Fundamentals of Data Science Course Project Paper

On

TWITTER AIRLINE SENTIMENT ANALYSIS

Department of Information Technology

Submitted By:-

1. Abhinav Singh 169005150

Project documentation submitted to fulfill the requirements for the class of Fundamentals of Data Science and in Partial fulfillment of the requirements for Master's Degree in Information Technology.

Date: - December 11, 2016

Session: - Fall '16

INTRODUCTION

1.1 Motivation

Airline service companies must interpret a substantial amount of customer feedback about their products and services. However, conventional methods to collect customers' feedback for airline service companies is to investigate through distributing and collecting questionnaires, which is time consuming and inaccurate. It needs labour to distribute and collect questionnaires to customers and also it will take too much effort to record and file those questionnaires considering how many passengers take flights every day. Beyond that, not all customers take questionnaires seriously and many customers just fill them in randomly and all of this brings noisy data into sentiment analysis. Unlike investigation questionnaires, twitter is a much better data source for sentiment classification for feedbacks of airline services. Because of the Big Data technologies, it has become very easy to collect millions of tweets and implement data analysis on them. This has saved a lot of labour costs which questionnaire investigations need. More than that, people post their genuine feelings on Twitter, which makes the information more accurate than investigation questionnaires. The other limitations for questionnaire investigations are that the questions on questionnaires are all set and it is hard to reveal the information which questionnaires do not cover.

As a result, sentiment analysis has become very popular in recent years for automatic customer satisfaction analysis of online services. Sentiment analysis is a sub domain of data mining, which are exploited to analyse large-scale data to reveal hidden information. Obviously, the advantages of automatic analysis of massive datasets make sentiment analysis preferable for airline companies.

Sentiment classification techniques can help researchers and decision makers in airline companies better understand customer feedback and satisfaction. Researchers and decision makers can utilize these techniques to automatically classify customers' 2 feedback on micro-blogging platforms like Twitter. Business analysis applications can be developed from these techniques as well.

There have been much research on text classification and sentiment classification, but there has been little on Twitter sentiment classification about airline services. Except applying popular sentiment classification approaches to tweets on airline services domain, it is also desirable to develop a new approach to further improve the classification accuracy.

1.2 Research Objectives

Twitter is a really good source to get customers' feedback and marketing information in airline services, but there has been no perfect solution to automatically classify the massive amount of tweets, which leaves room for doing research in this area. A sentiment

analysis job about the problems of each major U.S. airline. Planning on scraping Twitter data for a random month of 2015 and to classify positive, negative, and neutral tweets, followed by categorizing negative reasons (such as "late flight" or "rude service").

I have answered the questions:

- Text analysis of the user tweets to find out the reasons behind the user's sentiments.
- Find out which airlines which provide best and worst customer satisfaction.
- Get the most discussed topics among various airlines.
- Reasons for negative sentiment (and by airlines).
- Top negative and positive words.

For implementation purposes I used R and Python (Text classification) to infer about the dataset.

1.3 Challenges and Related Work on Twitter Data about

The challenges in twitter sentiment classification not only come from the fact that each post is not allowed to exceed 140 characters but also because the sentiment of the tweets can be very dependent on the scenarios the users are involved in but the context of the scenarios is not provided in the tweets. For example, "Cancelled again, It's the fourth time" can be a tweet with negative sentiment if it is about taking flights but also can be a neutral sentiment tweet if it is talking about the user frequently cancelling some subscriptions. Because of this, Twitter sentiment classifications are very domain dependent.

In sentiment classification, features are important because they are the attributes that determine texts' sentiments (Pang and Lee 2008). Features can be unigrams which are words, or N-grams. Twitter sentiment classifications are domain dependent because those features are domain dependent, and sentiment features in one domain may not be sentiment features in other domains at all. For example, in the stock market area, the word "bear" means negative sentiment since it is a term describing bad performances in the stock market but it means no sentiment at all in most other domains. So the unigram "bear" can be extracted as a feature in the stock market area but not in other areas such as airline services.

In the handbook of "Mining Twitter for Airline Consumer Sentiment", Jeffery Oliver illustrates classifying tweets sentiment by applying sentimental lexicons (Oliver 2012). This handbook suggests retrieving real time tweets from Twitter API with queries containing airline companies' names. The sentiment lexicons in this method are not domain specific and there is no data training process or testing process. By matching each tweet with the positive word list and the negative word list, and assigning scores based on matching result to each tweet, they can be classified as positive or negative according to the summed scores. The accuracy is unknown since it is not considered in this book.

Adeborna et al. adopted Correlated Topics Models (CTM) with Variational ExpectationMaximization (VEM) algorithm (Adeborna and Siau 2014). Their lexicons for classification were developed with AQR criteria. In Sentiment detection process, the performances of the SVM classifier, the Maximum Entropy classifier and Naive Bayes classifier were compared and Naive Bayes classifier was adopted. Besides that, tweets are categorized by topics using the CTM with the VEM algorithm. The result of this case study reached 86.4% accuracy in subjectivity classification and displayed specific topics describing the nature of the sentiment. In this research, the overall dataset they used contains only 1146 tweets, which includes only three airline companies. Besides, the author only used unigrams as sentiment classification features in Naive Bayes classifier, 13 which can cause problems because phrases and negation terms can change sentiment orientation of the unigrams in sentences. Besides that, their work did not present details about the classification approaches and comprehensive evaluations. However, my work not only contains the analysis of tweets with different sentiments but also includes the comparison of the performances of different approaches.

1.3 Timeline

Below is the tentative chronology of events as per my knowledge and experience. Hoping for everything going as per the timetable. Worst case estimate can be + 1 week.

Milestone	**Duration**
Data Collection	1 week
Data Pre-processing	2-3 weeks
Querying	2 weeks
EDA	2 weeks
Visualization	2 weeks
Final Project	1 week

DATA PREPARATION

There are an average of 6,000 tweets produced on Twitter per second. We think Twitter provides a great value to do sentiment analysis on text. Twitter posts are mostly public and can be used for such studies extensively. Also, frequent use of hashtags makes it more interesting to draw conclusions.

Sentiment analysis on airlines intrigues us since the industry is heavily price oriented. Often prices for tickets from different airlines are in similar range, putting emphasis on the quality of travel experience for the customer. But the customer understanding of the airline is commonly based on personal experience or general news. The customer is interested to know which airlines have a better reputation since the ticket prices are alike.

I did a sentiment analysis on tweets provided by the Twitter Search API in order to find the most preferable airline. I made a selection of airlines for the analysis. The selection was made at random. The other contributing factor was my initial analysis of the data collection process to verify that a sufficient amount of data is created for each airline. **The six airlines I ended up with were United, Virgin America, Southwest Air, Delta, JetBlue and American Air**.

My project aim was to focus on analysing the text of over 10,000 tweets about airlines using sentiment analysis and other methods to determine which airline receives the most positive or negative attention on Twitter, and what topics people are happy or sad about with regards to each airline.

2.1 Dataset Attributes Summary

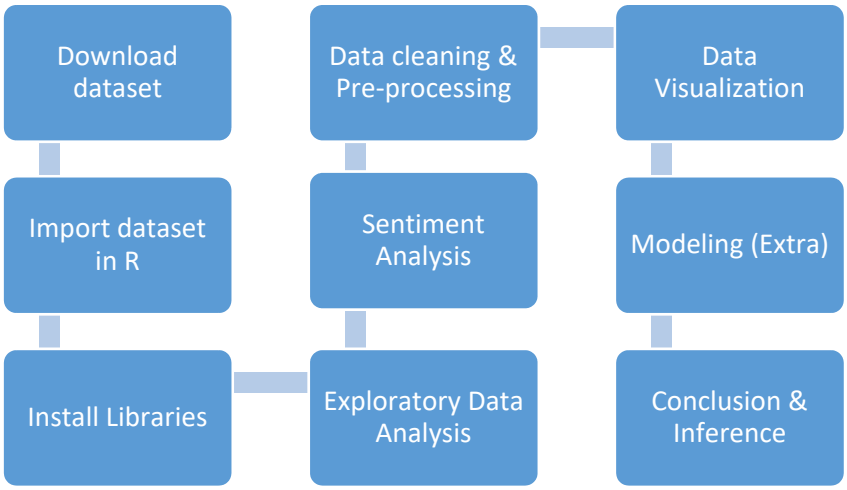
From this airline dataset I have primarily used the following attributes for the analysis:

Attributes	Data Type	Data Characteristics
TweetID	Int, Nominal	User's twitter ID
Airline_sentiment	String, Categorical	Can be either Neutral, Positive & Negative
airline_sentiment_confidence	Int, Nominal	Sentiment polarity associated with above column
negativereason	String, Categorical	Certain reason's which were extracted from tweets and categorized
negativereason_confidence	Int, Nominal	Polarity of the negative tweets
airline	String, Categorical	Labelling the tweets collected based on this column

userID	String, Nominal	Username's of the twitter users
retweet_count	Int, Nominal	Retweet count of the tweets
tweettext	String, Nominal	Tweet's of the user related to airlines
timestamp	Datetime, Nominal	Timestamp related to the tweets
user_timezone	String, Categorical	Time zone of the twitter users
numberofcharacters	Int, Nominal	Number of character's in the tweets
numberofwords	Int, Nominal	Number of character's in the tweets
tweet_location	String, Nominal	Location from where the user tweeted

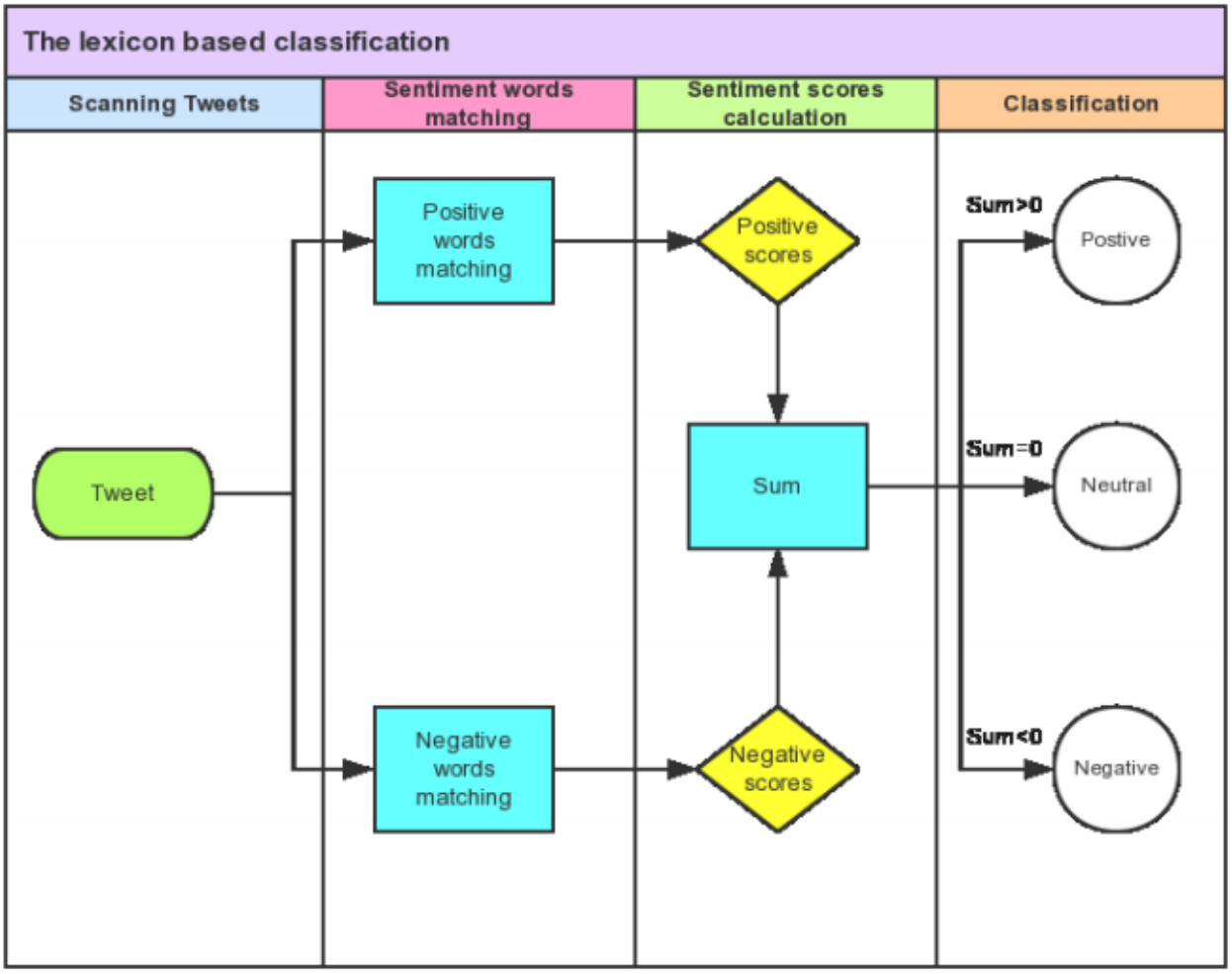
2.2 Project Workflow

The below is the presentation of the workflow in which I have approached this course project:



2.3 Sentiment Analysis

I used the Lexicon based classification for text pre-processing and cleansing. Using **tm** and **sentiment** library in R, I calculated values like airline_sentiment, airline_sentiment_confidence, negativereason and negativereason_confidence. Basically, this was how a lexicon based classification works:



2.3 Exploratory Data Analysis:

2.3.1 Columns containing NAs (no data)

```
##                       tweetID             airline_sentiment
##                             0                             0
## airline_sentiment_confidence                negativereason
##                             0                          5462
##    negativereason_confidence                       airline
##                          4118                             0
##       airline_sentiment_gold                        userID
##                         14600                             0
##          negativereason_gold                 retweet_count
##                         14608                             0
##                     tweettext                      latitude
##                             0                         13621
##                     longitude                     timestamp
##                         13621                             0
```

```
##                user_timezone           numberofcharacters
##                         4820                            0
##                numberofwords               tweet_location
##                            0                        13785
##                  tweet_coord
##                        13621
```

The apply command just checks which columns contain NA as well as returns the count of the same. **Airline_sentiment_gold** and **nevative_reason_gold** are mostly empty columns, i.e., they contain no information. So I can get rid of that.

2.3.2 ReTweet Analysis:

```
##
##     0     1     2     3     4     5     6     7     8     9
11    15
## 13873   640    66    22    17     5     3     3     1     1
1     1
##    18    22    28    31    32    44
##     1     2     1     1     1     1
```

I can see that most of the tweets are actually not retweeted. A very tiny fraction of them **(640/14640)** are tweeted only once. However, 4 tweets have been retweeted 44, 32, 31 and 28 times. Let's have a look and see why they say.

```
## [1] "@USAirways 5 hr flight delay and a delay when we land . Is
that even real life ? Get me off this plane , I wanna go home <f0>
<U+009F><U+0091><U+00A0><f0><U+009F><U+0091><U+00A0><f0><U+009F><U
+0091><U+00A0> (3 heel clicks)"
```

```
## [1] "@USAirways of course never again tho . Thanks for tweetin
ur concern but not Doin anythin to fix what happened. I'll choose
wiser next time"
```

```
## [1] "STOP. USING.THIS.WORD. IF. YOU'RE. A. COMPANY. RT @JetBlue
: Our fleet's on fleek. http://t.co/Fd2TNYcTrB"
```

```
## [1] "@USAirways with this livery back in the day. http://t.co/E
EqWVAMmiy"
```

The first 2 tweets show clear anger directed to US Airways. There was a substantial delay in the flight according to the first tweet, however the reason is not clear in the second tweet. The third tweet is directed towards Delta, although it is not clear what the message is. Being the curator of the dataset, I have identified this tweet as negative, because of the use of the word fleek which has negative polarity as per English dictionary. Finally, the fourth tweet is also targeted towards US Airways, the sentiment is neutral according to

me, because this is targeted towards the Airline Company or its officials (check livery definition).

2.3.3 Tweet location exploration

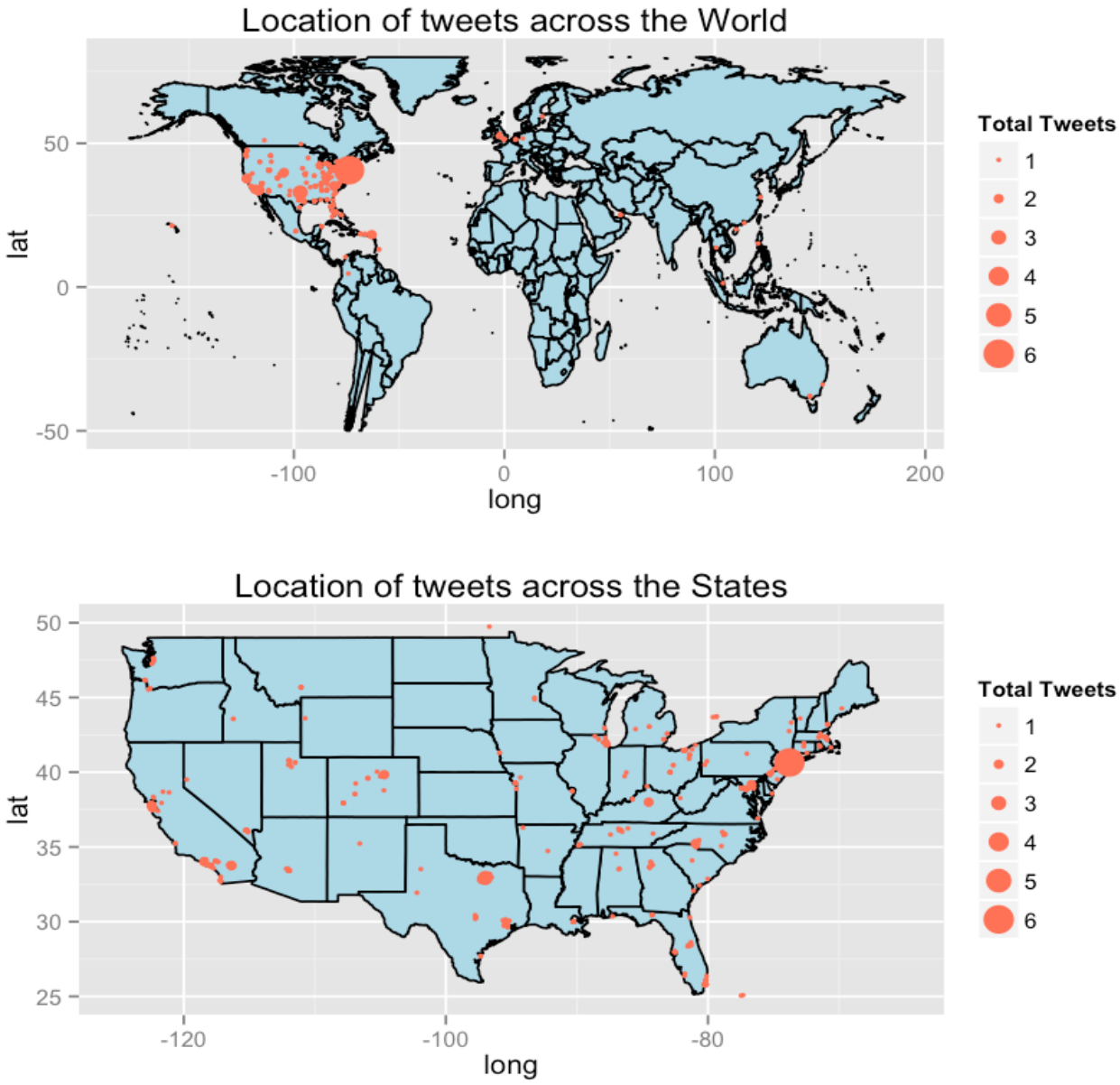
It would have been useful to know the location of the tweets to determine if certain areas are more prone to tweet, or to have one sentiment or the other. But as seen above, there are **13621** nulls in the latitudes and longitudes column. Hence we could only get the locations of **584** tweets and couldn't find any in particular favored location of the users. Only thing that could be inferred about this was most of them were tweeted within the airport premises. Maybe some insights can be shed upon time zone analysis.

2.3.4 Tweet Time-zone Analysis:

```
##                        timezone  Frequency
## 31  Eastern Time (US & Canada) 0.38126273
## 28  Central Time (US & Canada) 0.19663951
## 63  Pacific Time (US & Canada) 0.12301426
## 68                       Quito 0.07515275
## 13      Atlantic Time (Canada) 0.05061100
## 58 Mountain Time (US & Canada) 0.03757637
## 11                     Arizona 0.02331976
## 50                      London 0.01985743
## 3                       Alaska 0.01099796
## 77                      Sydney 0.01089613
```

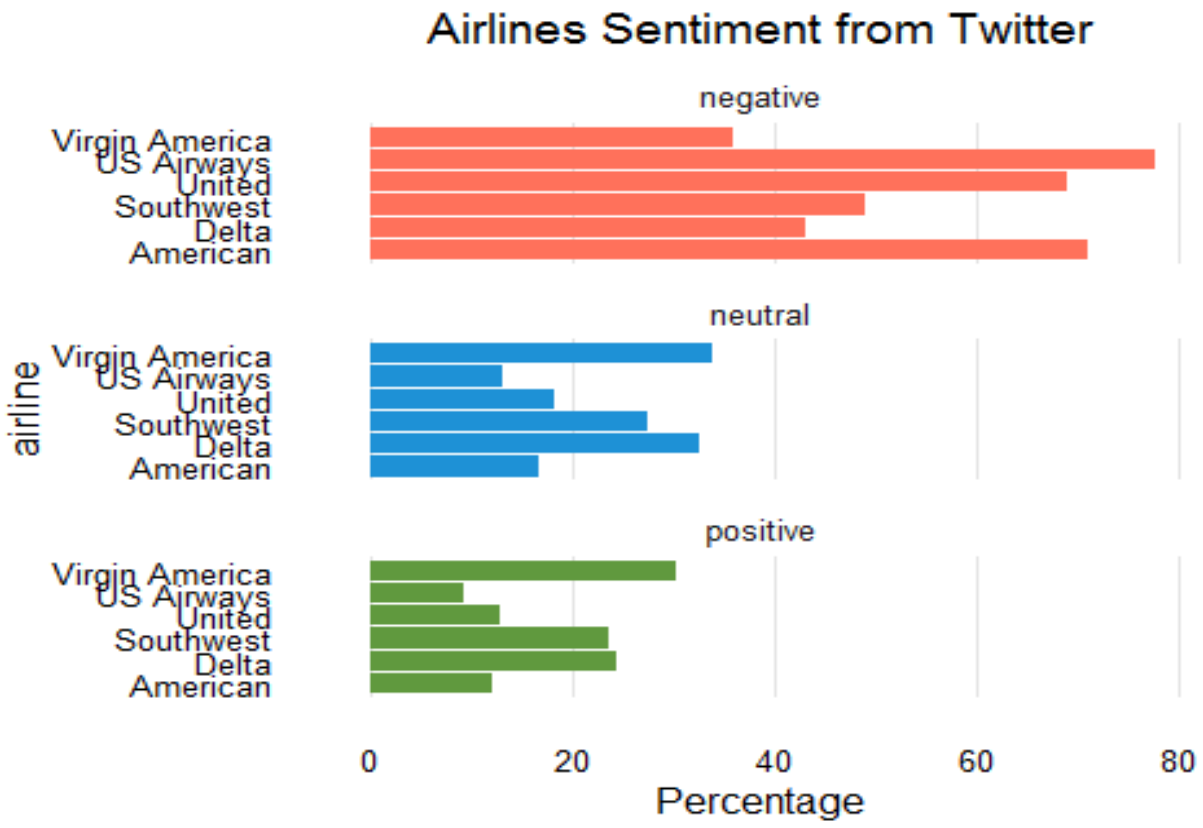
I found the great majority of tweets coming from Eastern Time zone, and almost all the tweets come from US & Canada time zone.

2.3.5 Visualization on maps



The above plots depicts the count of tweets over the US as well over the world. Basically the smaller the dot the less number of tweets and vice versa.

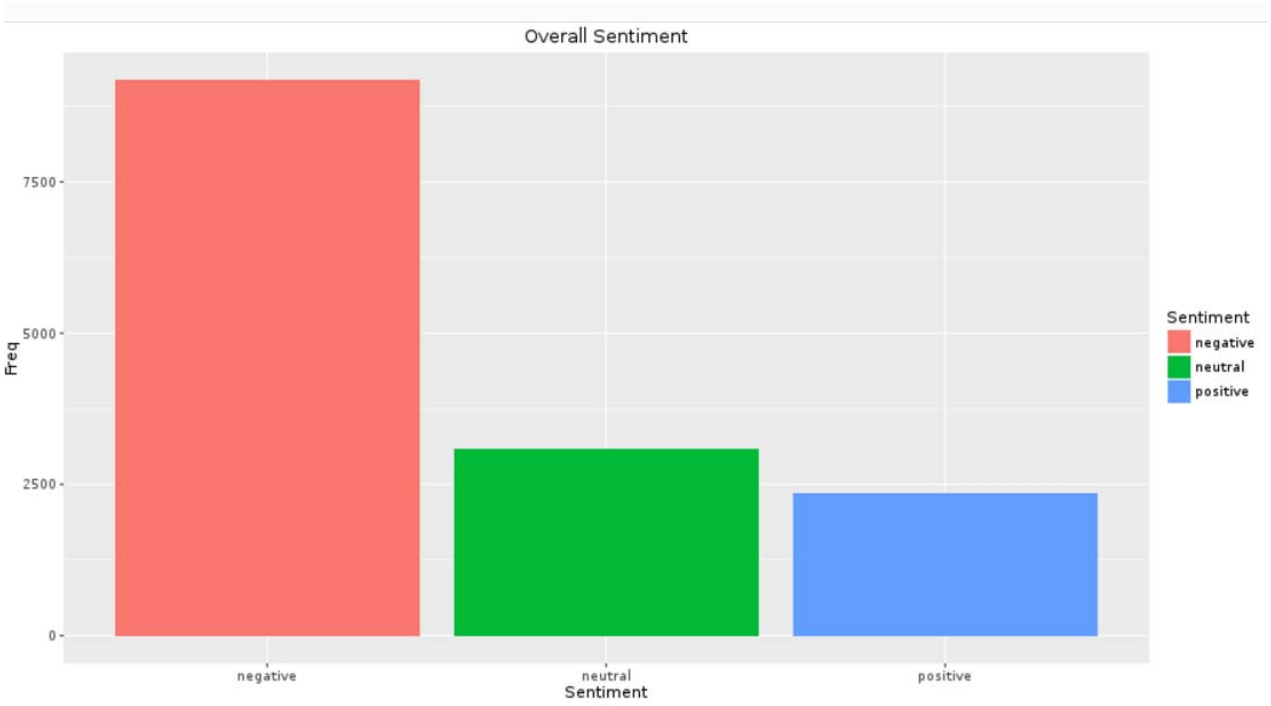
2.3.6 External Data Source Validation



The above graph shows distribution of tweets for a particular airline based on the tweets sentiment polarity with respect to the 6 different airlines. From this CNN link, I can see that Virgin America is number one and American is ranked lowest. As per my tweet dataset and sentiment analysis, I can see that Virgin America has most number of positive tweets and American has most negative number of tweets. US airways has the most negative tweets but, after 2015 as per this Wiki US Airways and American became one single carrier. Hence, the twitter data seems accurate.

RESULTS

3.1 Sentiment analysis of Airline Tweets based on Polarity

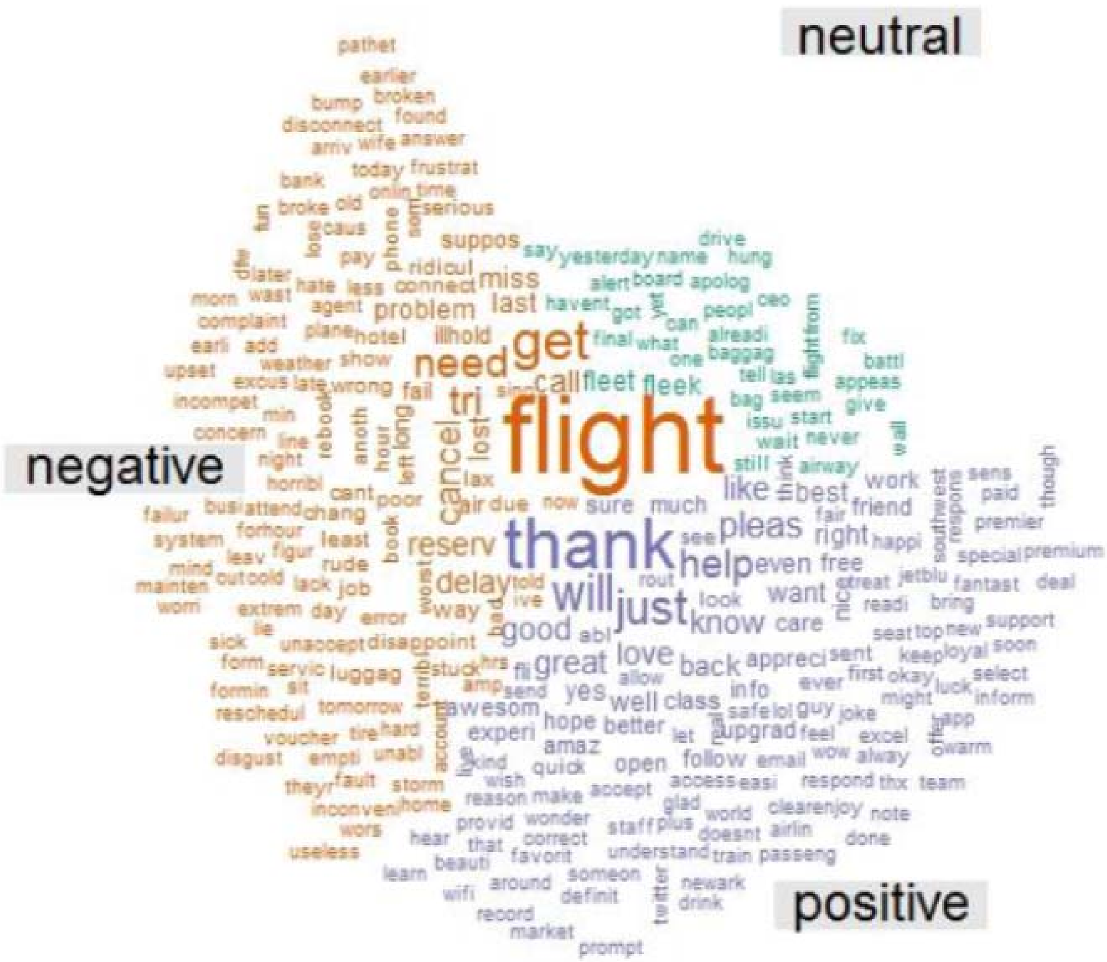


Sentiment Type	Tweets Count
negative	9178
neutral	3099
positive	2363

I have labelled the twitter texts against its sentiment classification – positive, negative and neutral. From the above illustration, it can be inferred that out of 14640 texts 9178 texts (63%) are related with traveller's negative emotional state and 2363 texts (16%) are related with positive emotional state. Remaining 21% texts are representing neutral state of emotion.

The finding confirms that there are issues related with travel experience which needs to be fixed. I will further analyse the negative tweets to identify the root cause of such negative experience of the travellers.

3.2 Comparison Cloud

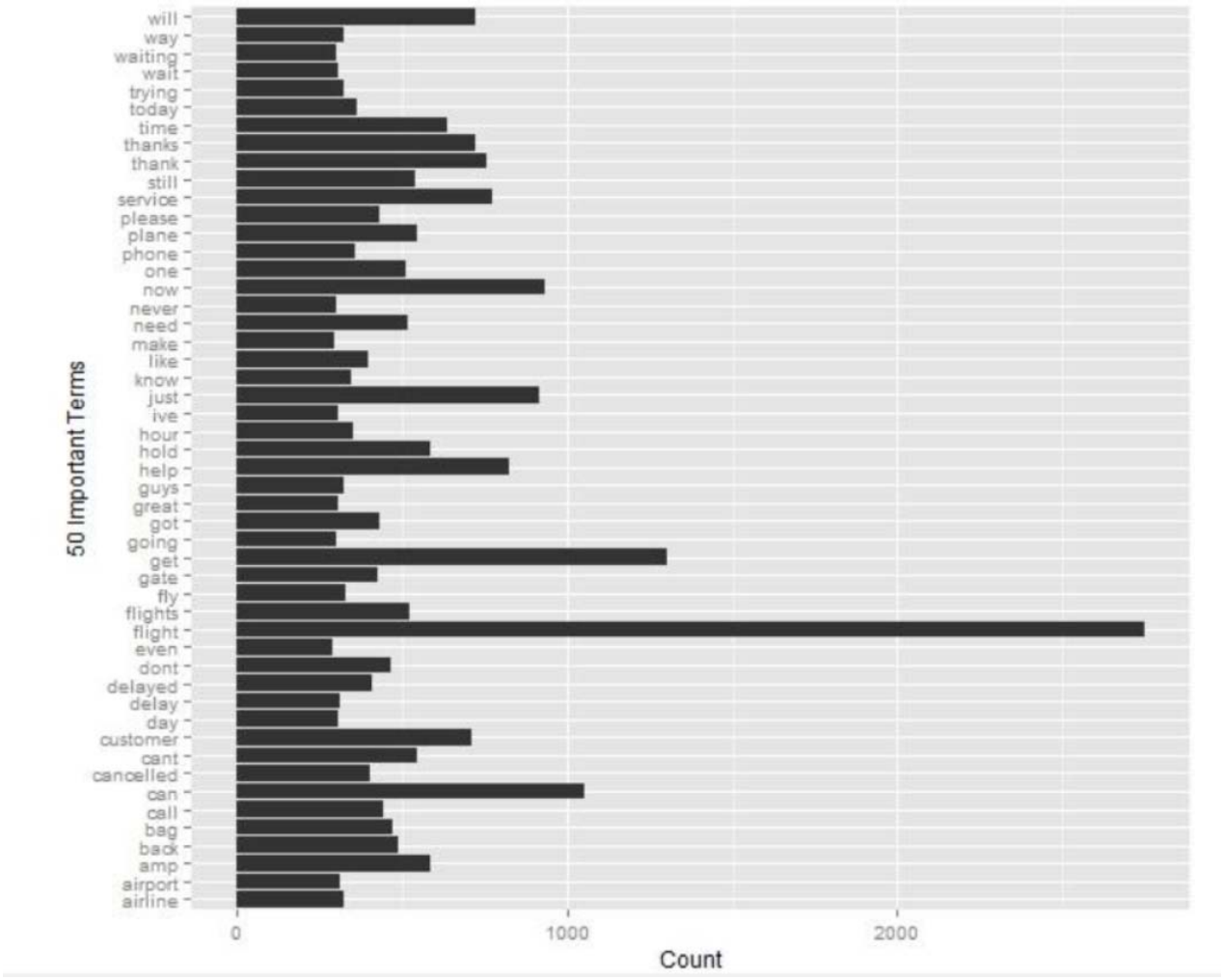


In this phase of my analysis I have performed text mining activities. To perform this task I have separated the texts by polarity, removed stop words, applied stemming and created a term document matrix with key words labelled by polarity and classification. The result is illustrated above through a comparison cloud.

From the above illustration, we can observe that a major area of the comparison cloud is highlighted as concern or focus area which includes all the terms associated with travellers negative emotional state. This result is driven by the experience that air travellers received during their flight journey.

If we further analyse this cloud, we can state that the most dominant term is **"Flight"** and it is highlighted in the negative polarity region. Other few important terms from negative polarity region are – "**Need", "Get", "Cancel", "Reserve", "Lost", "Delay", "Stuck"** etc. We can also observe terms from positive polarity region such as; "Thanks", "Help" etc. However, in this analysis we will primarily focus on terms from negative polarity region to justify our study purpose.

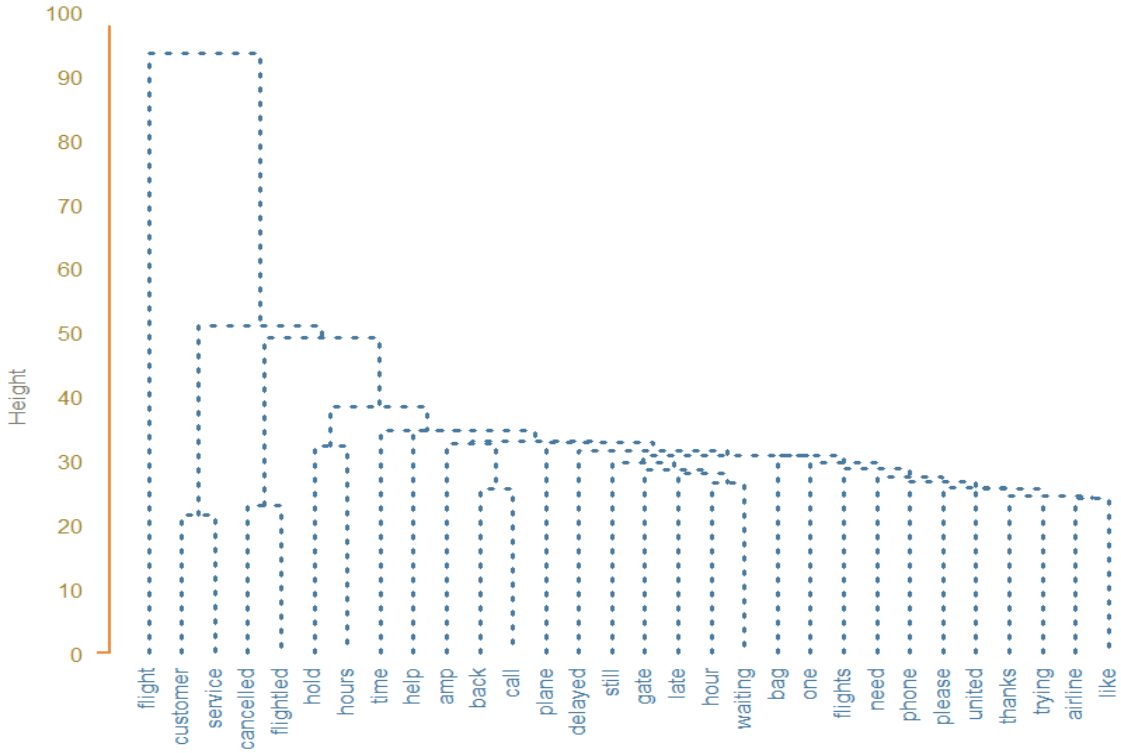
3.3 Top Words and Association Analysis:

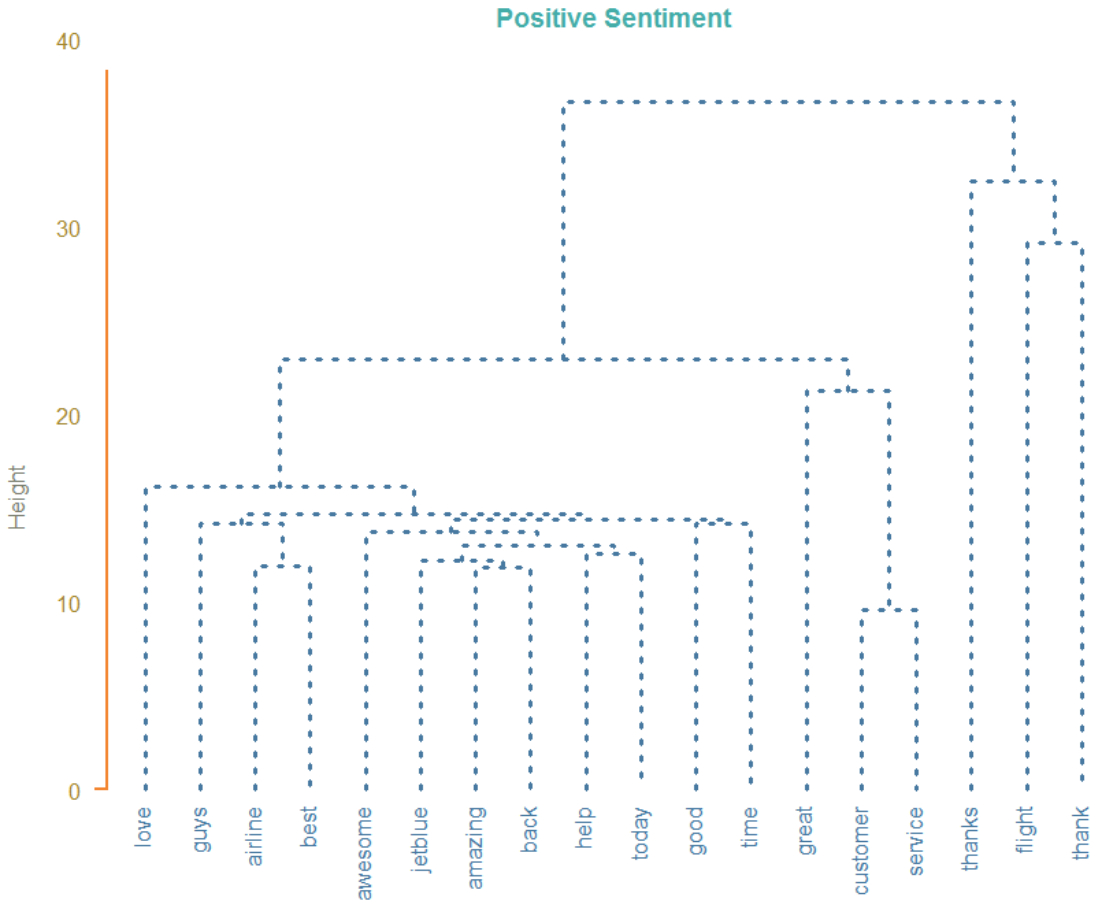


Sl.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Term	flight	get	can	now	just	help	service	thank	thanks	will	customer	time	hold	amp	cant	plane	still	flights	need	one
Frequency	2750	1303	1056	933	918	823	774	760	725	724	710	635	589	585	548	545	543	525	516	512
Sl.	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
Term	back	bag	dont	call	got	please	gate	delayed	cancelled	like	today	phone	hour	know	fly	airline	guys	way	trying	airport
Frequency	487	473	468	446	431	431	425	410	402	398	362	357	356	345	330	327	327	326	325	315
Sl.	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
Term	delay	great	day	ive	wait	going	waiting	never	make	even	flying	good	tomorrow	seat	change	last	want	new	check	weather
Frequency	314	310	309	309	308	303	303	300	299	289	286	274	271	270	269	263	263	261	260	257
Sl.	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
Term	really	told	work	first	take	another	travel	see	agent	email	getting	ticket	bags	due	worst	home	yes	love	much	lost
Frequency	253	253	245	244	242	236	235	233	232	232	232	232	231	229	227	224	223	218	218	217
Sl.	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
Term	people	someone	next	two	luggage	thats	crew	united	baggage	cancel	right	late	didnt	made	trip	ever	number	hours	let	canceled
Frequency	217	213	212	212	211	209	207	206	202	202	199	197	196	196	195	191	190	188	185	183

In this phase of my analysis I have performed text mining activities with top 100 selected terms. These are the terms which occurred most frequently in the analysis dataset. The second figure is about the top 100 words and their frequency in the dataset. The words marked in yellow are the most prominent ones occurring the dataset. This in turn helped in getting the **negative_reason** in my dataset using association analysis.

3.4 Clustering Analysis of Words:

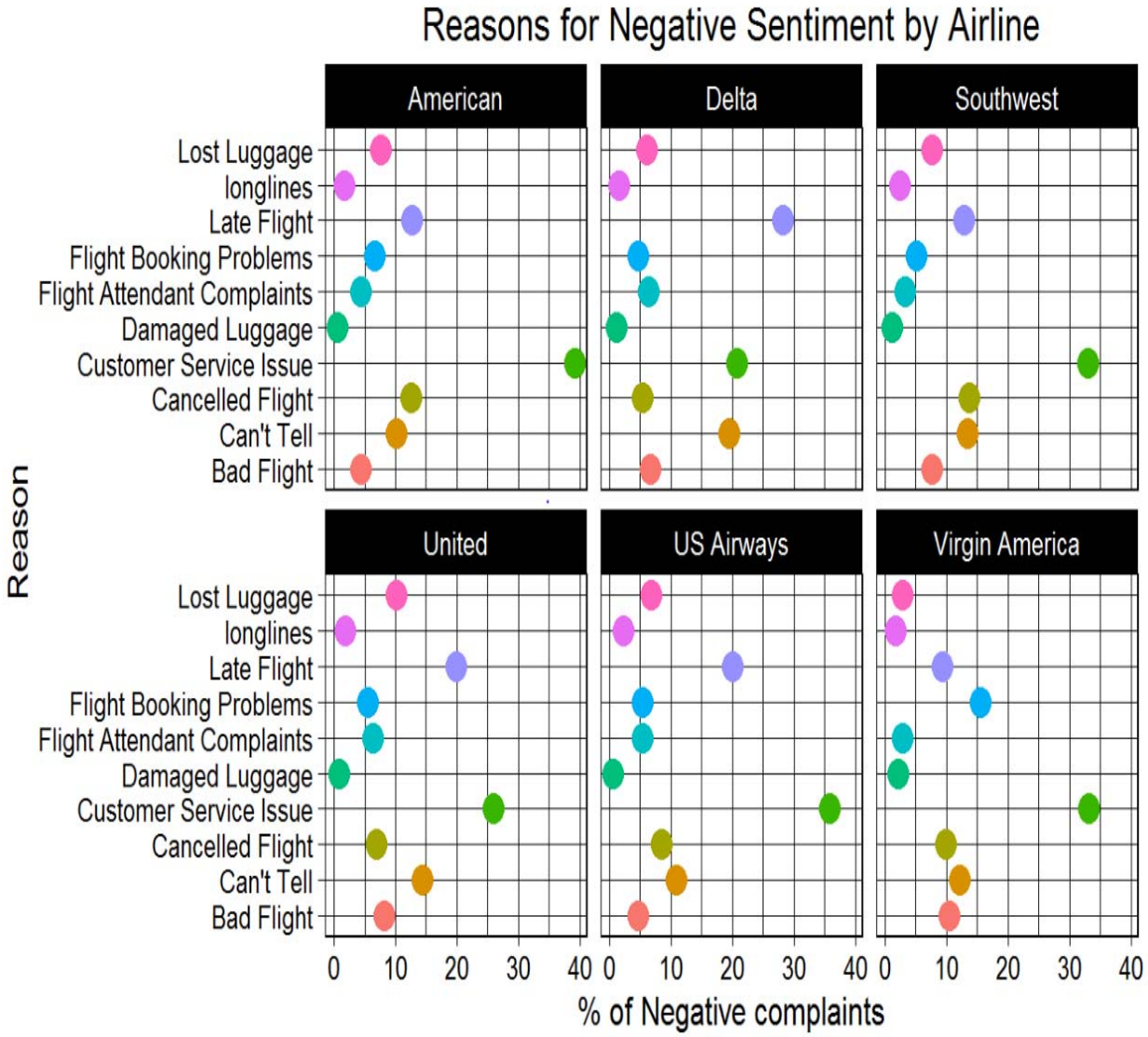




Although the negative sentiment dendrogram does not seem to be particularly informative, we observe again the association of words like customer and service, and cancelled flight. Words that reflect complains more generally, like waiting, bag (presumably lost), hours, time, hold, cluster altogether.

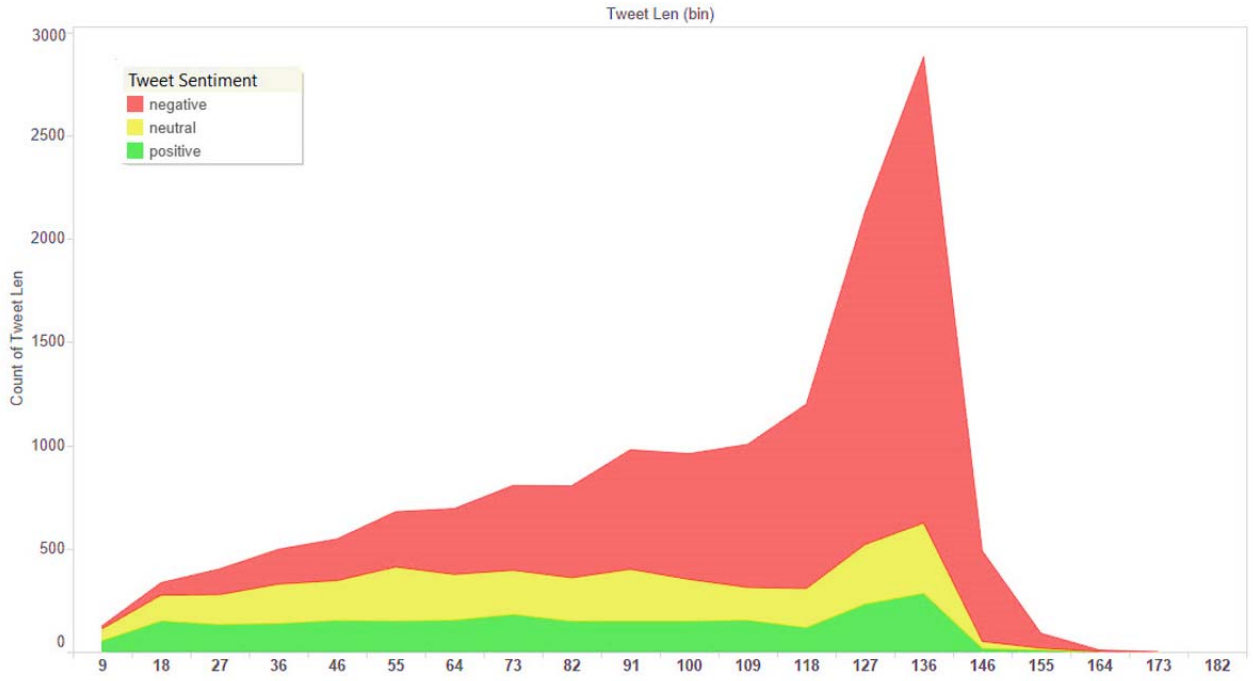
The positive tweet dendrogram is somewhat more informative. We can see the association of customer-service, and best-airline, or love-guys, good-time, which indicate more clearly, what the experience of the airline client was.

3.5 Negative sentiment Reason:



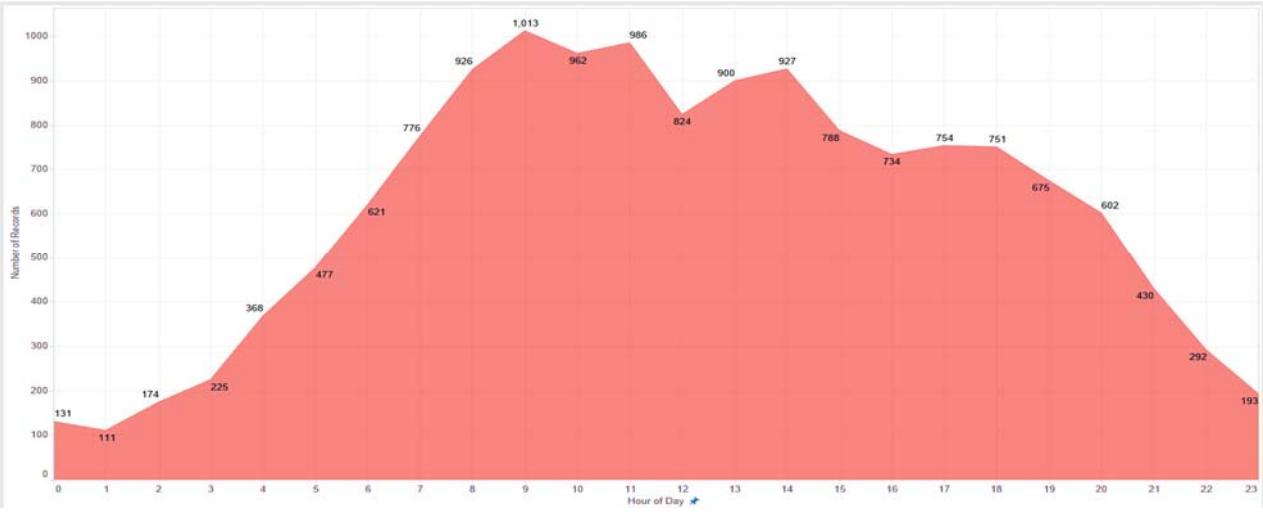
This part is just more elaborate part in analysing the negativity frequency in the negatively polarized tweets. It depicts frequency measures for reasons of the negative sentiment shared by the users per airline. The placement of the dot indicates the count of times the user has mentioned a particular phrase or set of words in his/her tweets related to dissatisfaction for that particular airline.

3.6 Length of Tweets



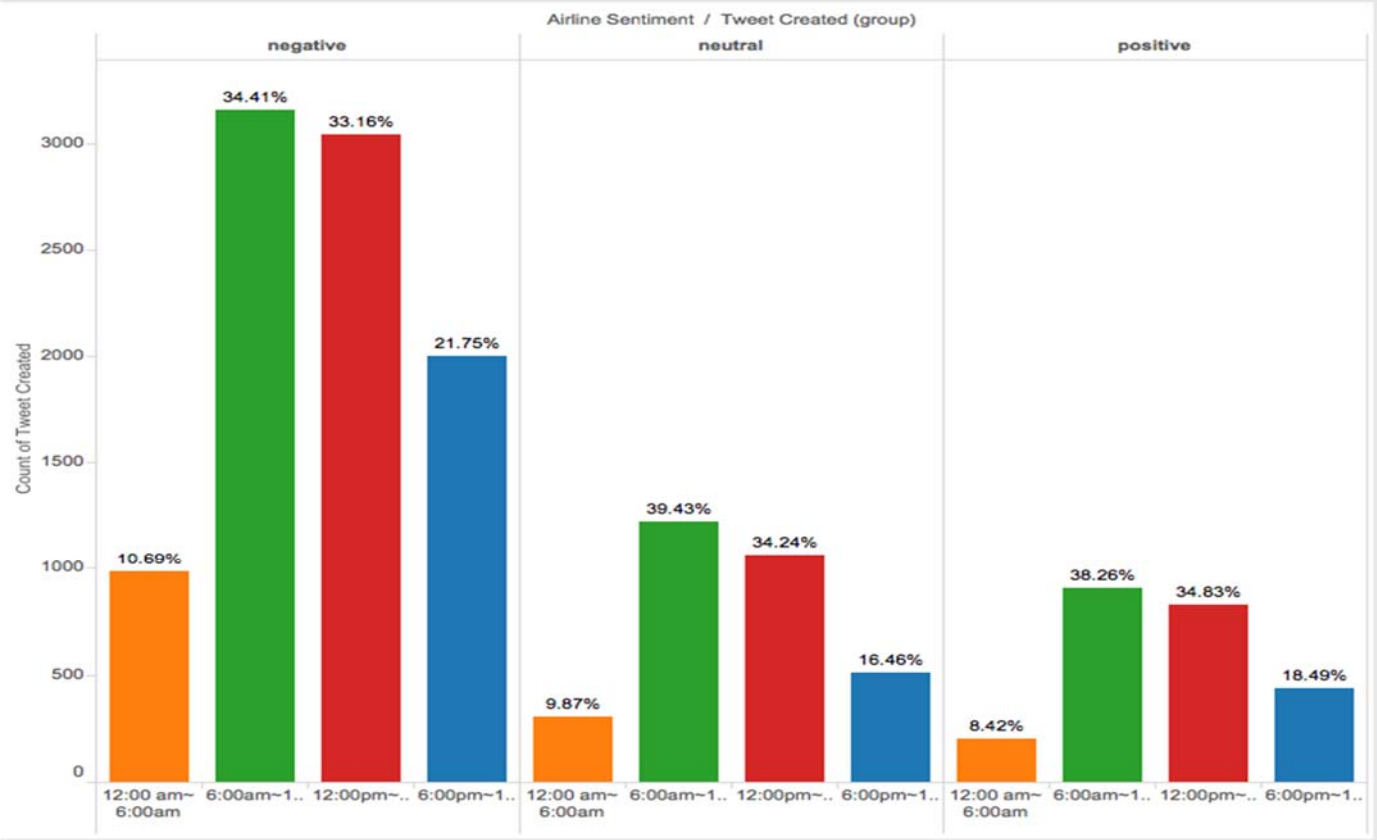
From the above bar plot, we can see that how tweet length varies for the tweets based on their sentiment classification. Negative sentiment tweets stands out the most here.

3.7 Tweet Timings:



The above is the tweet frequency distribution based on the timings of tweets. As we can see that most of the tweets are between **6:00 AM to 3:00 PM**. Airline companies should monitor the tweets during this time period in order to build its social image. More analysis into this based on sentiment and timings are done below.

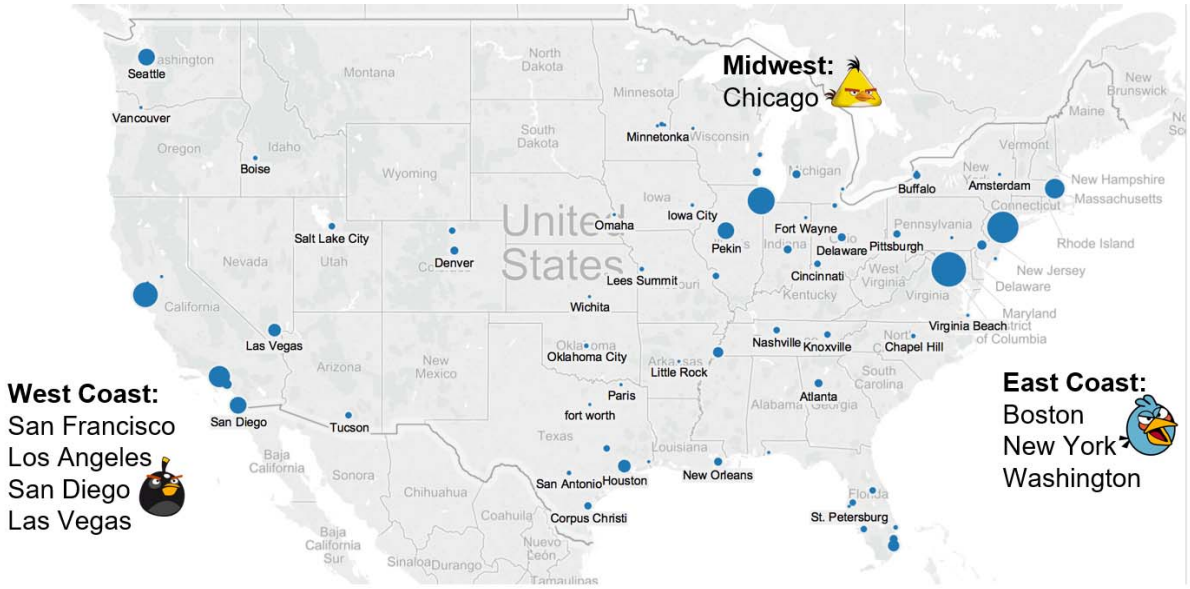
3.7.1 Tweet Timings II:



From this categorical clubbing of timings into 4 separate chunks and analyzing the frequency of the types of tweets based on their sentiment gives the above chart. According to the above chart, we can see that people tweet most in the morning between 6:00 am to 12:00 pm. Check this green section in the positive, negative and natural tweets.

About 69.84% of the total tweets is published from 6:00 am to 6:00 pm. Airline companies should monitor the tweets during this time period in order to build its social image

3.8 Negative Tweet Location:



From the map, we can see that most negative tweets come from big cities in mainly three parts of the US.

They come from east coast cities like Boston, New York and Washington and also from west coast cities such as San Francisco, Los Angeles, San Diego and Las Vegas. There are also some negative tweets come from mid-west like Chicago.

CONCLUSION & IMPLICATIONS

Sentiment classification has been intensively studied by researchers and professionals from different domains. Because of the wide applications in the business areas, many approaches have been developed for sentiment classifications. Every industry is getting into the big data era and applying data technologies to dig new opportunities to build better businesses. One of these technologies is the sentiment classification technology which can automatically classify the customer sentiments and provide comprehensive understanding of customer feedback from raw data on the Internet. In all of the social network platforms, Twitter has been one of the most popular sources for marketing information research and sentiment classification.

This course project makes empirical contributions to this research area by comparing the performance of different popular sentiment classification approaches and developing an ensemble approach, which further improves the sentiment classification performance. Besides that, the results of the experiments and the analysis on the tweets collected reveal much useful information for airline services improvements. Finally, the imbalanced accuracies of the classifiers in different sentiment class also reflects the customers' behaviours on Twitter.

All the interpretations of the plots have been mentioned alongside their plots. I'll again reiterate some of my analysis results, which didn't had any plot and simple R input. I conducted exploratory data analysis to understand and get familiar with the data at hand. I found that:

- Most tweets have negative sentiment (> 60%).
- Most tweets are targeted towards United, followed by American and US Airways.
- Virgin American receives very few tweets.
- Most of the tweets targeted towards American, United and US Airways contain negative sentiment.
- Tweets targeted to Delta, Virgin and Southwest contain roughly similar proportion of negative, neutral and positive sentiment.
- 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 re-tweeted.
- Most tweets come from US & Canada time zone
- Most tweets come from the States.

I also tried to test a couple of hypotheses. The first is that the more addresses in a tweet, the harsher (negatively polarized) its words. I did this by first counting the number of **@** symbols in the text of each tweet. I then used a few visualizations to investigate my theory further. Maybe it depends on the airline? My second hypothesis is that longer tweets are

also less likely to contain favourable language. **(Assignment - 4)** We see that negative tweets tend to be considerably longer than positive or neutral ones. In fact, it's interesting to see that ceiling effect of the 170 character limit among tweets directed at Virgin America. Something's seriously wrong with their services. I then conducted basic text analytics for the tweets. I displayed in word clouds with the frequency of words, the main topics of conversation in tweets with negative and positive sentiment. Then, we found associations between words that allowed us to better understand what the customers were complaining about, or why they enjoyed their flying experience.

Practical Implications

The objectives of the course project have been achieved by having:

- Did the text analysis of the tweets and found out various inferences, reasons for negative tweets, tweet classification based on their sentiment, polarity of each tweets, negative words associations in the tweets, sentiment based word cloud, top 100 word occurrences in the tweets and their associations.
- All the above text analysis and other correlations and plots helped in determining the best and worst airplane in customer service.
- Got the most negative discussed topics for different airlines as well the classification of these based on time-zones. One more thing worth analysing in this section was to determine the tweets frequency amongst the users based on timings and classify the same based on the tweet sentiment.
- Used text mining and **sentiment** library in R helped me in doing the positive and negative words analysis as well as the tweet associations with those words.
- With sentiment analysis, it helped me in discovereing that Twitter users like to express their complaints toward airline services in a sarcastic way. This reveals the linguistic customs on the Internet.
- This approach is applicable for the airline companies to analyse the twitter data about their services.

LIMITATIONS

There are might be certain limitations to this project as per my knowledge. First of all, the tweets collected from the Twitter API are not as pure as required for the sentiment classification. By searching tweets with keywords “flight” and the airline brand, there is a probability that 30% to 40% of dataset might be irrelevant tweets. Airline companies still need to further improve the accuracy of tweet data retrieval. Secondly, compared to the tweet data existing in Twitter, the dataset collected and used in this project is a very tiny part. This is a problem to solve for doing scale sentiment classification by applying big data techniques. However, it requires expensive infrastructure investment to do this kind research and application. Besides that, my tweet data are very messy and they contain a lot of typos and abbreviations. Even though adopting feature selection techniques to reduce the dimensionality, it still cannot group all words with the same roots into one stem. It is desirable to auto correct all the typos and to extend the abbreviations to regular words, which requires high level Natural Language Processing techniques. More than that, the balanced class distribution is not a real world case and it might cause over-fitting problem in positive class. In the future, more complicated models are expected to be built to solve this problem.

Moreover, for different airline brands, the features for sentiment classification might be different from each other and it is valuable to train sentiment classification models for different airline brands. To give more specific and valuable information for the decision makers of the airline companies, sentiment classification can be applied to the tweets about their airline services and produce detailed reports. Last but not least, there are also many further research directions, which can be worked on. Other information like the users who tweet them, the times of the retweets and other factors are also potentially useful. The time series analysis of the twitter sentiment about airline services is also an interesting topic and the time data is available in the tweets retrieved from the Twitter API.

ETHICS OF SOCIAL MEDIA RESEARCH

There are a number of ethical considerations to keep in mind when using social media data for this research. One of the key concerns hinges on the extent to which social media data should be treated as public versus private data. Even though social media data are publicly available, social media users may not intend or wish for their data to be used for research. Users may not be aware that their social media data is publicly available and may have expectations of privacy even in public settings. The distinction between public and private data becomes additionally complicated by the fact that machine learning algorithms can make inferences about private attributes, even if not explicitly stated in public data. These issues have been properly addressed in this research and no user information has been neither been collected nor has been used for analysis and inference.

LEARNING OUTCOMES

In addition to the analysis provided above, there's another big piece, which is the fact that Twitter sentiment analysis is a very novel problem from an academic standpoint. For a long time sentiment analysis was largely based on text samples that were a paragraph or larger. With the large amount of information in a paragraph of text it's much easier to achieve good performance.

When twitter first came out, it utterly broke most existing sentiment analysis approaches, mostly because it was based around very short, informal text. It became a very interested problem, not just from a business perspective, but also from an academic perspective. The other nice piece was that the Twitter dataset is MASSIVE. The huge amount of data that twitter makes available made it possible to explore a number of machine learning approaches, specifically in the realm of neural networks that were largely intractable on smaller datasets.

Ultimately, the synergy between academic curiosity and business interest led to the large amount of interest in twitter sentiment analysis seen publicly. Sentiment analysis was chosen largely because it is a simple binary classification problem that does away with many confounding factors like class imbalance that would impede more complex problems.

This project on an overall taught me to mine the sentiment of tweets just as I did for this project, and gave me an idea of how it can be done for other social media sites. Also, it helped me in understanding various aspects of R and Python programmatically as well as practically. The one thing that I could do differently is to use more robust method for tweets classification rather than using a library in R and python. A supervised model that has been built using all the dictionary words which helps in determining the polarity and sentiment classification for classifying the tweets is a preferred method.