CAPSTONE PROJECT REPORT
"Airline Twitter Sentiment Analysis"

Prepared for:
Syed Shariyar Murtaza
CKME 136 - Data Analytics: Capstone Course

Prepared by:
A. B. M. Naimul Huq
Ryerson ID: 500672338

June 14, 2015

Airline Twitter Sentiment Analysis

1. Introduction

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 (http://theacsi.org/) quantifies this impression. As a group, airlines fall close to the bottom of their industry rankings, just above the Postal Service, Health Insurance, Television and Internet Service companies.

Appendix: ACSI Industry Scores

1.1 Brief context about the problem

A number of airlines from different operators are operating in different routes and assisting people to travel their preferred destination. Due to a great demand of air travel and strong competition between airlines, travellers are getting an increased choice of airline, airport, price and service. However, the standard of services varies. On the contrary, it is important for airline operators to remain competitive in terms of price and being preferred by travellers based on the service quality. Therefore, it is important for airline operators to understand traveller's emotions and identify factors related to their services which might affect their brand preference.

1.2 Statement of the Problem

US airlines struggle to deliver a consistent, positive customer experience. Air travellers have stated issues related to their travel experiences in social media which needs to be analyzed so that air line operators can identify, understand and fix issues to improve the overall travel experience.

1.3 Proposal

The purpose of this report is to perform a **"Sentiment Analysis"** job about problems of each major U.S. airlines regarding their services. I will classify tweets based on polarity. The classification of polarity function will allow us to classify some text as positive or negative. I will tokenize the texts and separate key words based on polarity. I will prepare a model with tokenized terms which will assist airline operators to understand traveller's sentiment regarding their services. Finally, I will run Naïve Bayes to check the accuracy of the model.

- Perform sentiment analysis
- Perform exploratory data analysis
- Tokenize texts and build a model with unigrams
- Check the prediction of the model with reality through Naïve Bayes

2. Dataset

I have used **"Airline Twitter Sentiment"** dataset from **"CrowdFlower"** website.

A sentiment analysis job about the problems of each major U.S. airlines was performed based on this dataset. Twitter data was captured from Feb., 2015 and contributors were asked to first classify positive, negative, & neutral tweets, followed by categorizing negative reasons (such as "late flight").

From this airline dataset I have primarily used the following 6 attributes for the analysis.

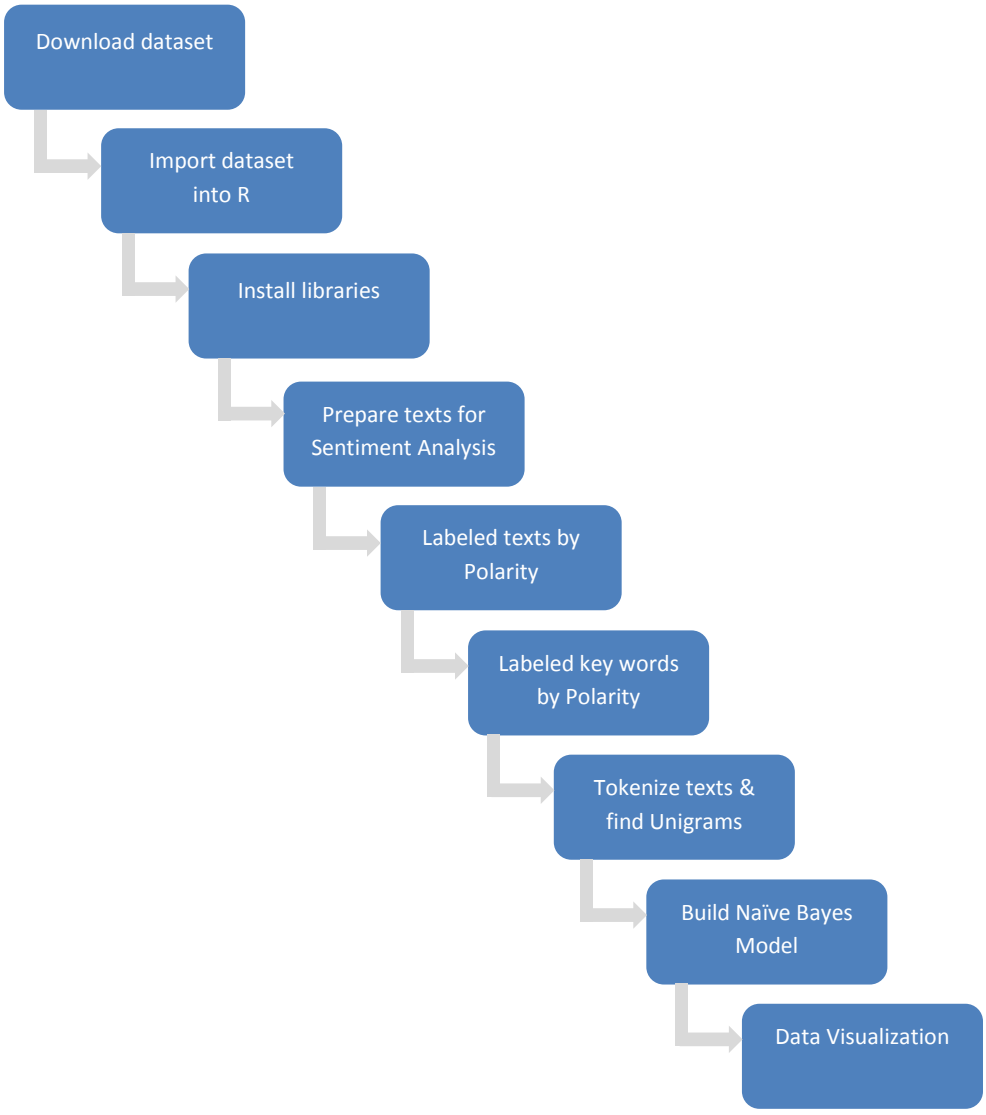
Attributes	Data Characteristics
_unit_id	Int
_id	Int
_country	Factor
_city	Factor
Airline	Factor
Text	Factor

However, the dataset has other 16 attributes which I have not considered for this analysis. These attributes are: created_at, golden, missed, started_at, tained, channel, trust, worker_id, region, ip, negativereason, airline_sentiment_gold, name, negativereason_gold, retweet_count, tweet_created, user_timezone, tweet_coord, airline_sentiment, tweet_id, tweet_location.

Source: http://www.crowdflower.com/data-for-everyone (added: February 12, 2015 by CrowdFlower)

3. Approach

An overview of the steps that I have considered for this sentiment analysis is stated below.



Step-1: Download dataset from CrowdFlower

I have downloaded the raw dataset of **"Airline Twitter Sentiment"** from **"CrowdFlower"** website. This dataset contains 27 attributes. However, I have selected 6 attributes with 55783 observations for the analysis and saved the new dataset as **"Airline"** in csv file format.

Step-2: Import dataset into R

After import the dataset into R I have first checked the data characteristics to understand the dataset and afterwards removed the duplicate data to avoid any data duplication.

Step-3: Install libraries

I have installed below libraries to perform the analysis.

3.1 Library (sentiment)
Sentiment is a R package with tools for sentiment analysis including Bayesian classifiers for positivity/negativity and emotion classification.

3.2 Library (tm)
This is a text mining package. This package will be used to perform below tasks.

#Clean unstructured twitter texts to perform sentiment analysis (e.g. remove numbers, remove punctuation, strip whitespace, remove words, stop words, stem document)

#Constructs or coerces to a term-document matrix or a document-term matrix after separating words from twitter texts (e.g. Term Document Matrix, Document Term Matrix)

#Remove Sparse Terms from a Term-Document Matrix (e.g. Remove Sparse Terms)

#Find associations in a document term or term document matrix (e.g. find assocs, find frequent terms)

3.3 Library (plyr)
Tools for splitting, applying and combining data (e.g. laply)

3.4 Library (Rstem)
Rstem package provides an interface to C code that performs stemming on words.

3.5 Library (SnowballC)
An R interface to the C libstemmer library that implements Porter's word stemming algorithm for collapsing words to a common root to aid comparison of vocabulary.

3.6 Library (RWeka)
Weka is a collection of machine learning algorithms for data mining tasks written in Java, containing tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

3.7 Library (ggplot2)
This is a data visualization package in R.

3.8 Library (wordcloud)
Plot a cloud comparing the frequencies of words across documents. (e.g. comparison cloud)

3.9 Library (RColorBrewer)
Provides color schemes for maps and other graphics.

3.10 Library (NLP)
Basic classes and methods for Natural Language Processing (e.g. tokenize texts).

3.11 Library (slam)
Sparse lightweight arrays and matrices (e.g. find frequency from unigrams).

3.12 Library (klaR)
Implementation of naive bayes in R.

3.13 Library (caret)
Experimental design.

3.14 Library (e1071)
Confusion matrix.

3.15 Library (MASS)
Support functions and datasets for Venables and Ripley's MASS.

3.16 Library (lattice)
Support Lattice Graphics

Step-4: Prepare text for sentiment analysis

In this step, I have cleaned the data (twitter texts) and prepared for sentiment analysis. This task required to perform a series of tasks which are stated below.

- removed retweet entities
- removed at people
- removed punctuation
- removed numbers
- removed html links
- removed unnecessary spaces
- defined "to lower error handling" function
- applied lower case
- removed NAs in air_txt
- erased characters that are not alphabetic, spaces or apostrophes

Step-5: Labeled tweets by polarity

Tweets can be labeled either by emotion or polarity. In this analysis I have considered polarity option. In this phase, I have performed a series of tasks which are stated below.

5.1 Classify polarity

I have used **"Voter"** algorithm to classify polarity. We can perform the same task through **"Bayes"** algorithm. I have labeled the twitter texts against – positive, negative and neutral.

5.2 Create data frame with the results and obtain some general statistics

After classification of polarity I have created a new data frame with 2 observations – text and polarity. So, this new data frame contains twitter texts which are labeled by polarity.

5.3 Plot distribution of polarity

I have used ggplot library to illustrate the distribution of polarity. The graphical representation will show us a good indication about the actual emotion of airline travellers regarding their travel experiences.

Step-6: Labeled key words by polarity

In this step, I have separated the texts by polarity, removed stop words, applied stemming and created a term document matrix with key words labeled by polarity.

From this matrix, I can now identify frequent words which are related to airline traveller's emotions.

I have used comparison cloud and word cloud to visualize the findings. With comparison cloud, I have illustrated key words against polarity. With word cloud, I have illustrated key 100 words that we have extracted from the texts.

Step-7: Tokenize texts

In this step, I have tokenized the texts from "Airline" dataset so that I can prepare a model with tokenized terms.

After tokenization, I have selected 100 most important terms based on frequency and created a data frame with unigrams labeled with polarity. This model will be used later as a sample of input to check the accuracy through Naïve Bayes algorithm.

Step-8: Build Model

In this final step, I have used Naïve Bayes algorithm to check the accuracy of my model based on unigrams. To check the model I have selected 100 most important terms / unigrams as sample.

8.1 Naive Bayes Model

The Naive Bayes makes the simplifying assumption that all the features are independent. In other words and in the context of Sentiment Analysis, each token (word or group of words) contributes independently to the sentiment of the whole sentence. Even if this assumption may seem too restrictive, Naive Bayes gives good results as it does not over fit.

8.2 Evaluation phase

- I have predicted the classes of the test instances. I have checked prediction against reality.

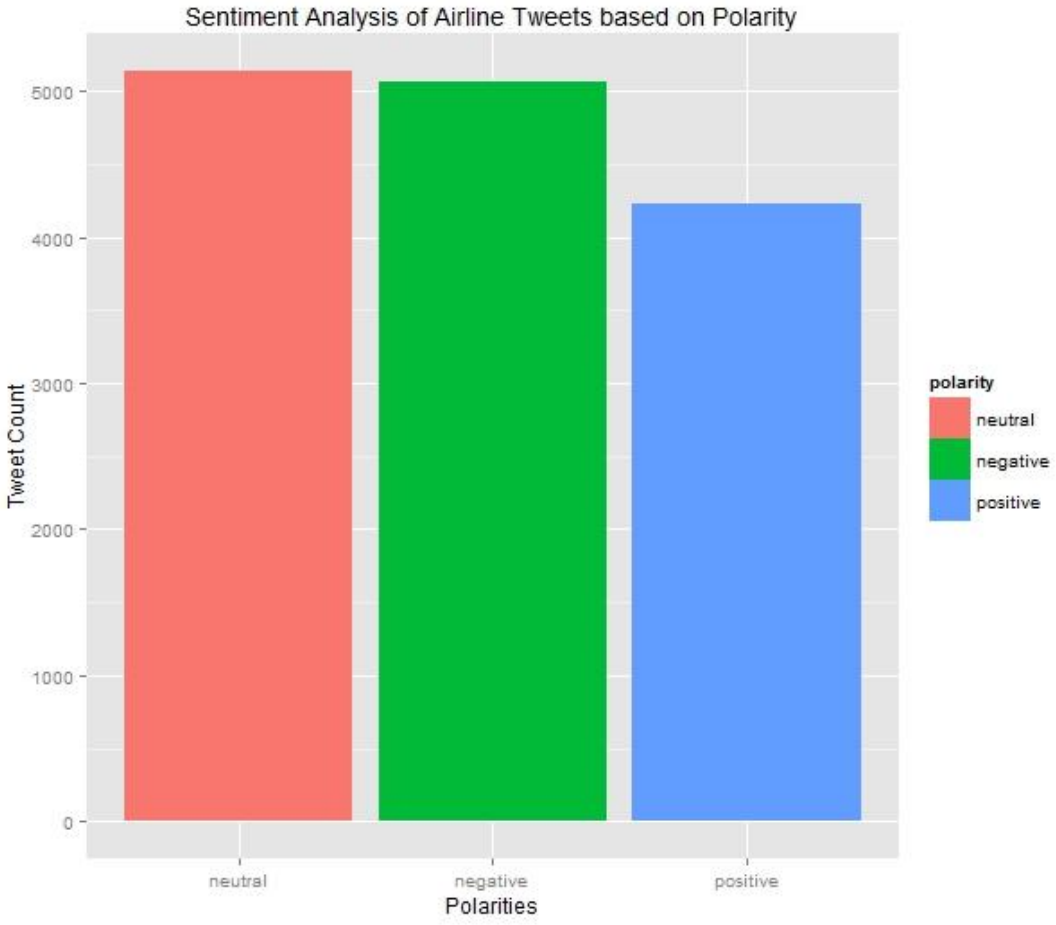
8.3 Performance criteria

- I have evaluated the performance through Confusion matrix

***Please note that I have performed my analysis in R and entire code can be viewed from the given link of my GitHub account (**https://github.com/naimulhuq/Capstone)

4. Result

4.1 Sentiment analysis of Airline Tweets based on Polarity



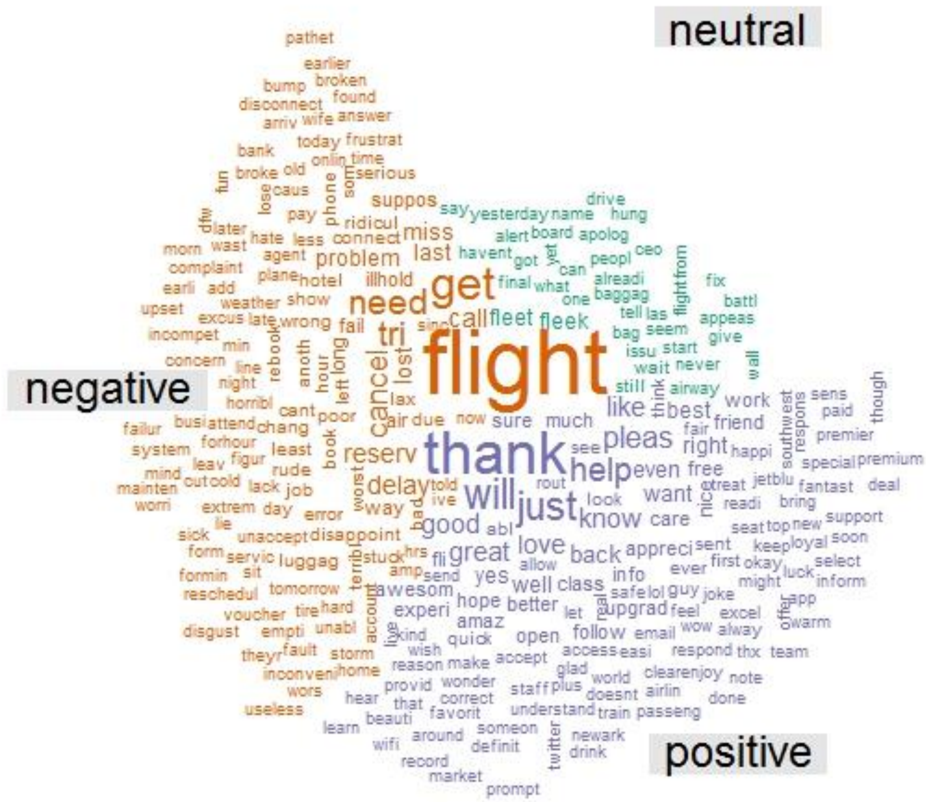

```
> table(air_sentiment$polarity)

 neutral negative positive
    5135     5064     4228
```

Interpretation

We have labeled the twitter texts against polarity – positive, negative and neutral. From the above illustration, we can inferred that out of 14427 texts 5064 texts (35%) are related with travellers negative emotional state and 4228 texts (29%) are related with positive emotional state. Remaining 36% texts are representing neutral state of emotion.

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

4.2 Comparison Cloud



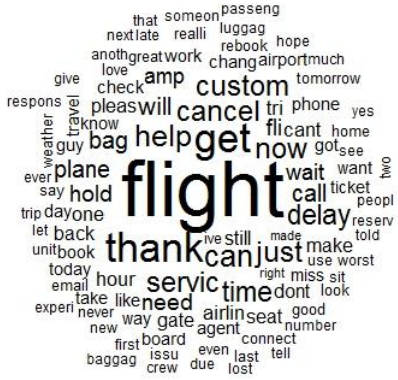
Interpretation

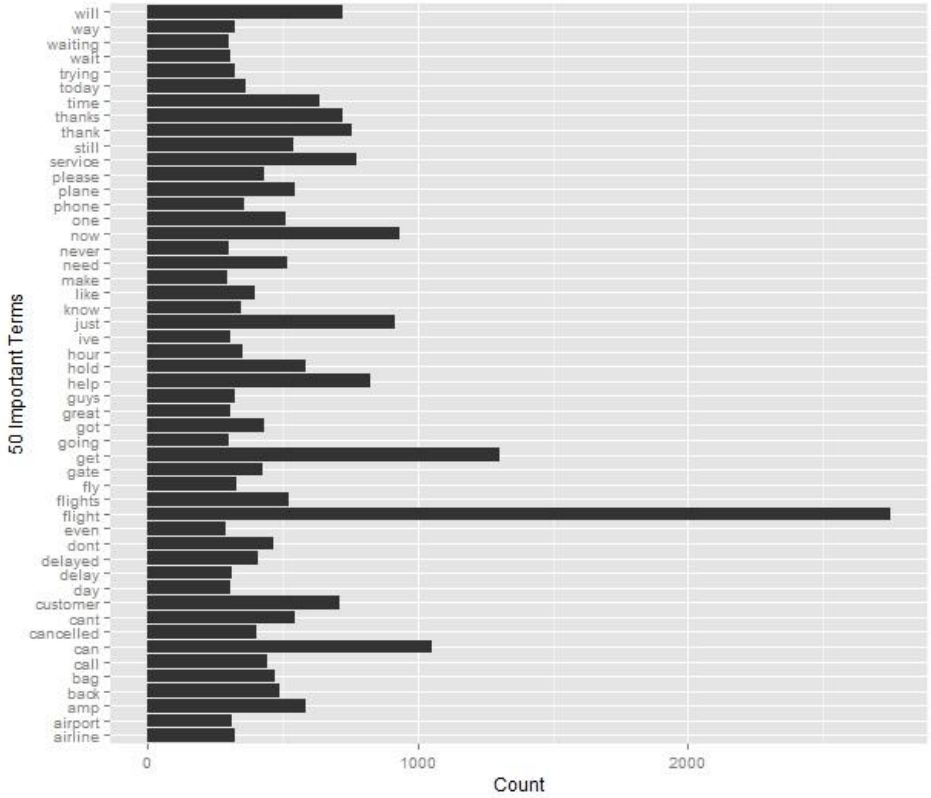
In this phase of my analysis I have performed text mining activities. To perform this task I have separated the texts by polarity, removed stop words, applied stemming and created a term document matrix with key words labeled by polarity. The result is illustrated above through a comparison cloud.

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 analyze 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.

4.3 Word Cloud





Interpretation

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.

4.4 Text Mining

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

Table-1: 100 Important Terms

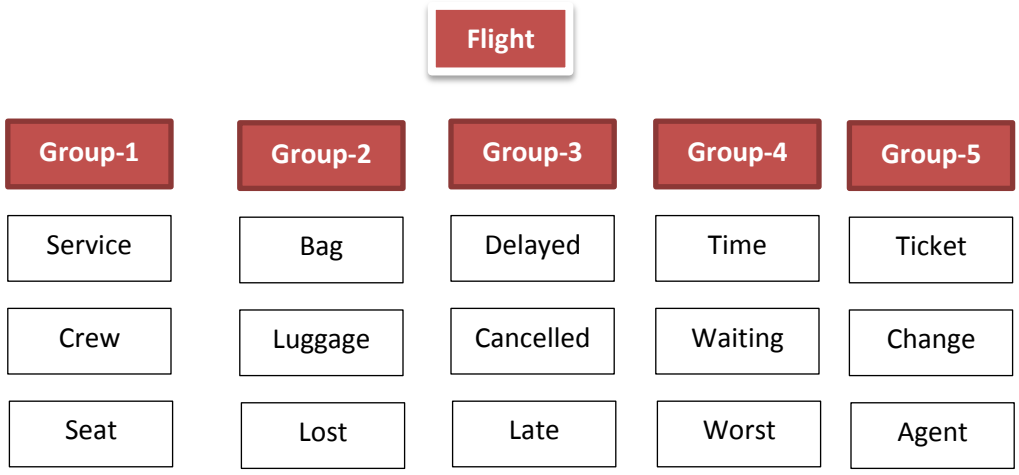


Table-2: Top 15 Terms

Interpretation

In this phase of my analysis, I have randomly selected 15 terms from the above table composed with 100 most important terms to determine association with airline's service quality. I will randomly pick tweets or comments from the dataset based on these 5 groups of unigrams. It will allow us to identify the root causes which are affecting the service quality of airlines.

4.5 Text Mining (Association of Key Terms with Tweets)

Keyword: Seat
"booked flight mnth ago seat confirmed small plane arrived and was not allowed to board bcs no seat horrible service**"** - US Airways

Keyword: Service / Crew
"serious display of poor customer service exhibited by flight crew today new American airline cheap slogan not motto" - US Airways

Keyword: Bag
"forces us to check our baby bag on overbooked flight complains to wife that we need to much for our baby united has no baby meals" - United

Keyword: Lost
"first you lost all my bags now you cancel my flight home min wait to talk to somebody poor service not good enough" – United

Keyword: Cancelled / Luggage
"after a canceled flight and delays you lost my luggage again you're the worst disgraceful awful company horrible service" - United

Keyword: Delayed
"the hotel you sent us to wouldn't take the voucher our flight was delayed then cancelled then delayed again hours and counting" - United

Keyword: Time
"possibly the worst airline gave them three chances second time in weeks a flight has been delayed and or cancelled due to mechanics" – United

Keyword: Worst
"you are the worst airline in the world from your crap website to your worthless app to your late flight you suck just shut down" - United

Keyword: Agent
"delayed again by united gate agent was borderline rude when i asked her a flight status question what is happening to united" - United

Keyword: Ticket
"i had to purchase a ticket that i never would have had to buy had the flight not been cancelled terrible customer service" - US Airways

From the above text mining analysis we have identified 3 broad areas where airline operators have scope for improvement. These areas are – Flight schedule, Customer Service, Baggage security. I have briefly touched all three broad areas as follows:

4.5.1 Flight Schedule

Most of the travellers have complaint about the effective schedule of airlines. It has been observed that most of the time negative experiences generated whenever a scheduled flight gets delayed or canceled. In that circumstance, passengers either required to pay extra amount to buy any alternate tickets or stay at airport for indefinite period.

4.5.2 Customer Service

Airline passengers faced rude and unfriendly behavior from customer service representatives/agents/air crews/other flight attendances. This type of behavior always generated negative customer experience and impacted the brand image of an organization. During an event of flight delay or cancellation, such unprofessional behavior can fuel the dissatisfaction even more.

4.5.3 Baggage Security

It has been observed that airline travellers have an issue with lost baggage followed by flight cancellation or delay. In some cases, traveller reached the destination without baggage. This type of incident always raised question about the credibility and reliability of service provider.

4.6 Recommendation

Following recommendation can be considered to resolve root causes and increase customer satisfaction.

- In case of any delay for a scheduled flight, airline operators should arrange meal for the travellers as a token of gesture. If the flight delays for a longer duration, hotel accommodation should be arranged with necessary meal voucher. In case of any cancellation, airline should take the responsibility to arrange an alternate plane from other available operators without any additional payment from the travellers. Most importantly, the customer service should update travellers on a frequent basis.

- It is important that customer service employees or any front line employees of airline industry show proper respect and attitude towards a customer. Airline industry should maintain service level KPI's where employees performance will be evaluated based on their achievement on service level KPI's. Adequate training on customer service is mandatory where employees should learn how to deal with a customer during regular and crisis period. In case of airline industry, employers should consider soft skills as important as subject matter knowledge during the hiring phase.

- When a passenger deposit his/her baggage it is airline's responsibility to keep it in a secure place and return it to the passenger upon arrival in the destination. The airline authority should take measures to ensure that all the baggage's transfer into the flight and handle carefully during shipment. In case of any lost occurrence, airline operators should take the responsibility to compensate the passenger and show their sincere effort to find it.

4.7 Naïve Bayes Model

In this final step, I have used Naïve Bayes algorithm to check the accuracy of my model based on the unigrams. To check the model I have selected 100 most important terms / unigrams as sample.

```
> confusionMatrix(predictions$class, predictions$class)
Confusion Matrix and Statistics

          Reference
Prediction negative neutral positive
  negative     4620       0        0
  neutral         0    6159        0
  positive        0       0     3648

Overall Statistics

               Accuracy : 1
                 95% CI : (0.9997, 1)
    No Information Rate : 0.4269
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: negative Class: neutral Class: positive
Sensitivity                   1.0000         1.0000          1.0000
Specificity                   1.0000         1.0000          1.0000
Pos Pred Value                1.0000         1.0000          1.0000
Neg Pred Value                1.0000         1.0000          1.0000
Prevalence                    0.3202         0.4269          0.2529
Detection Rate                0.3202         0.4269          0.2529
Detection Prevalence          0.3202         0.4269          0.2529
Balanced Accuracy             1.0000         1.0000          1.0000
```

Evaluation phase: I have predicted the classes of the test instances. I have checked prediction against reality.

Performance criteria: I have evaluated the performance through Confusion matrix. From the above summary table, we can observe that out of 14427 instances, we have correctly predicted all the 14427 instances. So, the accuracy level is 100%.

We can observe that out of total 14427 instances, 4620 are negative instances and 3648 are positive instances. We also have 6159 neutral instances.

So, we can summarize our result by stating that all these instances have an impact on the dataset or airline travelers travel experiences. Furthermore, we can state that it is a good model based on the accuracy result.

5. Conclusion

It is important for airline operators to understand traveller's emotions and identify factors related to their services which might affect their brand preferences. Continuous improvement and ability to maintain high standard of service quality are essential elements for airline operators to sustain their business growth in a service driven industry.

However, in recent times, it has been observed that US airlines are struggling to deliver a consistent, positive customer experience. Air travellers have stated numerous issues related to their travel experiences in social media. During this study, I have tried to identify the root causes behind such negative customer experiences by analyzing air traveller's relevant tweets.

Firstly, I have classified all the tweets based on polarity. Secondly, I have tokenized the texts and separated key words / terms based on polarity. Later on I have prepared a model with tokenized terms which assisted to understand traveller's sentiment regarding the services of airline operators. As an outcome of the sentiment analysis, 3 broad areas were identified where airline operators have scope for improvement. These areas are – Flight schedule, Customer Service and Baggage security.

At the end, I have stated recommendations focusing on the identified broad areas. For instance, in case of any delay or cancellation of a regular flight, travellers should be compensated by offering hotel accommodation, meal or arrangement of alternate airline. To ensure positive customer experience, airline employees should give adequate focus on soft skills while dealing with a customer. Adaptability of customer oriented mindset is an essential element in this respect. Furthermore, airline authority should take necessary measures to ensure the security of traveller's goods so that lost occurrences can be avoided or at least bring down to a minimum level.

Hence, I would like to conclude this report by stating the fact that it is important for airline operators to devise long term strategy with a customer focused mindset. **"Customer first"** mindset is the key to ensure positive customer experience and long term business sustainability in this industry.

. . . . . .