Twitter US Airline Sentiment

Neha Parulekar

## 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\_\_\_","\_\_betrayal",..: 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 1 ...  
## $ retweet\_count : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ text : Factor w/ 14427 levels "\"LOL you guys are so on it\" - me, had this been 4 months ago...â<U+0080><U+009C>@JetBlue: Our fleet's on fleek. http://t.co/LYcARlTFHlâ<U+0080>",..: 14005 13912 13790 13844 13648 13926 14038 13917 14004 13846 ...  
## $ tweet\_coord : Factor w/ 833 levels "","[-33.87144962, 151.20821275]",..: 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

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.

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))

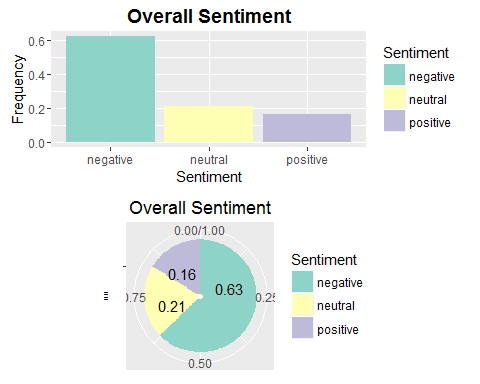
##   
## negative neutral positive   
## 0.6269126 0.2116803 0.1614071

We can see that most of the tweets contain negative sentiment, as given by the curators of the dataset.

# 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)) + scale\_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.title.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)

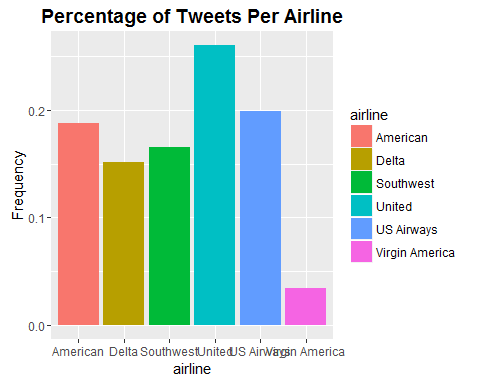


# Percentage of tweets per airline

# 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

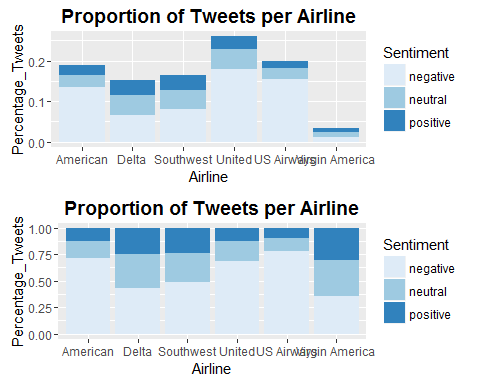
## airline Frequency  
## 1 American 0.18845628  
## 2 Delta 0.15177596  
## 3 Southwest 0.16530055  
## 4 United 0.26106557  
## 5 US Airways 0.19897541  
## 6 Virgin America 0.03442623

# 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('Proportion 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

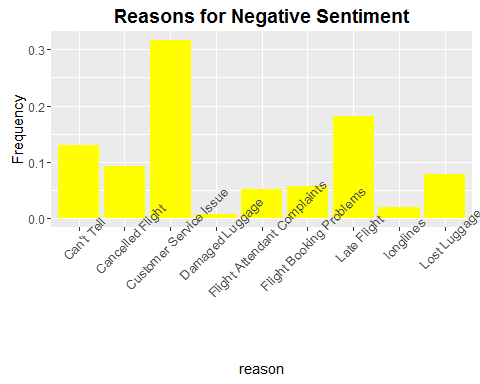
# Creating a data frame to get the reasons for negative tweets  
  
TweetDataFrame <- as.data.frame(prop.table(table(data$negativereason)))  
colnames(TweetDataFrame) <- c("reason", "Frequency")  
TweetDataFrame

## 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

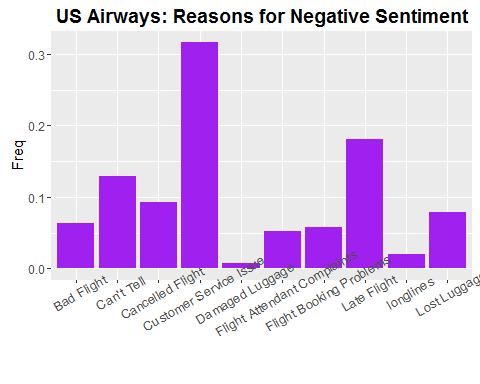
## reason Frequency  
## 2 Can't Tell 0.129657878  
## 3 Cancelled Flight 0.092285901  
## 4 Customer Service Issue 0.317062541  
## 5 Damaged Luggage 0.008062759  
## 6 Flight Attendant Complaints 0.052407932  
## 7 Flight Booking Problems 0.057637830  
## 8 Late Flight 0.181412072  
## 9 longlines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

NegReaplot <- ggplot(TweetDataFrame, aes(x = reason, y = Frequency)) + geom\_bar(stat = 'identity', fill = 7) + ggtitle('Reasons for Negative Sentiment') + theme(plot.title = element\_text(size = 14, face = 'bold', vjust = 1), axis.title.x = element\_text(vjust = -0.1), axis.text.x = element\_text(angle = 45, size = 10, vjust = 1))  
NegReaplot

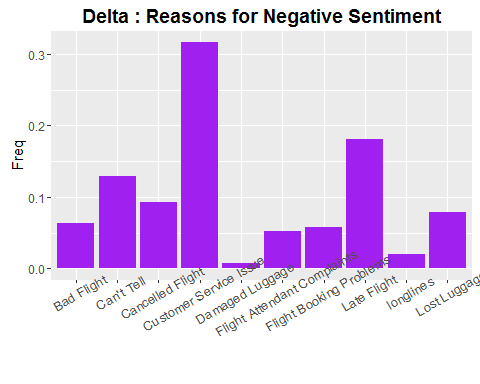


# Reasons For Negative Sentiment per airline

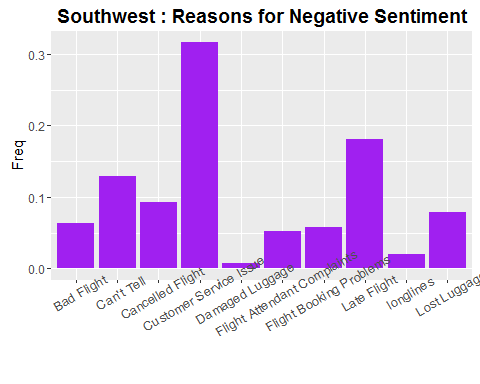
# First subset the data airline wise then plot to show the reason for negative sentiment  
  
# Subset the data for American Airlines  
AmericanAirline <- subset(data, airline = 'American')  
  
# Get the plot  
AAplot <- ggplot(as.data.frame(prop.table(table(AmericanAirline$negativereason))), aes(x = Var1, y = Freq)) + geom\_bar(stat = 'identity', fill = 'Purple') + ggtitle('American Airlines: Reasons for Negative Sentiment') + theme(plot.title = element\_text(size = 14, face = 'bold', vjust = 1), axis.title.x = element\_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') + ggtitle('US Airways: Reasons for Negative Sentiment') + theme(plot.title = element\_text(size = 14, face = 'bold', vjust = 1), axis.title.x = element\_blank(), axis.text.x = element\_text(angle = 30, size = 10, vjust = 1))  
USAplot



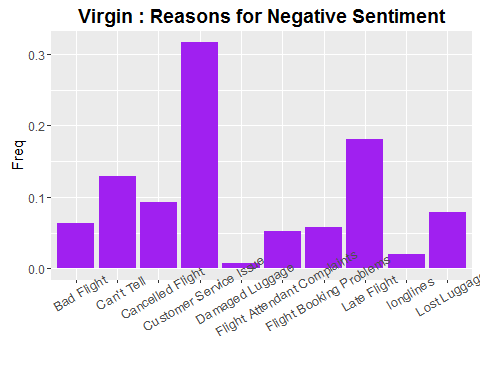
# Subset the data for Delta   
Delta <- subset(data, airline = 'Delta')  
  
# Generate the plot   
Delplot <- ggplot(as.data.frame(prop.table(table(Delta$negativereason))), aes(x = Var1, y = Freq)) + geom\_bar(stat = 'identity', fill = 'Purple') + ggtitle('Delta : Reasons for Negative Sentiment') + theme(plot.title = element\_text(size = 14, face = 'bold', vjust = 1), axis.title.x = element\_blank(), axis.text.x = element\_text(angle = 30, size = 10, vjust = 1))  
Delplot



# 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') + ggtitle('Southwest : Reasons for Negative Sentiment') + theme(plot.title = element\_text(size = 14, face = 'bold', vjust = 1), axis.title.x = element\_blank(), axis.text.x = element\_text(angle = 30, size = 10, vjust = 1))  
SWplot



# Subset the data for Virgin   
Virgin <- subset(data, airline = 'Virgin')  
  
# Generate the plot   
Virplot <- ggplot(as.data.frame(prop.table(table(Virgin$negativereason))), aes(x = Var1, y = Freq)) + geom\_bar(stat = 'identity', fill = 'Purple') + ggtitle('Virgin : Reasons for Negative Sentiment') + theme(plot.title = element\_text(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') + ggtitle('United : Reasons for Negative Sentiment') + theme(plot.title = element\_text(size = 14, face = 'bold', vjust = 1), axis.title.x = element\_blank(), axis.text.x = element\_text(angle = 30, size = 10, vjust = 1))  
Unplot

