Project Topic

“This study particularly aims at using text mining techniques to analyze the positive and negative tweets of six major US airlines and examine any possible correlations between negative sentiments expressed by the users to draw inferences about customer perception of airlines performance and service levels”

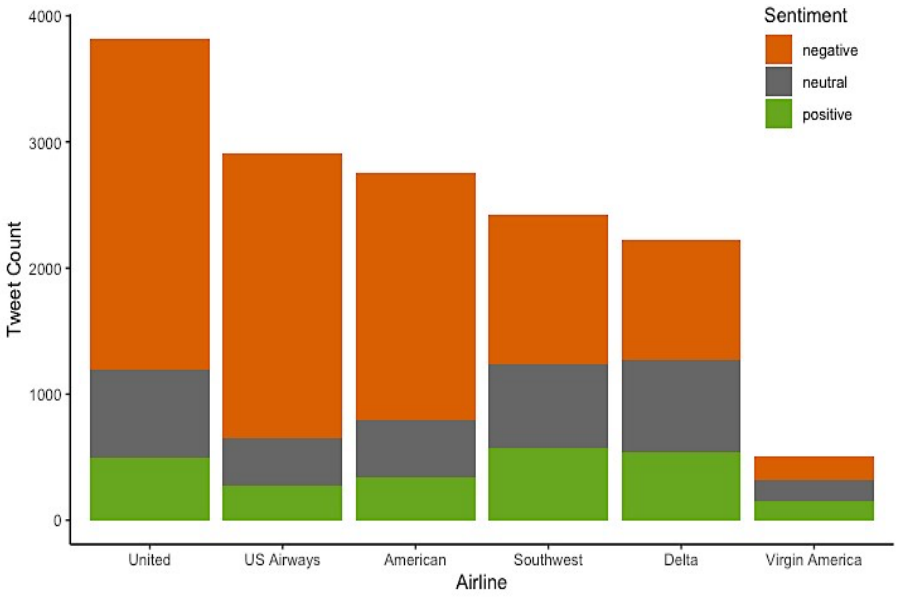
Null hypothesis: That is no relationship between the word frequencies of complaints (negative sentiments) between the airlines

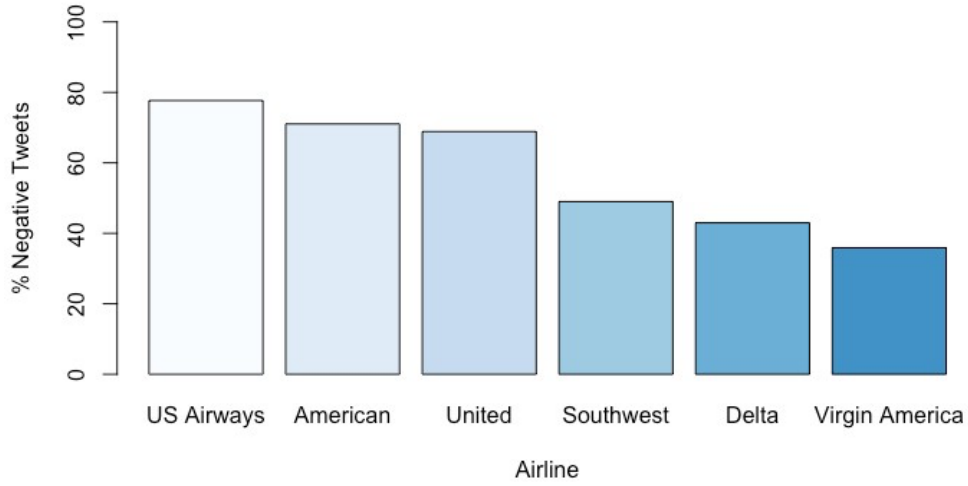
Alternative hypothesis: There is significant relationship between the sets of word frequencies between airlines

Data Description

- **Kaggle dataset**: A slight variation of the original source *(Crowdflower's Data for Everyone library)* comprised of Twitter data scrapped from February 2015
- 14,640 tweets of 6 major airlines in the United States with 15 variables
- One of the variables in the dataset contains the predicted tweet sentiment (*positive, neutral, negative*)
- 3 variables were excluded completely (*negativereason_gold, airline_sentiment_gold and tweet_coordinates*) as they had more than 90% missing data and do not provide any additional insights to our topic
- Data screened for accuracy, missing values & outliers in the key variables pertinent to our analysis in the dataset

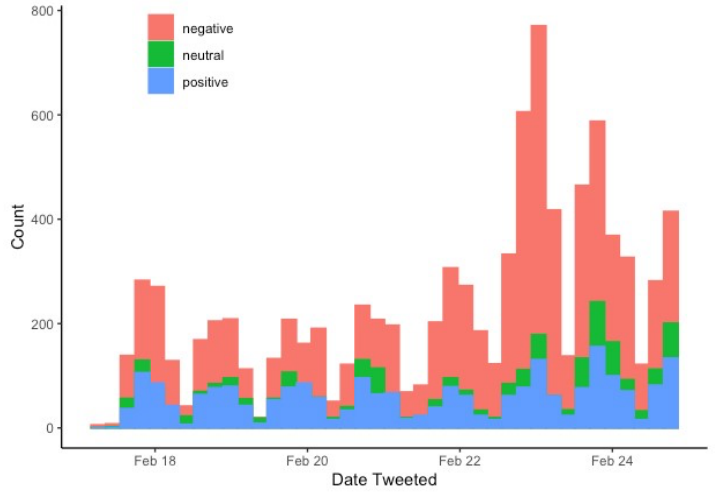
Exploratory Data Analysis



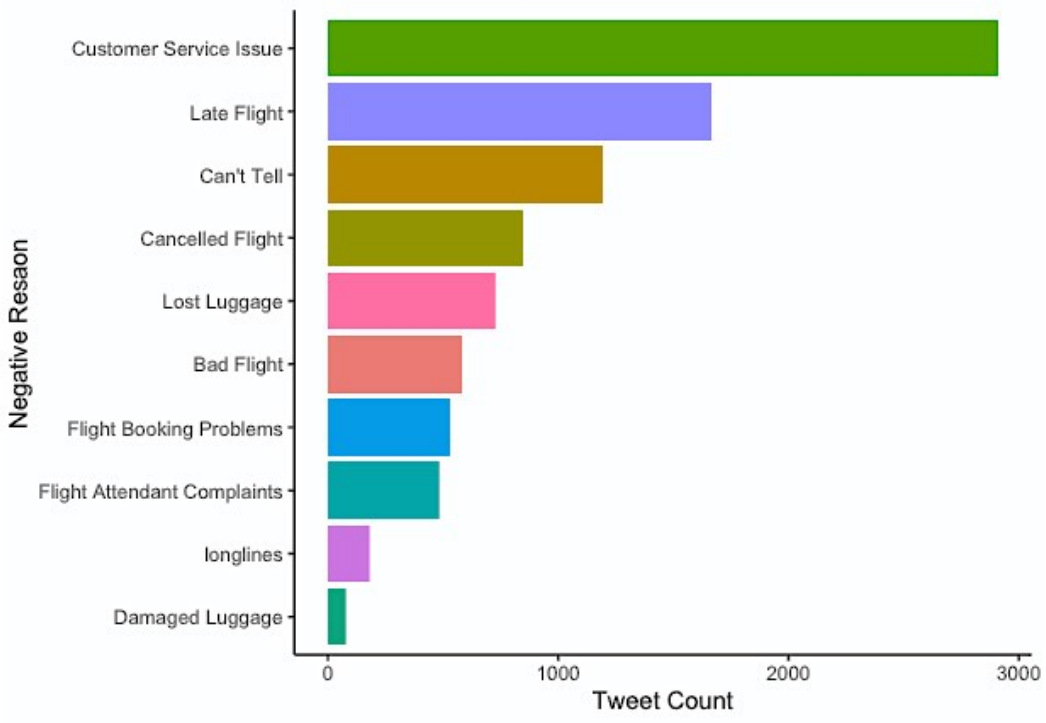


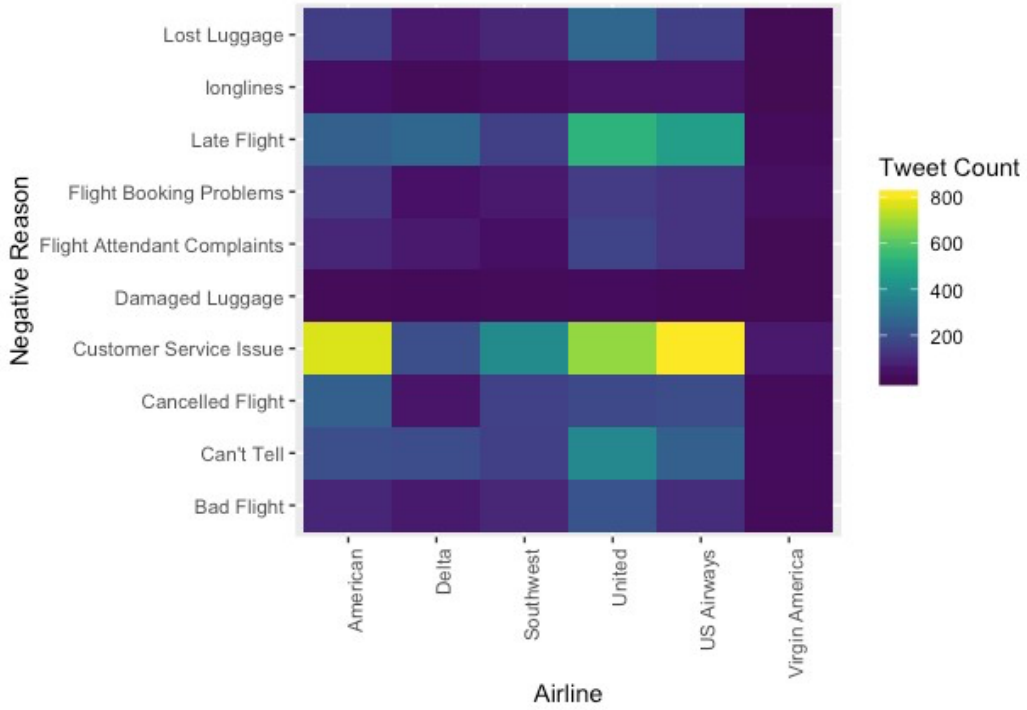
- 63% of the tweets expressed negative sentiments overall
- Although United has the most negative & total tweets in absolute count but drops to 3rd spot in % negative to overall count.
- Surprisingly, Delta seems to be better in comparison to its legacy competitors with 43% negative tweets and twice as much %positive tweets as its competitors.

- Couldn’t find any attributable reason for the spikes between Feb 23 – 26 and is assumed to be random



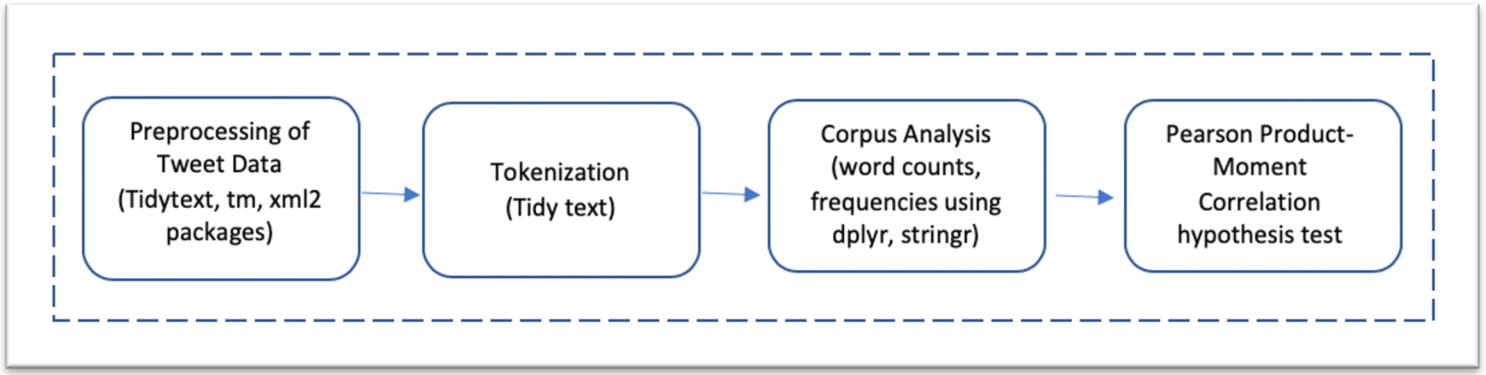
Exploratory Data Analysis





- "Customer Service Issue" *(topped the list of negative sentiment features with 32%)* along with late flights and cancellations together accounted for 60% of the reasons
- Interestingly, Delta being the World's second largest in terms of passengers carried and fleet size seems to have fewer cancellations related complaints and better flight experience

Problem Methodology



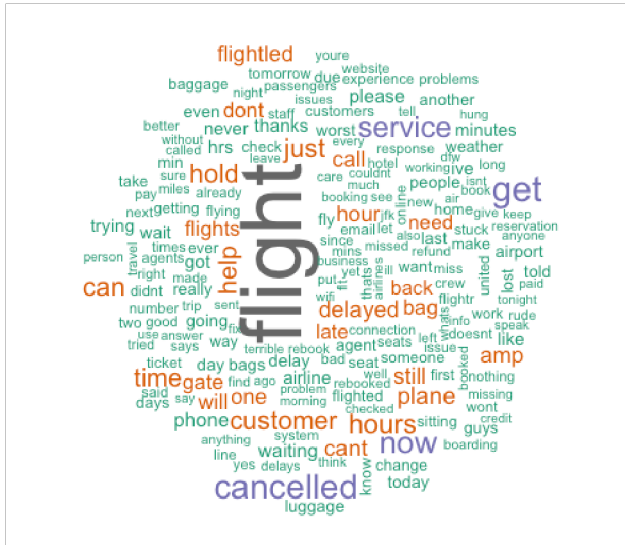

```
#Preprocessing the text data
library(tidytext)
library(xml2)
library(tm)
library(wordcloud)
TwCorpus <- Corpus(VectorSource(TwtRaw$text))
TwCorpus <- tm_map(TwCorpus, content_transformer(tolower))
removeMentions <- function(x) gsub("@\\w+","",x)
TwCorpus <- tm_map(TwCorpus, removeMentions)
removeHastag <- function(x) gsub("#\\w+","",x)
TwCorpus <- tm_map(TwCorpus, removeHastag)
removeUrl <- function(x) gsub("http\\w+","",x)
TwCorpus <- tm_map(TwCorpus, removeUrl)
TwCorpus <- tm_map(TwCorpus, function(x)removePunctuation(x, preserve_intra_word
ntractions = FALSE))
TwCorpus <- tm_map(TwCorpus, content_transformer(removeNumbers))
TwCorpus <- tm_map(TwCorpus, removeWords, stopwords("english"))
Clean_Text <- data.frame(text = get("content", TwCorpus))
names(Clean_Text) <- "Clean_Text"
CleanTwt <- cbind(TwtRaw,Clean_Text)
```

Preprocessing of tweets

```
# A tibble: 9,023 x 2
   Issues        n
   <chr>     <int>
 1 flight     2900
 2 get         984
 3 cancelled   913
 4 now         824
 5 service     741
 6 hours       653
 7 can         625
 8 just        617
 9 hold        608
10 customer    603
# … with 9,013 more rows
```

```
# A tibble: 65,921 x 2
   Issues                n
   <chr>             <int>
 1 customer service    444
 2 cancelled flightled 437
 3 late flight         217
 4 cancelled flighted  195
 5 flight cancelled    194
 6 late flightr        142
 7 cancelled flight    122
 8 flight delayed      114
 9 cant get            106
10 booking problems     97
# … with 65,911 more rows
```

Tokenization & Corpus Analysis



Negative & positive sentiment word cloud (visualization)

Pearson Product-Moment Correlation Results

American Airlines vs. Southwest

```
        Pearson's product-moment correlation

data:  ratio and Southwest
t = 227.43, df = 7854, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9287824 0.9346152
sample estimates:
      cor
0.9317589
```

United vs. Southwest

```
        Pearson's product-moment correlation

data:  ratio and Southwest
t = 178.11, df = 7854, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8908148 0.8995952
sample estimates:
      cor
0.8952919
```

Delta vs. Southwest

```
        Pearson's product-moment correlation

data:  ratio and Southwest
t = 153.93, df = 7854, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8610151 0.8720297
sample estimates:
     cor
0.866628
```

Virgin Airlines vs. Southwest

```
        Pearson's product-moment correlation

data:  ratio and Southwest
t = 111.38, df = 7854, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7738003 0.7909504
sample estimates:
      cor
0.7825238
```

For each of the correlations between Southwest (reference) and the rest of the airlines, namely United, Delta, American and Virgin respectively, the p-value of the correlation test is < 0.01. Hence, we reject the null-hypothesis and accept the alternate hypothesis that there is relationship between the sets of word frequencies between the airlines as far as customer complaints (negative sentiments) are concerned. American Airlines (Correlation Coefficient of 0.93 and Coefficient of determination of 0.87) is more correlated than United, Delta and Virgin Airlines (in the decreasing order of correlation)

Conclusion

- The study does show evidence that "**customer service**" seems to be the single major issue *across the commercial airlines industry* from the customer's perspective along with "**schedule adherence**" (delays, cancellations, boarding)
- This information can be used to support new initiatives & better loyalty programs by the commercial airline companies to better position themselves in the industry and gain competitive advantage
- Scope of the study can be further advanced by deep diving into positive or neutral sentiments too as they may provide existing solutions to current state problems
- Also, we could analyze tweets from different timeframes and / or adopt different machine learning techniques for improved text predictions to expand the scope