|  |  |
| --- | --- |
| X` | MS Business Analytics & Project Management |

**OPIM 5671 Data Mining and Business Intelligence**

**Summer 2016**

**Final Project**

***“Twitter US Airline Sentiment”***

**Manaswini Rayappureddi**

Table of Contents

[Executive Summary 3](#_Toc456175222)

[Business Problem 3](#_Toc456175223)

[Data Overview 4](#_Toc456175224)

[Data Pre-Processing 4](#_Toc456175225)

[Modelling in SAS Miner 5](#_Toc456175226)

[SEMMA- Sampling 5](#_Toc456175227)

[SEMMA- Text Exploration 5](#_Toc456175228)

[SEMMA- Modifying 7](#_Toc456175229)

[SEMMA- Modeling 8](#_Toc456175230)

[**Text-Topic Creation** 8](#_Toc456175231)

[**Modelling:** 8](#_Toc456175232)

[SEMMA- Assess 10](#_Toc456175233)

[**Model Comparison** 10](#_Toc456175234)

[**Scoring Data:** 10](#_Toc456175235)

[Model Recommendation 11](#_Toc456175236)

[Comparison between SAS and R 12](#_Toc456175237)

[Text Topic Comparison for Negative Sentiments 12](#_Toc456175238)

[Topic Visualization using SAS and R 13](#_Toc456175239)

[Model Comparison between SAS and R 14](#_Toc456175240)

[Insights 14](#_Toc456175241)

[Recommendations 18](#_Toc456175242)

[Appendix 20](#_Toc456175243)

[Code for Web-Scraping (Extracting Data from Twitter) 20](#_Toc456175244)

[Code For Modelling in R 20](#_Toc456175245)

[Code For Text Topic(LDA) in R 20](#_Toc456175246)

[Word Cloud 20](#_Toc456175247)

[Bibliography 21](#_Toc456175248)

# Executive Summary

The goal of this project was to find a dataset that could be analyzed given the concepts we covered in this course, with emphasis on applying deeper concepts. The dataset we chose is from a Kaggle competition which focused on sentimental analysis. They provided a dataset with Twitter tweets related to airlines and wanted sentimental analysis on the classification of the tweet, if it was positive, negative, or neutral, and then to focus on the reason why a tweet was negative. Our main focus was to analyze the key topics in texts of Airline tweets using Text modelling in SAS across six different airlines; American, JetBlue, Southwest, United, US Airways, and Virgin America. We performed Text mining in R as well to compare the model accuracy and Latent Dirichlet Algorithm (LDA) using Gibbs sampling process to recreate the documents in corpus by adjusting the relative importance of topics either from previous analysis or from hit and trial method through iterative process and compare the relevant topics in SAS.

We used SAS Enterprise Miner in SAS to analyze 14,640 tweets and created our stop words data dictionary for text parsing to remove the unwanted topics that were not providing relevant ideas about the sentiments. A Decision Tree model was used to predict the overall sentiments obtained from text modeling after using text filtering, which provided us the information about little to no information terms in tweets. After getting the best model, we used the scoring modelling process to analyze the sentiments further on United Airlines data as we found that United Airlines has the most negative sentiments. We have used web scraping to extract the current reviews from Twitter related to United Airlines 2016 data to compare customer feedback with 2015 data. We have compared the models and topics obtained from SAS and R. Common trends for United Airlines were bad customer service, flight delays, flight cancellations, and luggage issues.

# Business Problem

Airline companies will always be the one of the top contenders when looking at reports on the most disliked companies in America. It is an impossible business where one can never make the customers always happy, even though the reasons can be anything, the customers will always hold the airline accountable.

It is also challenging to make the peace in the era of social media. Whenever something bad used to happen on a flight, we would simply share our story with our friends and families. Now, with the advent of Twitter, complaints can now be socialized with a much larger audience. That’s why focus on airline customer service has largely risen to the task of operating a customer service portal through social media. To analyze the problems of each major US airlines further, Twitter data was scraped from February of 2015. With this data, we have tried to capture the actual customer experience and feedbacks across these six airlines using tweets which can help these airlines improve their services because, unhappy customers are the greatest source of learning.

# Data Overview

The data originally came from Crowd Flower’s Data open library source but we took the reformatted version of original data source from Kaggle. It contains the sentiments for six airlines. It has a number of fields including, retweeted count, and the text of the tweets which helped us in analyze the reason for the negative, neutral, or positive sentiment. The Tweet\_location field helped us in visualizing the sentiments across different regions. The dataset consisted of:

* 15 variables and 14,640 observations
* Tweets about 6 different airlines (American, Delta, Southwest, United, US Airways, Virgin America) and their respective unique Tweet Ids
* Variables such as airline sentiment, sentiment confidence, retweeted count, text etc.

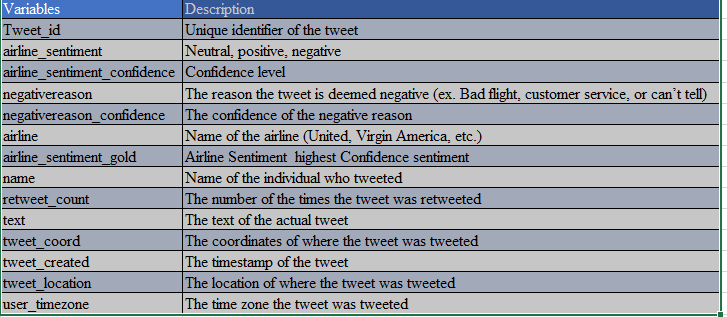


Figure 1 – Variable Description

# 

# Data Pre-Processing

Documents which we wish to analyze comes in different formats so to find the relevant topics in texts we need to text preprocessing in R and SAS. In order to pre-process the data in SAS, we removed the text containing “@airline\_name” from each tweet in the Text field, as we have a separate column for airline name corresponding to the reviews. We observed some missing values in the ‘tweet\_location’ and ‘user\_timezone’ columns but as these variables are not used to predict the sentiments of the data we did not impute these rows, although we did use this data to visualize the outcomes by regions. We converted the Target variable values ‘neutral’, ’positive’, ’negative’ to ‘0’,’1’ and’-1’ by creating a new variable “Sentiments” in dataset for applying additional models such as regression. For further analysis, we used only the Text variable in order to predict the target variable and removed the insignificant terms in Text Parsing node by creating our own data dictionary.

In R, for text preprocessing we created a corpus for the collection of texts and to perform our analysis further, then we have cleaned the data using package tm(text mining package) and have removed numbers ,punctuations, English stop words, stemming terms and terms having less than 3 words which is creating unwanted noise in the data.

# Modelling in SAS Miner

## SEMMA- Sampling

In order to assess the model generalization, we partitioned the data source. We have 14,640 observations related to tweets texts and the sentiment is our target variable which contains the values of -1 for negative sentiments, 0 for neutral sentiments and 1 for positive sentiments which helped us in predicting the reasons related to each sentiment.  We divided the dataset to contain 75% training data for preliminary model fitting as the hold-out sample itself is often split into two parts, validation data and test data. We have considered 25% of data as validation to prevent over-fitting the training data (model fine-tuning), and to compare prediction models. The test data set is used for a final assessment of the chosen model for which we have used web scraping method to extract the current data of US airlines from twitter.

By default, Enterprise Miner uses either simple random sampling or stratified sampling, depending on our target. As our target is a class variable, then SAS Enterprise Miner stratifies the sample on the class target.

## SEMMA- Text Exploration

In terms of text mining, we called exploration of data as Text Visualization. To explore our dataset and examine the variable distribution and statistics, we used Stats Explore in SAS Enterprise Miner. The results showed that airline is the most worth variable for predicting our target. Other variables like name and tweet location showed a worth variable with .015%, but per our understanding these values can help us in visualizing the data, but not much in modelling so we have rejected these variables in our metadata node.

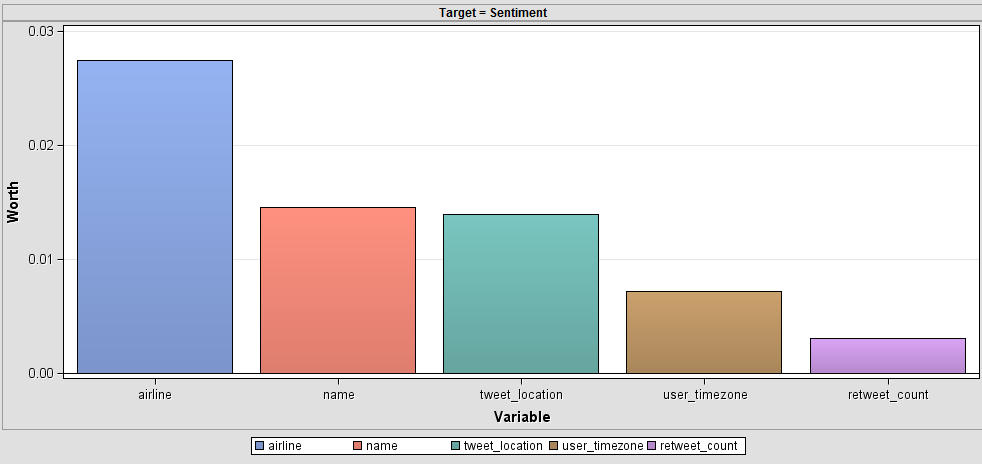


Figure 2 – Worth Variable

We also used the Graph Explore node to explore large volumes of data and found that negative sentiments are 60% more in comparison to positive and neutral sentiments. We also determined United airlines has the most number of negative sentiments and therefore have chosen to make that our primary focus for this project

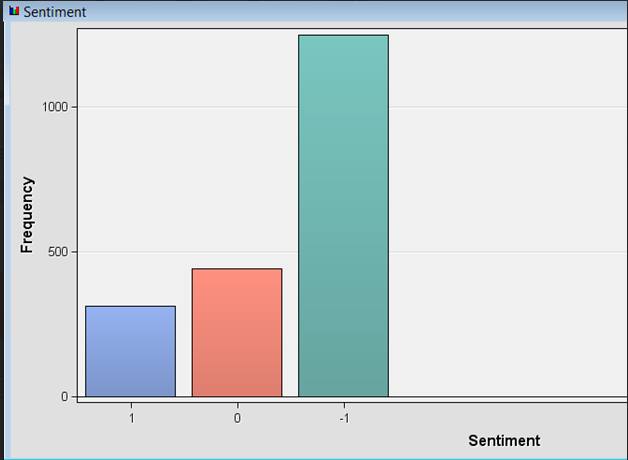


Figure 3 – Sentiment Frequency

## SEMMA- Modifying

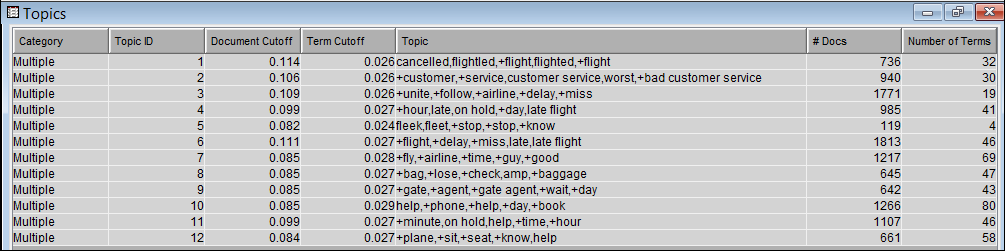
Our data is related to tweets text so we performed text preprocessing to transform the data into a format that will be more easily and effectively processed for the purpose of the user. The following steps are what we performed in SAS as a part of text preprocessing:

1. **Text Parsing:** After extensive exploration of the values SAS added to the stop list, we determined there were terms missing from their list. We also chose to ignore additional parts of speech, such as numbers, conjunctions, propositions, interjections, pronouns, etc. which does not provide any insight in predicting the target variable. After repeated iterations we created our stop list “Airline Stop list “and appended it to default stop list in text parsing node.
2. **Text Filter:** Based on the size of the dataset, we changed the term weight to Inverse document frequency and also changed the Maximum number of topics to 6 to filter out the terms that are not used in at least 6 documents in corpus collection.
3. **Text Preprocessing in R:** We also performed text pre-processing in R by first converting the complete text to lower case and then converted the document to Plain text. Further we have removed the digits in user texts and removed all the punctuations and whitespace to get the most relevant topics in the document. Stop words are also added to remove the most common words in the document along with other frequent terms such as can’t, can, will, have etc. spaces and special characters.

## SEMMA- Modeling

### **Text-Topic Creation**

The main focus in this step was to identify the most frequently occurring terms and topics in the various texts that people tweeted related to airlines feedbacks. This was achieved by using different iterations to text topic node such as setting the multi-term topics to 25 followed by 20, 15, 12 and 10. We found that 12 had the most significant topics.

Figure 4 – Text Topic in SAS

As we see the above screenshot of the text topic results, 12 topics are formed with classifying the data into 12 major topics. Some of the important topics are given below

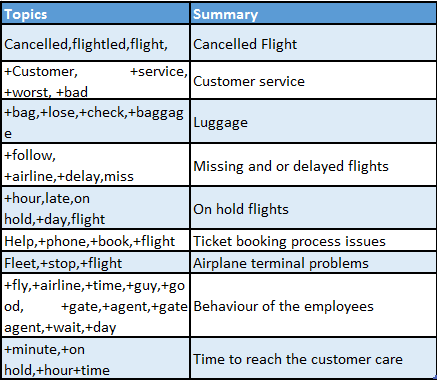
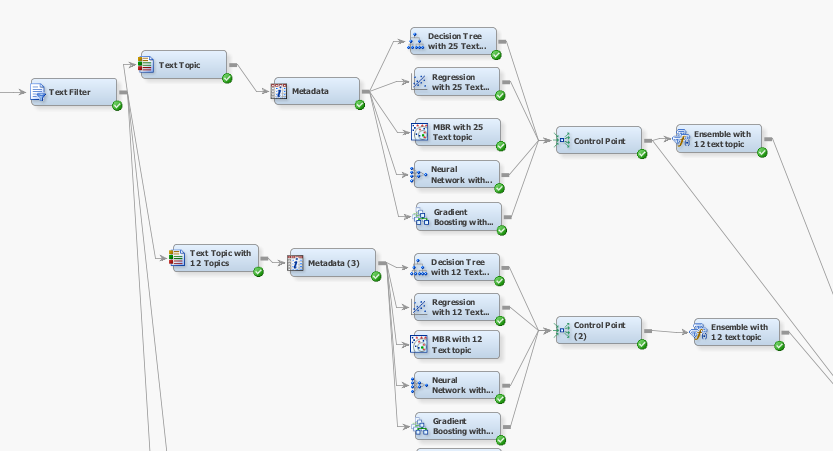


Figure 5 - 12 Significant Topics

**Modelling:** The modelling is done on the output of the text topic and text cluster node. Taking the texttopic\_raw and text cluster SVD as input multiple model are run in SAS using their respective nodes. All the models are run with both default settings and changing the properties and the best models are chosen. Below are the different models performed to predict the target variables.

* Decision tree:
* Regression:
* MBR
* Neural Network
* Gradient Boosting
* Ensemble



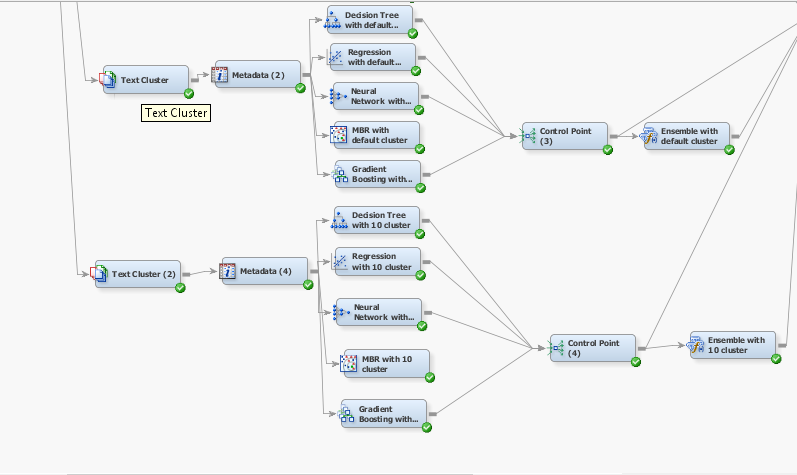


Figure 6 - All of the models

## SEMMA- Assess

### **Model Comparison**

After performing all the models, we compared them using the model comparison node. Looking at the accuracy and misclassification rates all the model were performing almost similar with the misclassification rate ranging from 0.31- 0.37.

Naturally, one would want to choose their model based mainly on the misclassification rate and as per the results of our model comparison we found Neutral Network with 10 cluster has the lowest misclassification rate but Text Cluster node is suited for documents that generally focus on a particular topic because when multiple concepts are present in a document, the chosen theme could be 'biased' (for lack of a better word) that’s why we decided to go with Topic modelling and found that Decision Tree with Text topic 25 having 34.8% misclassification error but keeping complexity and being able to explain our model to the business, we decided to choose the Decision Tree with 12 Text Topics model.

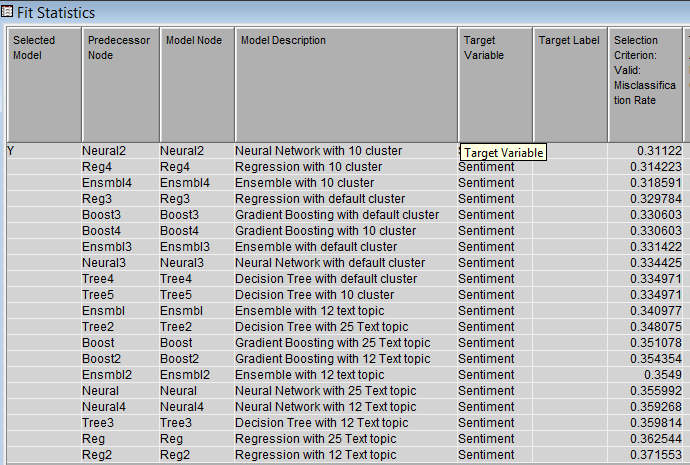
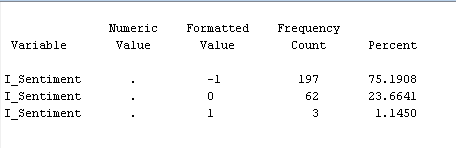


Figure 7 - 12 Text Topics

### **Scoring Data:**

To evaluate the modelling results, we performed web scrapping of twitter data to extract the 2016 tweets for United Airlines from Twitter because we found the most negative sentiments are about United Airlines after analyzing the dataset. We attached a score node with new score data set that we pulled directly from Twitter in order to evaluate our model and have classified these reviews based on the best model we selected. The Score data consists of 262 twitter reviews about United Airlines. After scoring the data using the decision tree model with 12 text topics, the data is predicted as 197 negative reviews, 62 neutrals and 3 positive reviews.



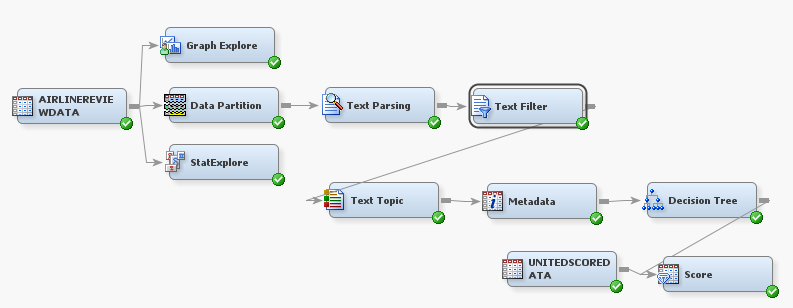
 Figure 8-Score Data Outcome

Figure 9 – Final Model

Model Recommendation

We recommend the Decision tree with twelve Text topics as the best model. Even though the misclassification rate is larger, the model is not complex compared to the other MBR, ensemble models. And also the model is easy to interpret with the text topics. Decision tree clearly segregates the text topics based prediction into appropriate sentiment based on SVD weights and helps a user to understand the process how the model functions. These factors also make this type of model easily sellable to the customer and any iterations or modifications to the model are can easily be interpreted by the customer as well.

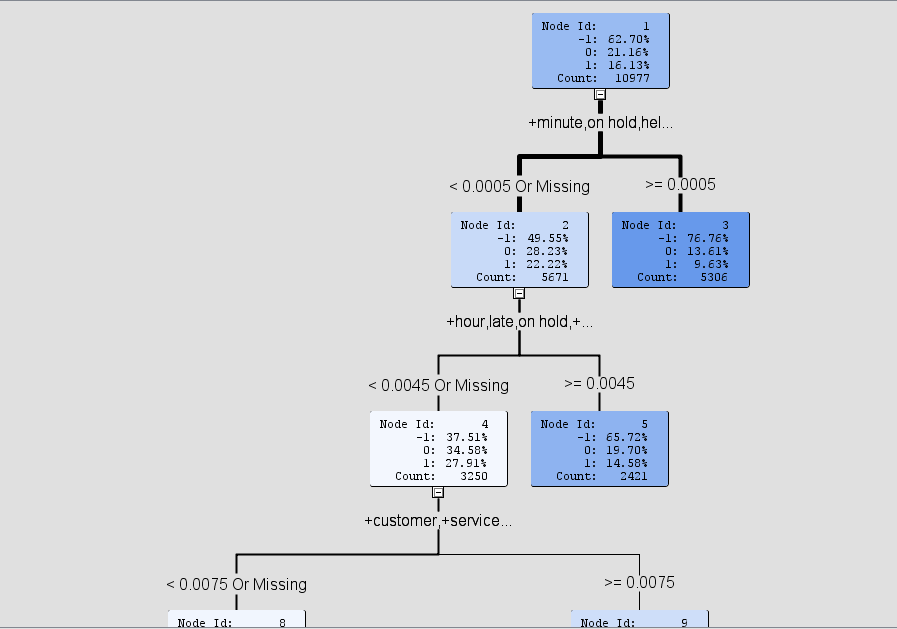


Figure 10 - Decision Tree

# Comparison between SAS and R

# Text Topic Comparison for Negative Sentiments

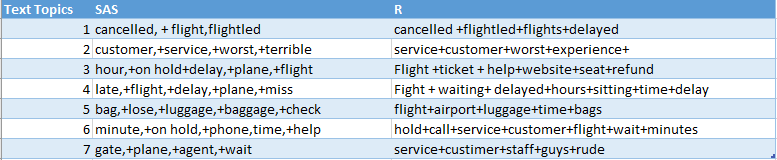
When comparing the text topics, we found very significant and similar topics in SAS and R related to the negative reviews, yet R has given slightly better information. For example, bag+ lose +luggage shows that baggage loses but in R flight + airport + luggage + time + bags shows that baggage loss on airport causes delay also in seventh text topic Agent, gate, plane, wait does not provide much information but in R - service, customer, staff, guys, rude shows that customer service guys are rude. 

Figure 11 - SAS vs. R Text Topics

## Topic Visualization using SAS and R

We used LDAvis package for visualization of topic and visualization is based on Json script. It is divided into two parts where left part represents the overview of the topic model where each topic is plotted as circles whereas right part of visualization is in the form of horizontal chart where each bars represents the significant terms used in determining the topic on the left part.

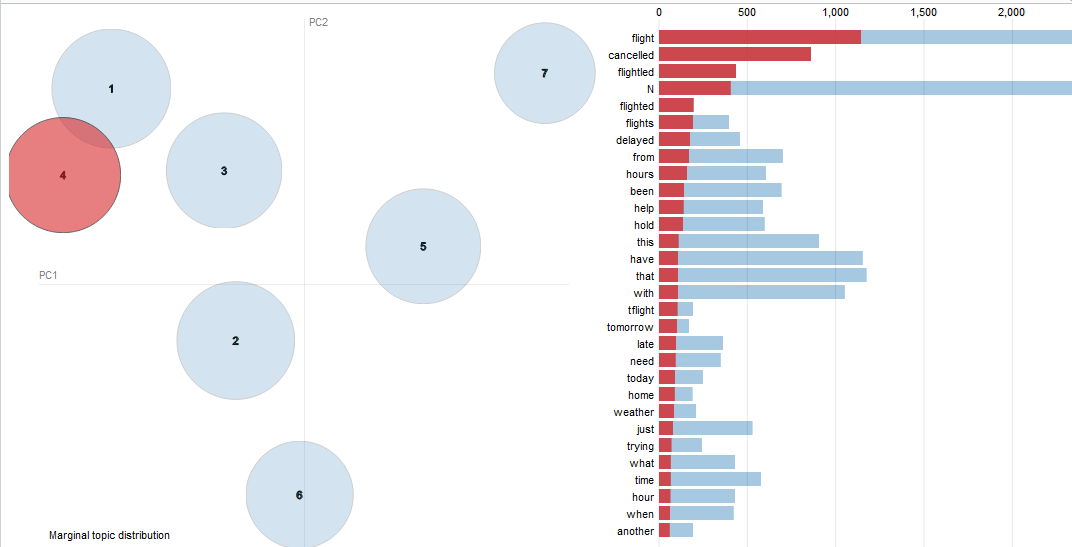
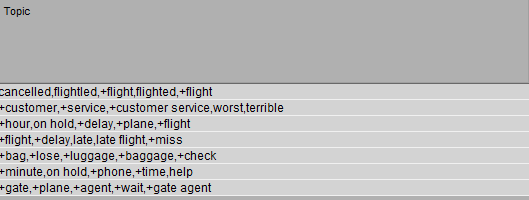


Figure 12 - LDA Distribution of Negative Reviews in R

Figure 13 - Text Topics for Negative Reviews in SAS

## Model Comparison between SAS and R

We compared the models between SAS and R and found that SAS performs better in terms of accuracy, yet the models perform similarly. Regression in SAS gave higher accuracy over R in our testing dataset. Random forest is not currently available with SAS Enterprise Miner but our experience shows that Gradient Boosting provides as good or better results in less time on the dataset.

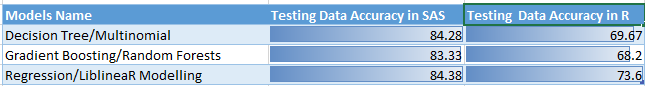


Figure 14 - Accuracy between R & SAS

# Insights

The dashboard below provides a visual way to quickly conclude on valuable insights. For example, at a glance you can see that United Airlines has the most negative sentiments, followed by US Airways and American Airlines. You can also quickly discern the locations which are receiving the most negative tweets. From United Airlines perspective, they can see that Washington DC has drastically higher negative comments than positive. That is valuable information as that will drive their business strategy on how to resolve these issues.

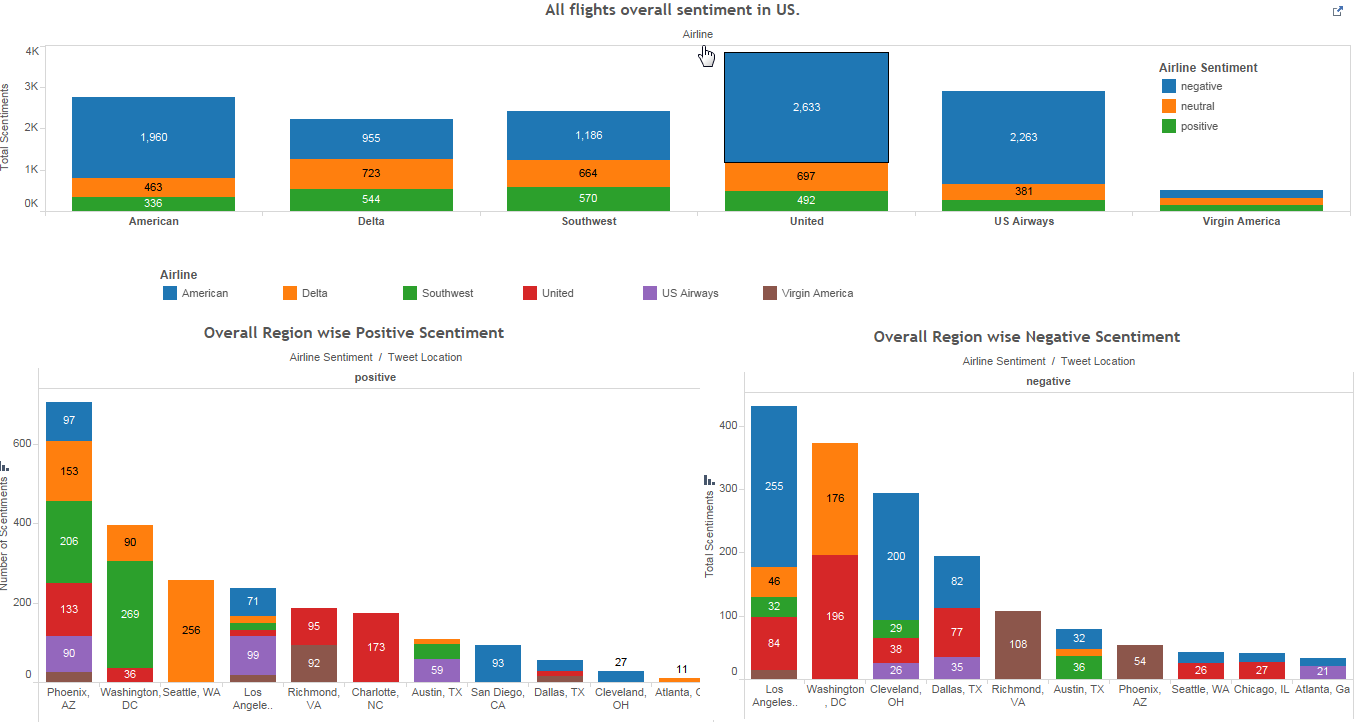


Figure 15 - Dashboard 1

The next graph shows each airline’s most prominent negative reason, which again will equip them the information required to solve their problems. For example, if US Airways looked at this, they should address their customers concerns about long phone lines.

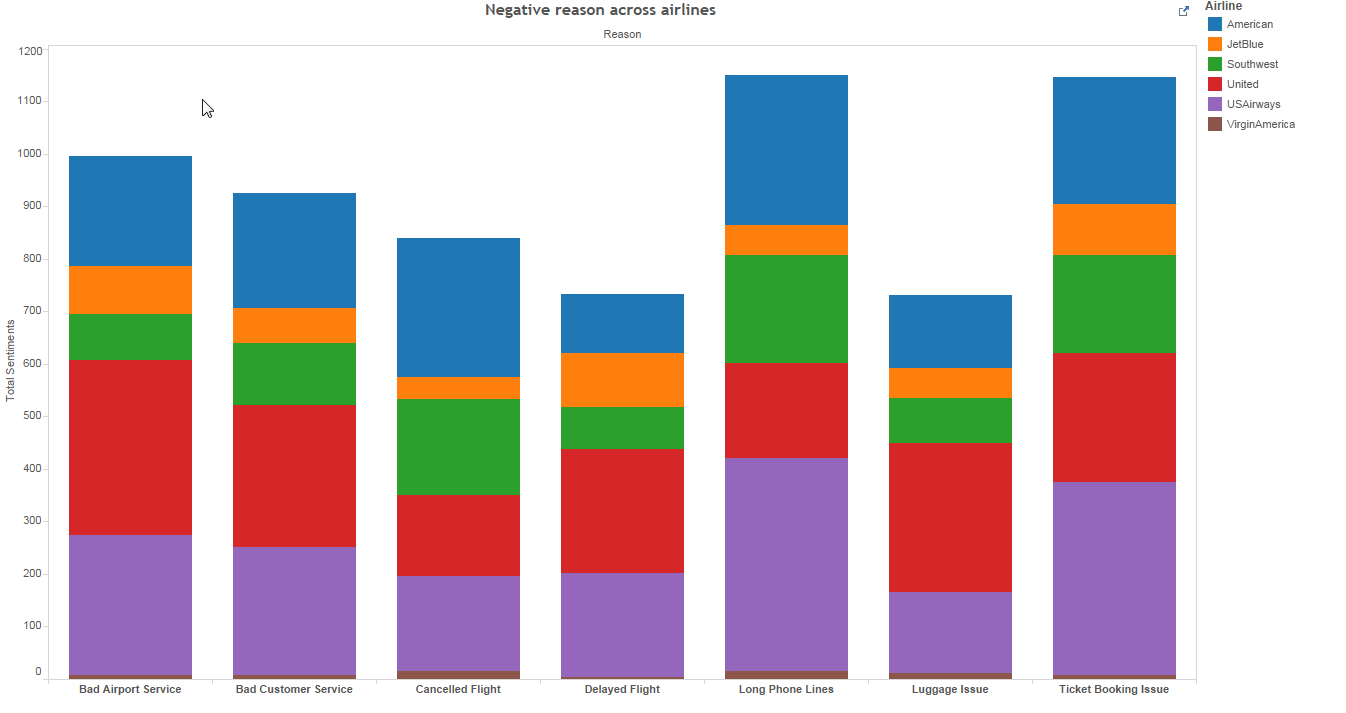


Figure 16 - Dashboard 2

As mentioned previously in this report, we decided to take a closer look at United Airlines as they have the most significant negative sentiments out of all the airlines. We put together additional graphs that will provide quick insights. The graphs below show United that their top four most tweeted about regions are Pheonix, Charlotte, Richmond, and Washington DC. Of those, they should all be explored further as they have prominently negative tweets coming from those locations.

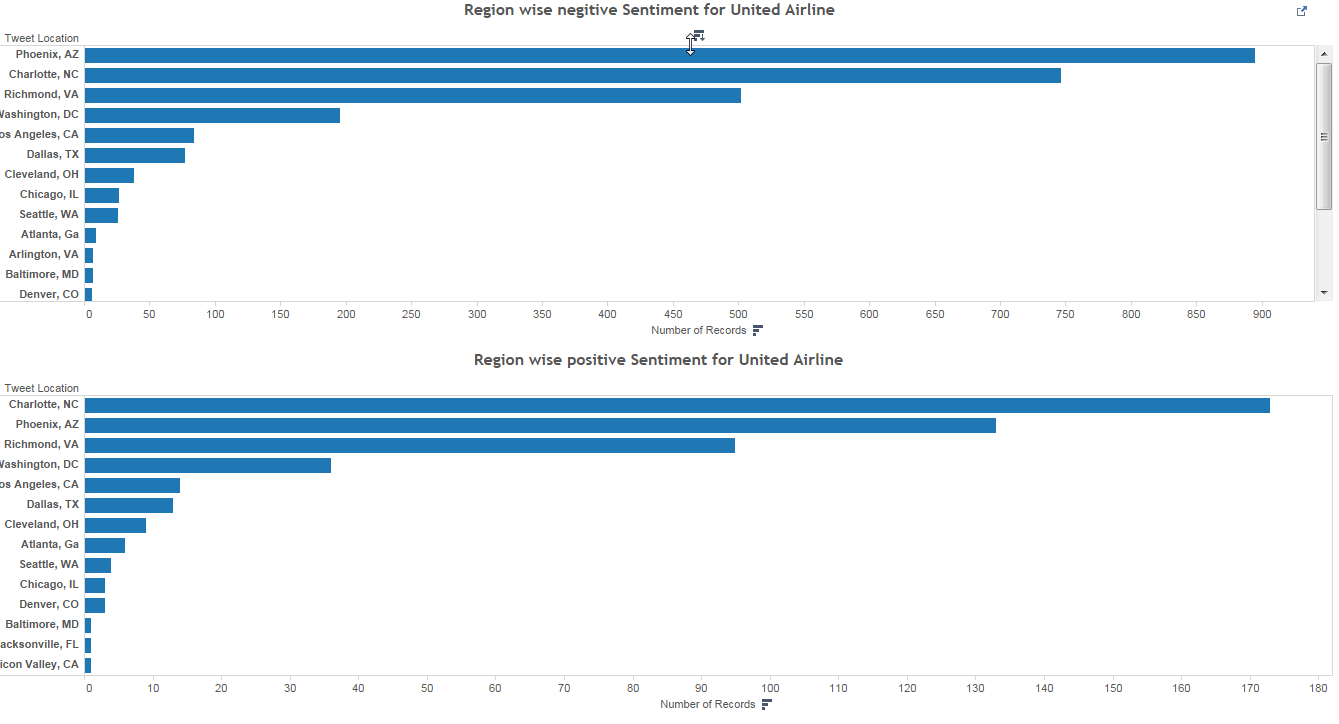
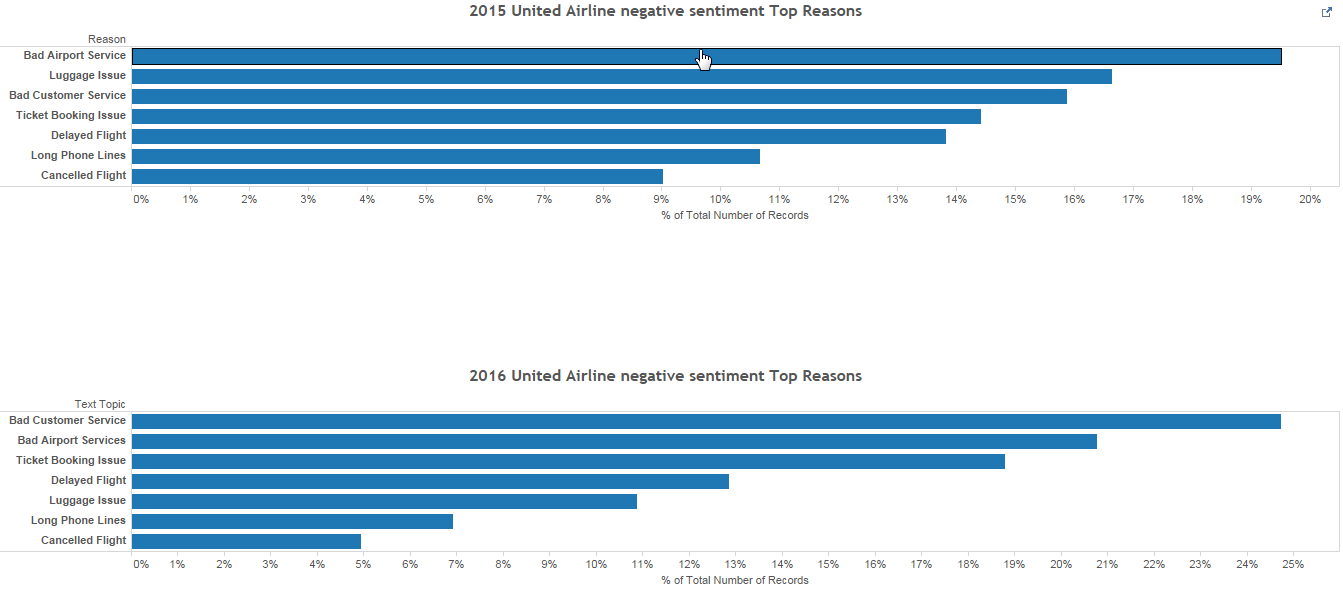


Figure 17 - Dashboard 3

The next graph equips United Airlines with the information needed to see the change in primary negative tweet reasons from 2015 to 2016. We have classified 2015 data into the seven Negative reasons based on the text topics formed and analyzed the data in the below graph. For the score data also we applied the text topics are scores the 2016 data taken from twitter to classify them into seven negative reasons. Based on the output of the score node, graphed the data as below.

Figure 18 - Dashboard 4

# Recommendations

The analysis of customer sentiments provides very deep insights into what’s driving the customer sentiment and hence might be impacting the revenues of the company. The airline can analyses this data based on various dashboard provided both for the current data and future predictions. For example, United airline can also look at why they are being blamed for bad airport services, and identify what they are they doing differently from other airlines in the same airport. The data also shows us that the luggage issues have decreased, but are not completely resolved. They can further investigate this and determine what is driving the reduction and then concentrate on enhancing that. Issues with ticket booking are also very prominent and look as though they are increasing this year compared to last year. It would be a good strategic decision to quickly look into what’s causing it, web application, poorly trained staff, etc. to resolve them. IT could further assist them in doing Time series analysis based on data by individual text topic and make the airline aware of what’s coming for them if they decide or not decide to focus on a specific reason for –ve customer feedback and working on a way to remediate it.

# Appendix

## Code for Web-Scraping (Extracting Data from Twitter)

****

## Code for Modelling in R



## Code for Text Topic(LDA) in R



## Word Cloud

Performed the word cloud in R using the word cloud package for the twitter review and we can see that the terms ‘flight’,‘cancel’, ’thank’, ’delay’, ’service’, ’custom’ and ‘call are significant followed by the terms ‘hold,’book’,’gate’.



# Bibliography

<https://www.kaggle.com/crowdflower/twitter-airline-sentiment>

<https://eight2late.wordpress.com/2015/09/29/a-gentle-introduction-to-topic-modeling-using-r/>

https://support.sas.com/resources/papers/proceedings14/1288-2014.pdf

http://www.sas.com/content/dam/SAS/en\_my/doc/training/TASM.pdf

http://www.slideshare.net/ajayohri/twitter-analysis-by-kaify-rais