data$tweet_coord
class(Location)
proj4string(Location)
Location <- Location[complete.cases(Location)]
Location <- as.data.frame(Location)
Location$count = 1
Location$Location = as.character(Location$Location)

# remove the duplicate coordinates and count the no of time they appear
Location <- aggregate(count ~ Location, data = Location, FUN = sum)
Location <- Location[-5]
coords <- strsplit(Location$Location,",")

# Now separate longitude and latitude

lat <- NULL
long <- NULL

for (i in 1:length(coords)) {

  lat = c(lat, substring(coords[[i]][1], 2))
  long = c(long, coords[[i]][2])
}

Location$lat = lat
Location$long = long

Location$long = substr(Location$long, 1, nchar(Location$long)-1)

head(Location)
dim(Location)
```

```
# PLot these on the map of United States

UStates <- map_data("state")
USplot <- ggplot()

USlpot <- USplot + geom_polygon(data=UStates, aes(x=long, y=lat, group = group), color
="black", fill = 'lightblue') + ggtitle("Location of tweets across the United States")
USplot <- USplot +  geom_point(data=Location, aes(x=long, y=lat, size = count), color
= "coral1") + scale_size(name="Total Tweets")

USplot <- USplot + xlim(-125, -65) + ylim(25, 50)
USplot
```

The Above code was for plotting the tweets on the Map of United states. That would help infer more clearly from which part were the most tweets from. The code does not execute when you try to combine poly and points together. It shows an error.

Analysis of Text Content

```
install.packages("dplyr"); install.packages("tm");
install.packages("Snowball"); install.packages("wordcloud");
install.packages("cluster");
```

```
require(dplyr)

## Loading required package: dplyr

## Warning: package 'dplyr' was built under R version 3.2.5

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

require(tm)

## Loading required package: tm

## Loading required package: NLP

require(SnowballC)

## Loading required package: SnowballC

require(wordcloud)

## Loading required package: wordcloud
```

```
## Warning: package 'wordcloud' was built under R version 3.2.5

## Loading required package: RColorBrewer

require(cluster)

## Loading required package: cluster
```

loading the data

```
getwd()

## [1] "C:/Users/Neha/Desktop"

setwd("C:/Users/Neha/Desktop")
data = read.csv('Tweets.csv')
data = select(data, airline_sentiment, negativereason, airline, text)
head(data)

##   airline_sentiment negativereason        airline
## 1           neutral                Virgin America
## 2          positive                Virgin America
## 3           neutral                Virgin America
## 4          negative     Bad Flight Virgin America
## 5          negative     Can't Tell Virgin America
## 6          negative     Can't Tell Virgin America
##
text
## 1
@VirginAmerica What @dhepburn said.
## 2                                                                    @VirginA
merica plus you've added commercials to the experience... tacky.
## 3                                                                     @Virgin
America I didn't today... Must mean I need to take another trip!
## 4           @VirginAmerica it's really aggressive to blast obnoxious "ente
rtainment" in your guests' faces &amp;amp; they have little recourse
## 5
@VirginAmerica and it's a really big bad thing about it
## 6 @VirginAmerica seriously would pay $30 a flight for seats that didn't ha
ve this playing.\nit's really the only bad thing about flying VA

# Removing the @

data$text <- gsub("^@\\w+ *", "", data$text)
head(data)

##   airline_sentiment negativereason        airline
## 1           neutral                Virgin America
## 2          positive                Virgin America
## 3           neutral                Virgin America
## 4          negative     Bad Flight Virgin America
```

```
## 5          negative     Can't Tell Virgin America
## 6          negative     Can't Tell Virgin America
##
text
## 1
What @dhepburn said.
## 2                                                                    plus you
've added commercials to the experience... tacky.
## 3                                                                     I didn'
t today... Must mean I need to take another trip!
## 4           it's really aggressive to blast obnoxious "entertainment" in y
our guests' faces &amp;amp; they have little recourse
## 5
and it's a really big bad thing about it
## 6 seriously would pay $30 a flight for seats that didn't have this playing
.\nit's really the only bad thing about flying VA
```

```
# Dividing tweets based on positive and negative sentiment

posTweets <- subset(data, airline_sentiment == 'positive')
dim(posTweets)
```

```
## [1] 2363    4
```

```
NegTweets <- subset(data, airline_sentiment == 'negative')
dim(NegTweets)
```

```
## [1] 9178    4
```

```
# Removing these words seemed to be necessary as they are repeated a lot
wordsToRemove = c('get', 'cant', 'can', 'now', 'just', 'will', 'dont', 'ive',
'got', 'much')

# analyse corpus
analyseText = function(text_to_analyse){
    CorpusTranscript = Corpus(VectorSource(text_to_analyse))
    CorpusTranscript = tm_map(CorpusTranscript, content_transformer(tolower),
lazy = T)
    CorpusTranscript = tm_map(CorpusTranscript, PlainTextDocument, lazy = T)
    CorpusTranscript = tm_map(CorpusTranscript, removePunctuation)
   CorpusTranscript = tm_map(CorpusTranscript, removeWords, wordsToRemove)
    CorpusTranscript = tm_map(CorpusTranscript, removeWords, stopwords("engli
sh"))
    CorpusTranscript = DocumentTermMatrix(CorpusTranscript)
    CorpusTranscript = removeSparseTerms(CorpusTranscript, 0.97) # keeps a ma
trix 97% sparse
    CorpusTranscript = as.data.frame(as.matrix(CorpusTranscript))
    colnames(CorpusTranscript) = make.names(colnames(CorpusTranscript))

    return(CorpusTranscript)
}
```

```r
# Analysing positive and Negative tweets
Nword <- analyseText(NegTweets$text)
dim(Nword)
```

```
## [1] 9178   30
```

```r
Pword <- analyseText(posTweets$text)
dim(Pword)
```

```
## [1] 2363   18
```

```r
# Determining the Frequency of negative words and creating a word cloud

Freq_Nword <- colSums(Nword)
Freq_Nword <- Freq_Nword[order(Freq_Nword, decreasing = T)]
head(Freq_Nword)
```

```
##    flight cancelled   service     hours      help      hold
##      2900       920       740       644       610       607
```

```r
wordcloud(freq = as.vector(Freq_Nword), words = names(Freq_Nword), random.ord
er = FALSE, random.color = FALSE, colors = brewer.pal(9, 'RdPu')[4:9] )
```

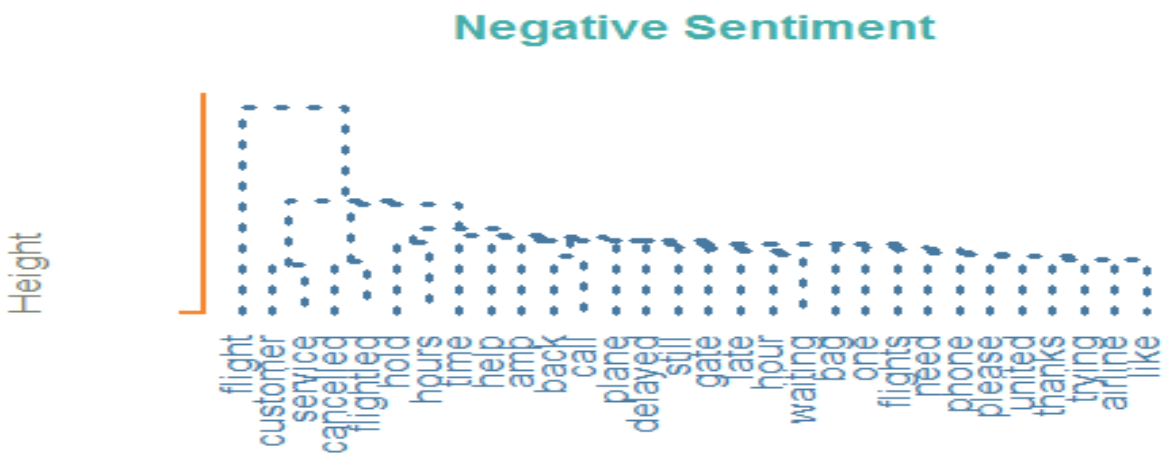

```r
# Analysing Negative words generally mentioned in tweets

analyseText2 = function(text_to_analyse){

    CorpusTranscript = Corpus(VectorSource(text_to_analyse))
```

```
    CorpusTranscript = tm_map(CorpusTranscript, content_transformer(tolower),
lazy = T)
    CorpusTranscript = tm_map(CorpusTranscript, PlainTextDocument, lazy = T)
    CorpusTranscript = tm_map(CorpusTranscript, removePunctuation)
    CorpusTranscript = tm_map(CorpusTranscript, removeWords, wordsToRemove)
    CorpusTranscript = tm_map(CorpusTranscript, removeWords, stopwords("engli
sh"))
    CorpusTranscript = DocumentTermMatrix(CorpusTranscript)
    CorpusTranscript = removeSparseTerms(CorpusTranscript, 0.97) # keeps a ma
trix 97% sparse

    return(CorpusTranscript)
}

Nword <- analyseText2(NegTweets$text)
findAssocs(Nword, c("flight", 'customer', 'gate', 'phone'), .07)
```

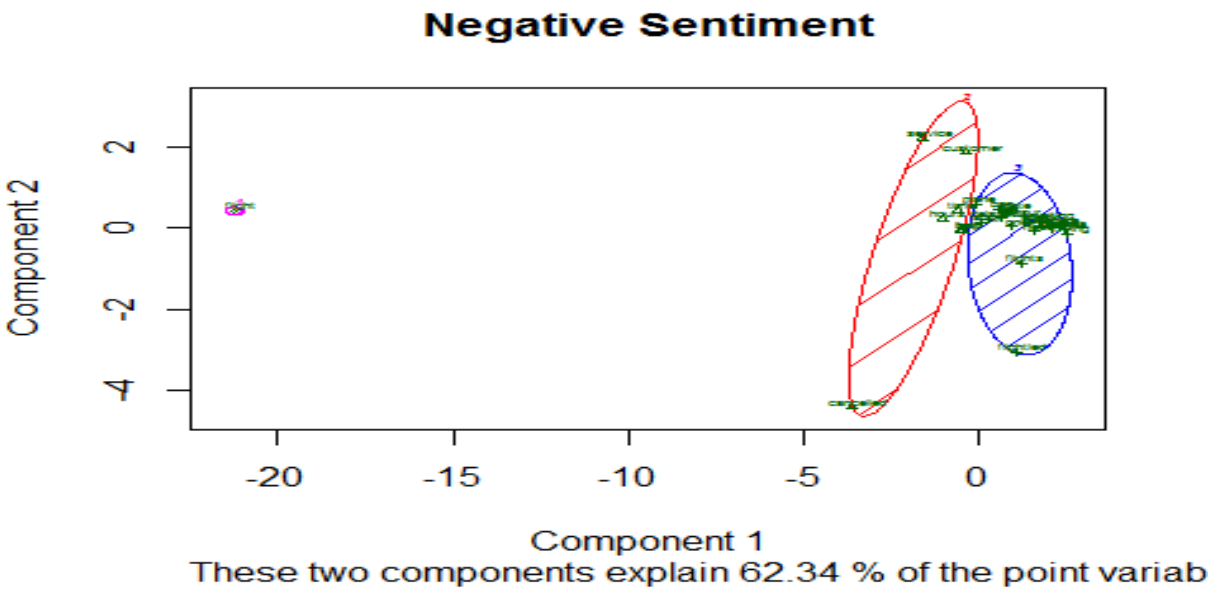
```
## $flight
## cancelled      late flightled   delayed
##      0.36      0.25      0.23      0.16
##
## $customer
## service
##    0.65
##
## $gate
## waiting   plane
##    0.09    0.08
##
## $phone
## help
## 0.07
```

Further understanding the association with Clustering Analysis

```
# hierarchical clustering
d = dist(t(as.matrix(Nword)), method = 'euclidean')
fit = hclust(d = d, method = 'ward.D')

# plotting the graph
plot(fit, col = "#487AA1", col.main = "#45ADA8", col.lab = "#7C8071", main =
'Negative Sentiment', xlab = '', col.axis = "#F38630", lwd = 3, lty = 3, sub
= "", hang = -1, axes = FALSE)

# add axis
axis(side = 2, at = seq(0, 400, 100), col = "#F38630", labels = FALSE, lwd =
2)
```




```
# k-mean clustering
d = dist(t(as.matrix(Nword)), method="euclidean")
kfit = kmeans(d, 3)
clusplot(as.matrix(d), kfit$cluster, color=T, shade=T, labels=2, lines=0, cex
= 0.4, main = 'Negative Sentiment')
```



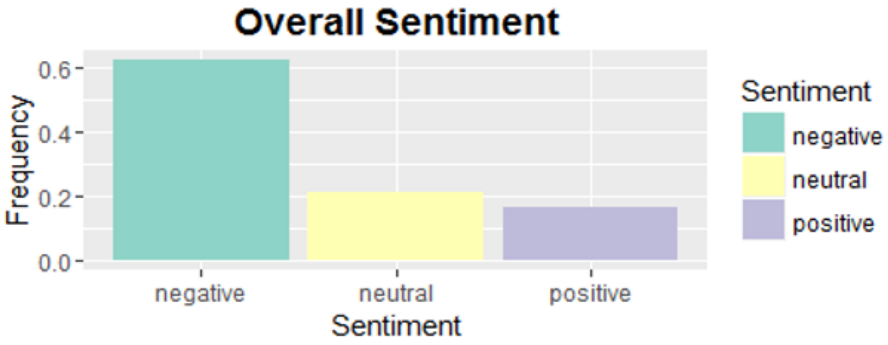
Results:

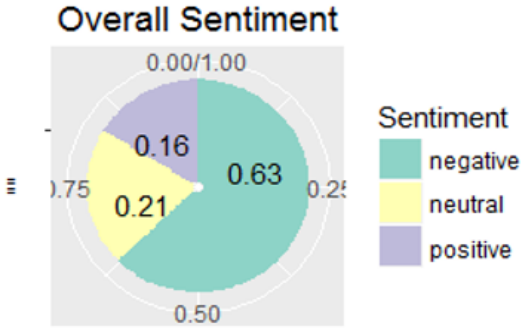
Exploratory data analysis: checking the columns containing no data (NAs):

```
##                                checkdata
## tweet_id                               0
## airline_sentiment                      0
## airline_sentiment_confidence           0
## negativereason                      5462
## negativereason_confidence           4118
## airline                                0
## airline_sentiment_gold             14600
## name                                   0
## negativereason_gold                14608
## retweet_count                          0
## text                                   0
## tweet_coord                        13621
## tweet_created                          0
## tweet_location                      4733
## user_timezone                       4820
```

We can see that columns like airline_sentiment_gold and negativereason_gold are mostly empty columns with NAs and have no information. Columns like negativereason, tweet_location and user_timezone has partial data.

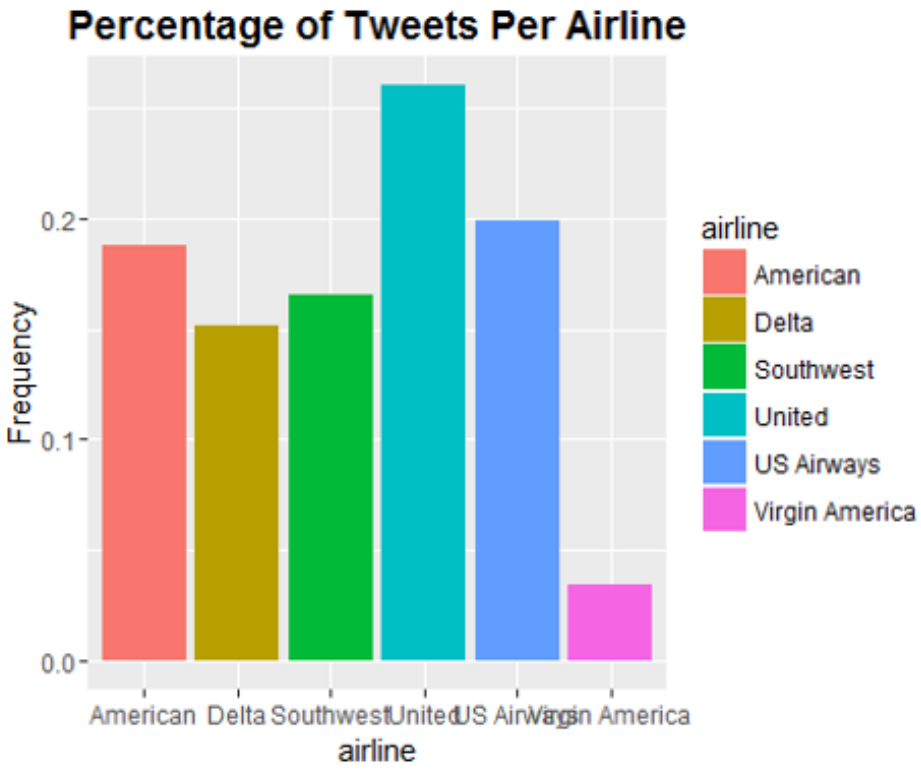
Sentiment Analysis:



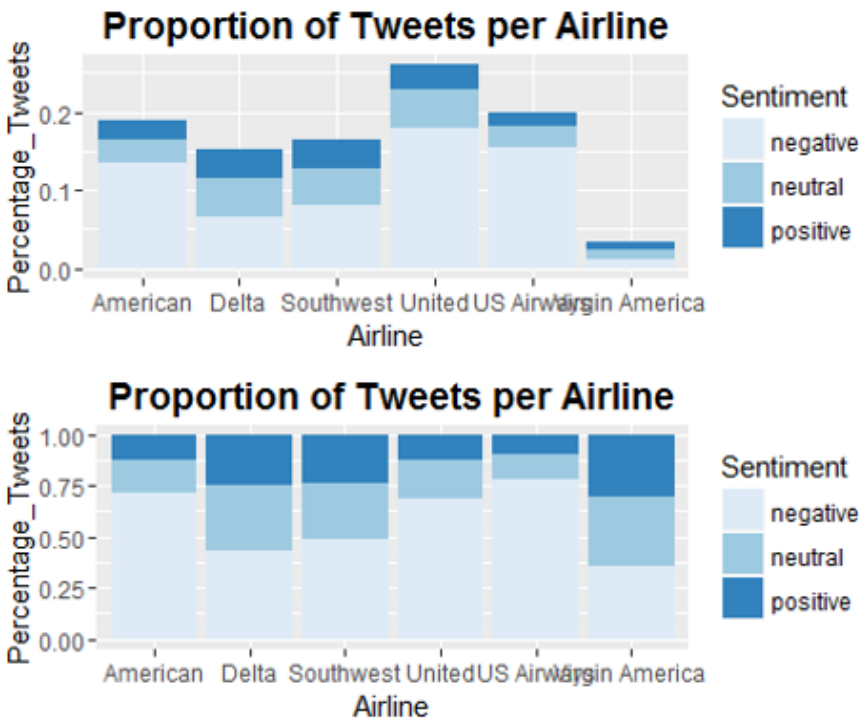


We can see that the bar plot and the pie graph that most tweets contain negative sentiment

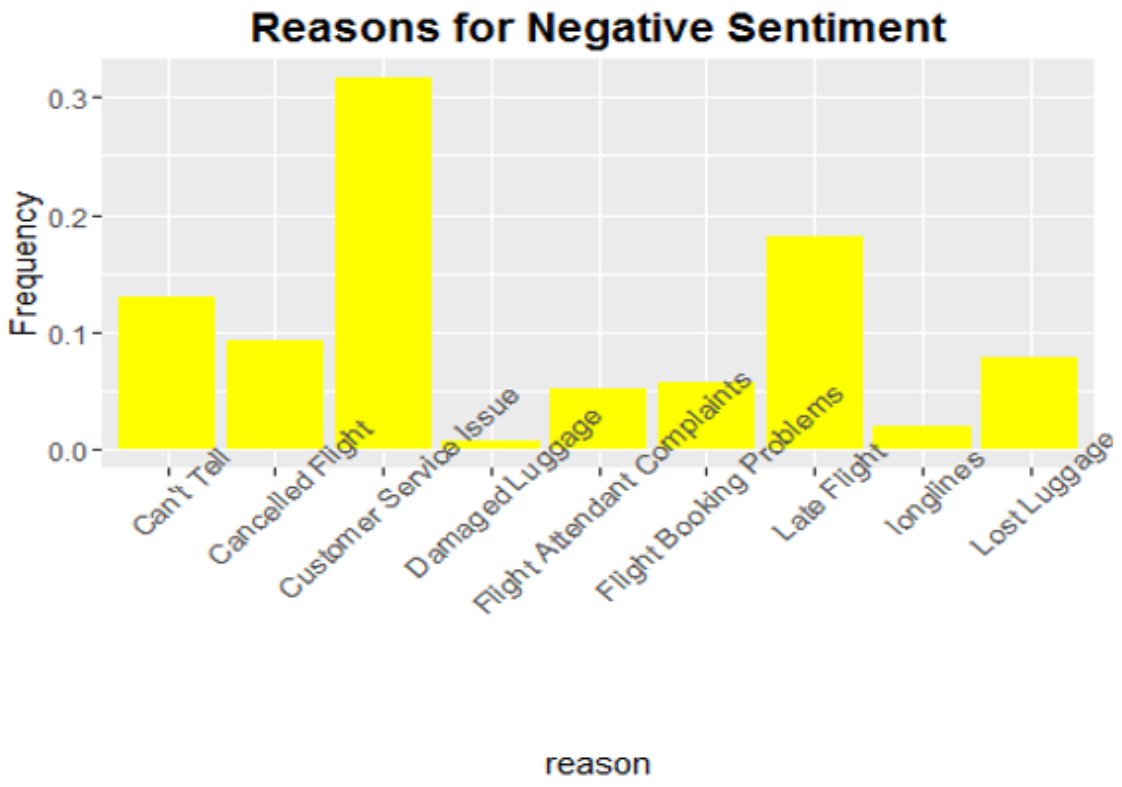
Percentage of Tweet Per Airline:



Most of the tweets are directed towards United Airlines, which are followed by American and US Airways. It can also be seen that very few are targeted towards Virgin America

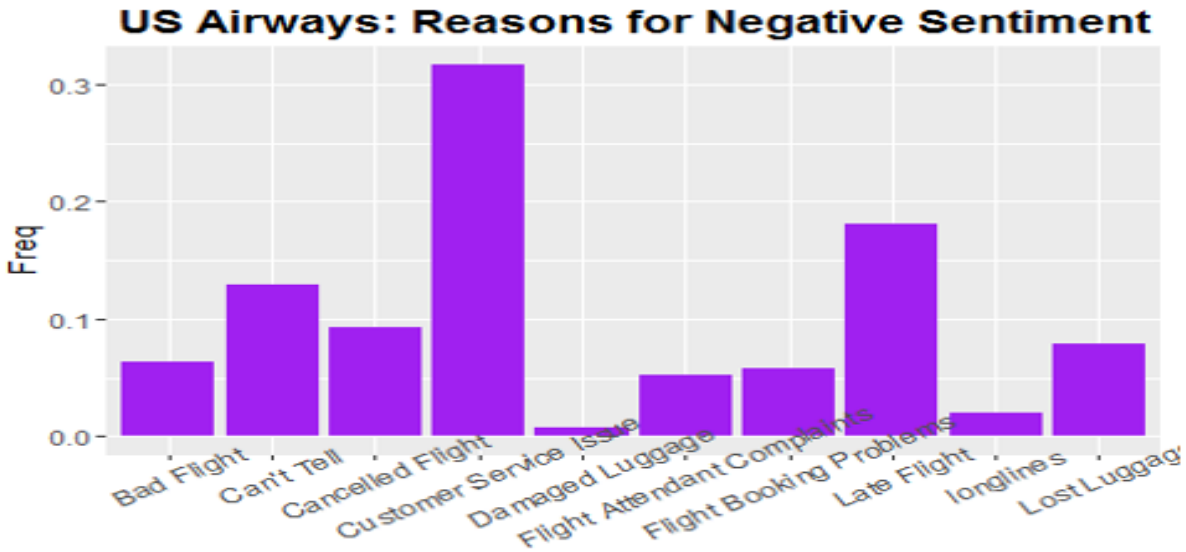


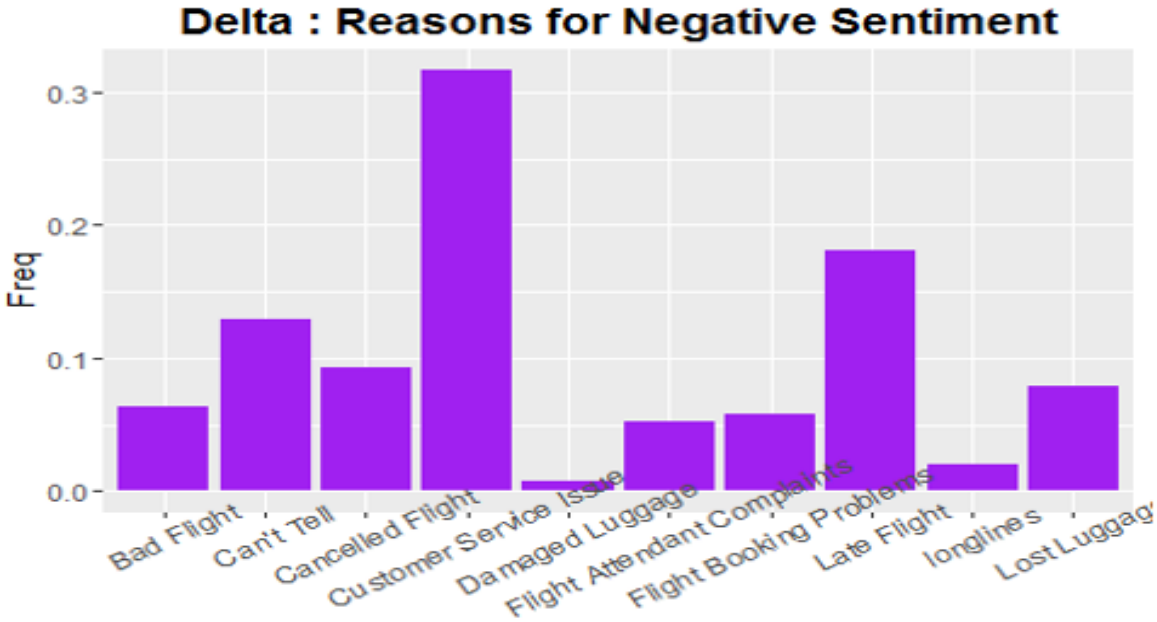
The above plots show proportion of tweets per Airline. The second plot is more informative, as it allows us to see the proportion of negative sentiment tweets per airline. We can see that the tweets directed towards American, United and US Airways are mostly negative. While those towards Delta, Southwest and Virgin have good proportions of all the sentiments

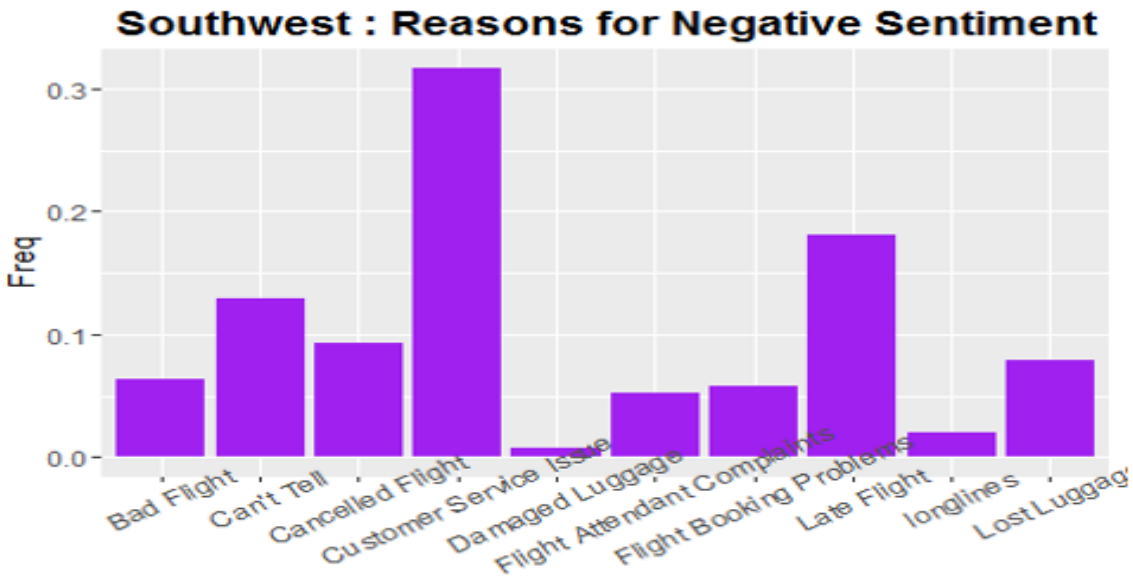


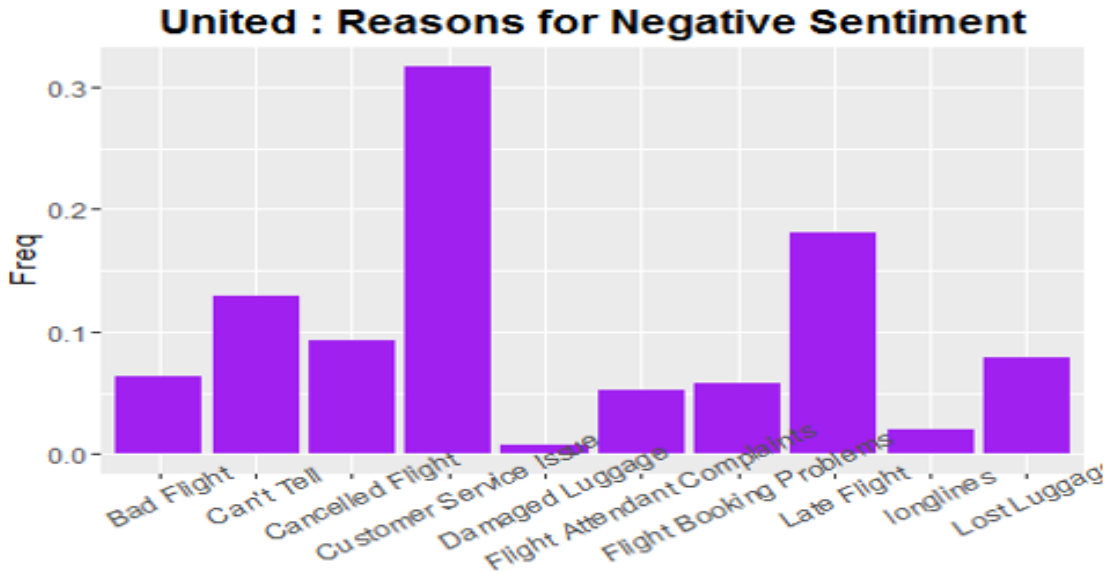
From the above we can infer that most of the negative tweets are elicited by Customer Services Issues like bad customer service. Also can be seen that late flights are a cause of Late Flight.

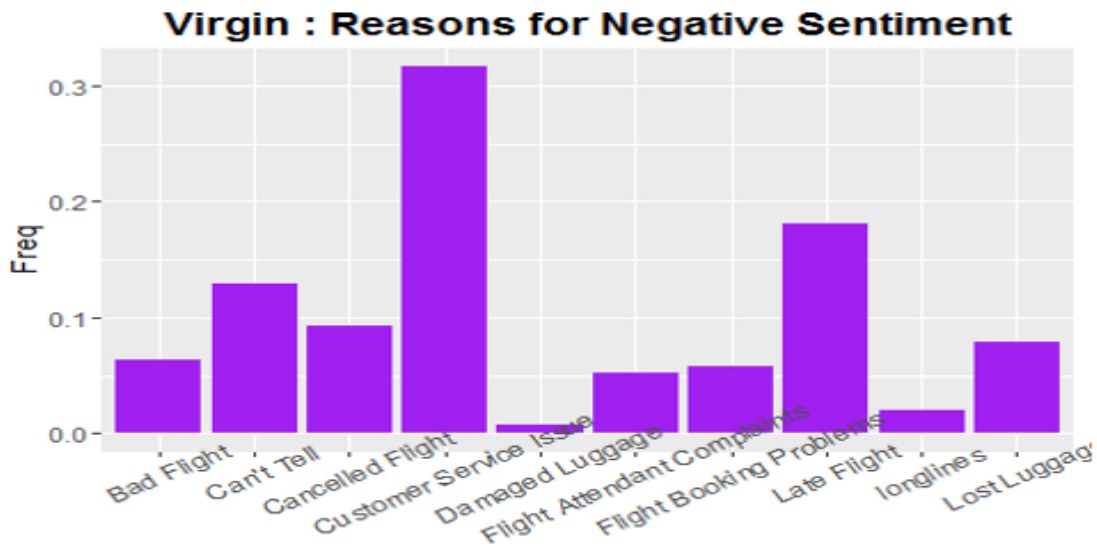
Below are the plots depicting the reasons for negative sentiments per airline:


US Airways: Reasons for Negative Sentiment
Freq
0.3
0.2
0.1
0.0
Bad Flight
Can't Tell
Cancelled Flight
Customer Service Issue
Damaged Luggage
Flight Attendant Complaints
Flight Booking Problems
Late Flight
longlines
Lost Luggag


Delta : Reasons for Negative Sentiment
Freq
0.3
0.2
0.1
0.0
Bad Flight
Can't Tell
Cancelled Flight
Customer Service Issue
Damaged Luggage
Flight Attendant Complaints
Flight Booking Problems
Late Flight
longlines
Lost Luggag


Southwest : Reasons for Negative Sentiment
Freq
0.3
0.2
0.1
0.0
Bad Flight
Can't Tell
Cancelled Flight
Customer Service Issue
Damaged Luggage
Flight Attendant Complaints
Flight Booking Problems
Late Flight
longlines
Lost Luggag


United : Reasons for Negative Sentiment
Freq
0.3
0.2
0.1
0.0
Bad Flight
Can't Tell
Cancelled Flight
Customer Service Issue
Damaged Luggage
Flight Attendant Complaints
Flight Booking Problems
Late Flight
longlines
Lost Luggag



We can observe in the above plots that American Airlines has most negative tweets on customer services related issues and hardly any because of late flights. Thus we can say that American flights were mostly on schedules. The same can be said about Southwest and Virgin airlines. Virgin airlines mostly gets negative tweets elicited because of their booking system.

US Airways and United both have a number of negative tweets because of Customer services and late flights.

On the contrary, we can see that Delta has negative tweets due to flight delays, though they seem to have a good customer services.

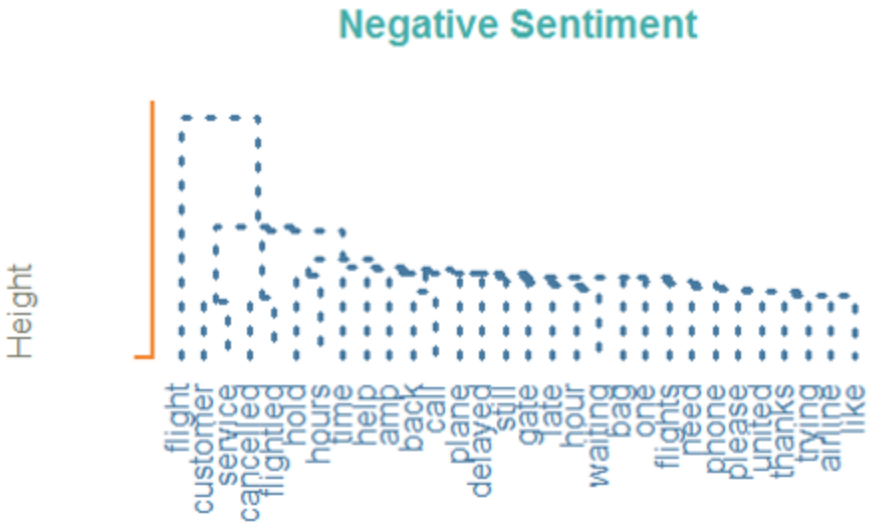
Analysis of text content of tweets:

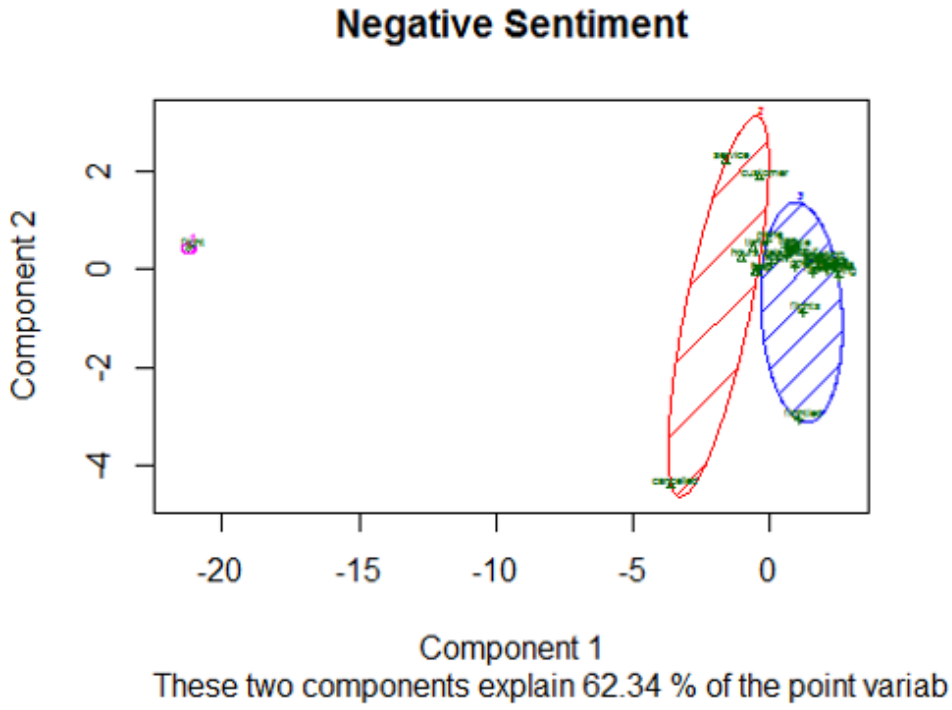
Word Cloud:

The word cloud provide a nice visual representation of the word frequencies for negative sentiment. The size of the word correlates with its frequency accross all tweets. We can get an idea of what people are talking about. For example, for negative sentiment, people seem to complain about cancelled or delayed flights, and hours waiting.

Understanding the association with Clustering Analysis:

- **hierarchical clustering**
- **k-mean clustering**


Negative Sentiment
Height
flight
customer
service
cancelled
flighted
hold
hours
time
help
amp
back
call
plane
delayed
still
gate
late
hour
waiting
bag
one
flights
need
phone
please
united
thanks
trying
airline
like


Negative Sentiment
Component 2
2
0
-2
-4
-20
-15
-10
-5
0
Component 1
These two components explain 62.34 % of the point variab

Through the above dendogram we observe the association of words like customer and service and flight canceled. The words that reflect complains are more generally like waiting, hours, bags, hold clustered together.

Through K-mean clustering we determine the proximity of words. The main aim of k-mean was to show that we can do more in terms of basic text analysis.

Discussion:

Through the Twitter data used for analyzing Sentiments using exploratory data analysis and Analysis of text content of tweets we could conclude that most of the tweets have negative sentiments, which was more than approx. 60% of the tweets. Most of the tweets were directed towards united airline followed by American and US Airways.

The main reason for the negative sentiments are Customer Service Issues and late flights. This could be seen in both the analyzing methods used.

As the data mainly deals with experiences of travelers, it was wise to do the exploratory analysis. However if one is to use APIs to collect data through twitter a wide range of data could be collected for analyzing, in which case clustering methods would be more approachable.

Though the above analysis does not show a broad usage of clustering, as twitter has a collection of wild ranging data would result in inaccurate data, thus it could be said that clustering is necessary to use in order to help discover data with similarities.

References:

ggmap: Spatial Visualization with ggplot2 by David Kahle and Hadley Wickham

http://shinyapps.org/apps/RGraphCompendium/index.php

http://www.cookbook-r.com/Graphs/Colors_(ggplot2)/

http://www.inside-r.org/howto/mining-twitter-airline-consumer-sentiment

http://www.cs.columbia.edu/~julia/papers/Agarwaletal11.pdf

http://andybromberg.com/sentiment-analysis/