Twitter US Airline Sentiment Analysis

Suma Gummadapu

29/06/2022

Summary

The customer of airline services is always provided with a platform on Twitter where they can tweet any views or opinions regarding the services ad traveling experience on a particular flight. This makes Twitter contain a large amount of data and information related to the services that airlines provide to their customers. To track customers' satisfaction, there is always the need to explore sentiments from tweets made by customers about the airline services. Therefore, this project's main objective is to analyze the US Airline twitter dataset to find the best and worst airline services and make predictions on the most common issues that have occurred during the airline services. The project creates a word cloud of negative, positive, and neutral sentiments from tweets and a comparison word cloud.

Introduction

For airline companies, customer feedback concerning their services and experience is essential. Recently, the popularity of Twitter sentiment analysis has increased as it is possible to analyze customer satisfaction for online services. Therefore, this project includes the analysis of the US Twitter airline data set, and the main problems that have been experienced during airline services have also been predicted.

Literature Review

In their study, Mohan & Venu (2016) plotted a sentiment mining graph from which sentiment analysis was performed in a trip review, movie review, social discussion, and product review form. The analysis involved making a difference between computers and humans. Sentiment analysis in this study was based on neutral, negative, and positive commands. In the study by Da Silva et al. (2014), an ensemble sentiment classification system was developed to perform analysis for Twitter data airline services. The authors also utilized a lexicon-based approach through a lexicon dictionary. With lexicon, it is possible to calculate the sum of positive and negative sentiment words appearing in a text file. In a project by Hakh et al. (2017), the authors indicate that there is a rapid growth in online media, hence creating a need for sentiment analysis. With sentiment analysis, data analysts and data scientists can extract customer feedback and reviews and identify what customers like and dislike about a given product. Sentiment analysis is also essential as the reactions, feedback, and opinions towards a specific product of people can be extracted. In another study, Prabhakar & Sugashini (2018) indicated that there are two types of sentiments: lexicon-based and learning sentiment. Lexicon sentiment uses a lexicon dictionary and can collect feedback from airline companies' customers.

Theory

H1: Delays in flights lead to negative sentiments from the airline customers.

Data

The dataset has been retrieved from <https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment/download> and it contains:

## tweet\_id airline\_sentiment airline\_sentiment\_confidence negativereason

## 1 5.703061e+17 neutral 1.0000

## 2 5.703011e+17 positive 0.3486

## 3 5.703011e+17 neutral 0.6837

## 4 5.703010e+17 negative 1.0000 Bad Flight

## 5 5.703008e+17 negative 1.0000 Can't Tell

## 6 5.703008e+17 negative 1.0000 Can't Tell

## negativereason\_confidence airline airline\_sentiment\_gold name

## 1 NA Virgin America cairdin

## 2 0.0000 Virgin America jnardino

## 3 NA Virgin America yvonnalynn

## 4 0.7033 Virgin America jnardino

## 5 1.0000 Virgin America jnardino

## 6 0.6842 Virgin America jnardino

## negativereason\_gold retweet\_count

## 1 0

## 2 0

## 3 0

## 4 0

## 5 0

## 6 0

## text

## 1 @VirginAmerica What @dhepburn said.

## 2 @VirginAmerica plus you've added commercials to the experience... tacky.

## 3 @VirginAmerica I didn't today... Must mean I need to take another trip!

## 4 @VirginAmerica it's really aggressive to blast obnoxious "entertainment" in your guests' faces &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 have this playing.\nit's really the only bad thing about flying VA

## tweet\_coord tweet\_created tweet\_location

## 1 2015-02-24 11:35:52 -0800

## 2 2015-02-24 11:15:59 -0800

## 3 2015-02-24 11:15:48 -0800 Lets Play

## 4 2015-02-24 11:15:36 -0800

## 5 2015-02-24 11:14:45 -0800

## 6 2015-02-24 11:14:33 -0800

## user\_timezone

## 1 Eastern Time (US & Canada)

## 2 Pacific Time (US & Canada)

## 3 Central Time (US & Canada)

## 4 Pacific Time (US & Canada)

## 5 Pacific Time (US & Canada)

## 6 Pacific Time (US & Canada)

Display the names of the variables in the dataset.

names(airline\_tweets)

## [1] "tweet\_id" "airline\_sentiment"

## [3] "airline\_sentiment\_confidence" "negativereason"

## [5] "negativereason\_confidence" "airline"

## [7] "airline\_sentiment\_gold" "name"

## [9] "negativereason\_gold" "retweet\_count"

## [11] "text" "tweet\_coord"

## [13] "tweet\_created" "tweet\_location"

## [15] "user\_timezone"

Identify if there are any missing values in the dataset.

colSums(is.na(airline\_tweets))

## tweet\_id airline\_sentiment

## 0 0

## airline\_sentiment\_confidence negativereason

## 0 0

## negativereason\_confidence airline

## 4118 0

## airline\_sentiment\_gold name

## 0 0

## negativereason\_gold retweet\_count

## 0 0

## text tweet\_coord

## 0 0

## tweet\_created tweet\_location

## 0 0

## user\_timezone

## 0

There are 4118 missing values which have been removed as shown below.

airline\_tweets <- airline\_tweets[complete.cases(airline\_tweets), ]

All the rows with missing values have been removed

sum(is.na(airline\_tweets))

## [1] 0

Change the format of the tweet\_created column to Date format and name it date as a new column in the dataset.

airline\_tweets$date <- as.Date(airline\_tweets$tweet\_created)

From the text column which contained the tweets by customers, we removed stopwords, removed numbers, and removed any capitalisation.

docs <- Corpus(VectorSource(airline\_tweets$text))

docs <- tm\_map(docs, content\_transformer(tolower))

## Warning in tm\_map.SimpleCorpus(docs, content\_transformer(tolower)):

## transformation drops documents

docs <- tm\_map(docs, removeNumbers)

## Warning in tm\_map.SimpleCorpus(docs, removeNumbers): transformation drops

## documents

docs <- tm\_map(docs, removeWords, stopwords("english"))

## Warning in tm\_map.SimpleCorpus(docs, removeWords, stopwords("english")):

## transformation drops documents

docs <- tm\_map(docs, removeWords, c("usairways" ,"united", "flight" , "americanair" , "jetblue" , "southwestair"))

## Warning in tm\_map.SimpleCorpus(docs, removeWords, c("usairways", "united", :

## transformation drops documents

docs <- tm\_map(docs, removePunctuation)

## Warning in tm\_map.SimpleCorpus(docs, removePunctuation): transformation drops

## documents

docs <- tm\_map(docs, stripWhitespace)

## Warning in tm\_map.SimpleCorpus(docs, stripWhitespace): transformation drops

## documents

Next, Term Document Matrix has been created.

dtm <- TermDocumentMatrix(docs)

m <- as.matrix(dtm)

v <- sort(rowSums(m),decreasing=TRUE)

d <- data.frame(word = names(v),freq=v)

head(d, 10)

## word freq

## get get 1117

## cancelled cancelled 989

## now now 880

## can can 789

## service service 786

## just just 709

## help help 689

## hours hours 667

## time time 648

## customer customer 642

Methodology

Since we have a clean dataset, we will now analyze the distribution of sentiments for the airline dataset. First, we get to find the positive or negative distribution of the airline sentiment.

Pos\_Neg = airline\_tweets %>% group\_by(airline\_sentiment) %>% dplyr::summarise(count = n())

Pos\_Neg

## # A tibble: 3 × 2

## airline\_sentiment count

## <chr> <int>

## 1 negative 9178

## 2 neutral 1014

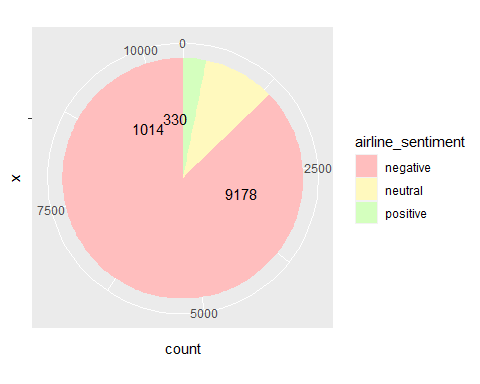
## 3 positive 330

Use a pie chart to visualize the distribution identified above.

ggplot(data=Pos\_Neg , aes(x="" , y=count , fill=airline\_sentiment))+geom\_bar(width=1,stat="identity")+

geom\_text(aes(y = count/3 + c(0, cumsum(count)[-length(count)]),

label =count), size=4)+coord\_polar("y") + scale\_fill\_manual(values=c("#ffbebe", "#fff9be", "#d4ffbe"))

 Next, we find the distribution of positive or negative sentiments by airlines.

## Using date as value column: use value.var to override.

## Aggregation function missing: defaulting to length

airlines\_Pos\_Neg$negPer = airlines\_Pos\_Neg$negative / (airlines\_Pos\_Neg$negative + airlines\_Pos\_Neg$positive + airlines\_Pos\_Neg$neutral)

airlines\_Pos\_Neg = airlines\_Pos\_Neg[order(-airlines\_Pos\_Neg$negPer),]

airlines\_Pos\_Neg

## airline negative neutral positive negPer

## 5 US Airways 2263 154 53 0.9161943

## 1 American 1960 179 51 0.8949772

## 4 United 2633 274 93 0.8776667

## 3 Southwest 1186 191 68 0.8207612

## 2 Delta 955 174 55 0.8065878

## 6 Virgin America 181 42 10 0.7768240

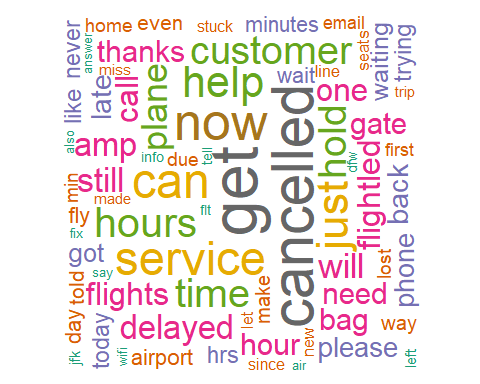
To identify the most used words, a wordcloud has been created.

options(warn=-1)

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"))



Results

From the above performed analysis, there are some observations that were made. The distribution of the positive and negative airline sentiments indicate that more than half of the tweets in the dataset are negative.

## # A tibble: 3 × 2

## airline\_sentiment count

## <chr> <int>

## 1 negative 9178

## 2 neutral 1014

## 3 positive 330

The observations also made from the distribution of positive and negative sentiments by airlines illustrated that more than 50% of negative tweets for United, American, and US Airways.

## airline negative neutral positive negPer

## 5 US Airways 2263 154 53 0.9161943

## 1 American 1960 179 51 0.8949772

## 4 United 2633 274 93 0.8776667

## 3 Southwest 1186 191 68 0.8207612

## 2 Delta 955 174 55 0.8065878

## 6 Virgin America 181 42 10 0.7768240

The observations for the reasons for negative tweets for the airline services indicate that late flights and customer service issues are the main reasons for negative tweets for most airlines.

Neg\_Tweets <- airline\_tweets %>% filter(airline\_sentiment=="negative")

Neg\_tweets\_reason\_airlines <- Neg\_Tweets %>% group\_by(airline,negativereason) %>% dplyr::summarise(count=n())%>% arrange(airline,desc(count))

## `summarise()` has grouped output by 'airline'. You can override using the

## `.groups` argument.

Neg\_tweets\_reason\_airlines

## # A tibble: 60 × 3

## # Groups: airline [6]

## airline negativereason count

## <chr> <chr> <int>

## 1 American Customer Service Issue 768

## 2 American Late Flight 249

## 3 American Cancelled Flight 246

## 4 American Can't Tell 198

## 5 American Lost Luggage 149

## 6 American Flight Booking Problems 130

## 7 American Bad Flight 87

## 8 American Flight Attendant Complaints 87

## 9 American longlines 34

## 10 American Damaged Luggage 12

## # … with 50 more rows

Implications

This project did not involve full analysis, and there are areas of improvement for future research. These areas include the researchers should consider performing text mining and identifying the directions of flights. This is because the customer complaints may be from specific airports. Another area of improvement is putting into consideration the type of traveler.

**Conclusion**

The theory that flight delays lead to airline customers' negative sentiments is true. Late flights were identified as a major factor contributing to the customers' negative sentiments.

References

Mohan, V., & Venu, S. H. (2016). Sentiment Analysis Applied to Airline Feedback to Boost Customersâ€™ Endearment. International Journal of Applied and Physical Sciences, 2(2), 51-58.

Hakh, H., Aljarah, I., & Al-Shboul, B. (2017). Online social media-based sentiment analysis for us airline companies. New Trends in Information Technology, 176.

Prabhakar, E., & Sugashini, K. (2018). New Ensemble Approach to Analyze User Sentiments from Social Media Twitter Data. The SIJ Transactions on Industrial, Financial & Business Management (IFBM), 6(1), 7-11.