Aspect based sentiment analysis for United States of America Airlines

MSc Research Project
Data Analytics

Swapn Joshi
x15043517

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Swapn Joshi
Student ID:	x15043517
Programme:	Data Analytics
Year:	2016
Module:	MSc Research Project
Lecturer:	Dr. Catherine Mulwa
Submission Due Date:	11/12/2017
Project Title:	Aspect based sentiment analysis for United States of America Airlines
Word Count:	6043

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	11th December 2017

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Aspect based sentiment analysis for United States of America Airlines

Swapn Joshi
x15043517
MSc Research Project in Data Analytics

11th December 2017

Abstract

Around the world, internet plays a significant role when it comes to decisions making. Nowadays, a large population of people shares their opinions and views on the internet through social media, blogs and other online platforms. This leads internet to be full of information both relevant and irrelevant. Therefore, in order to get the desired information, it is not possible to go through each document present on the internet. Here, Sentiment analysis acts as a panacea to this problem. This research aims to provide a decision support for the customers for selecting the best fit US-based Airline, by providing an Aspect level sentiment analysis from the other customer's opinions present in micro blogging site Twitter and online review site Skytrax. The proposed research will follow a modified Knowledge discovery and data mining (KDD) methodology. Several machine learning algorithms are applied in order to find out the best-fit algorithm for the system. Also, evaluation is measured based on the performance matrix for the system.

Keywords: Aspect-based sentiment analysis, Machine learning, Performance matrix.

1 Introduction

In this innovative and quick moving world there is a freedom for each and everyone to express themselves, their views or opinions on any platform. People express their thoughts from print media to social media platforms in order to render their experience about the product they use, about the gadget they buy, the food they eat and the places they visit. This not only gives them a chance to share their view but also gives other users an advantage or benefit before they visit, buy or try things in future.

Consequently, people utilize other people's opinion and take suggestions before making any decisions in order to get a better and a hassle-free experience. This not only saves their time but also their money, energy, and efforts. Since anyone can post his/her views on the internet, sometimes it becomes very difficult for people to go through each article, tweet post or review in order to get the information they wanted. Considering a lot of information available on the internet available is irrelevant and is not useful. Hence, a lot of time and energy is wasted if we are not able to find out the relevant information

which we are looking for.
In these scenarios, Sentiment analysis plays a critical role. A sentiment analysis, also known as opinion mining is a process which extracts the polarity of a text, sentence or an opinion by applying text mining and natural language processing (NLP). The polarity is in the form of positive, negative and neutral based on the tonality of the text. Also, sentiment Analysis lies in the domain of computational linguistics and due to the exponential use of social media, it has been getting a considerable degree of consideration in the last few years (Chen and Zimbra; 2010). It can also be seen in the Google trend graph show in Figure 1 for the last decade.

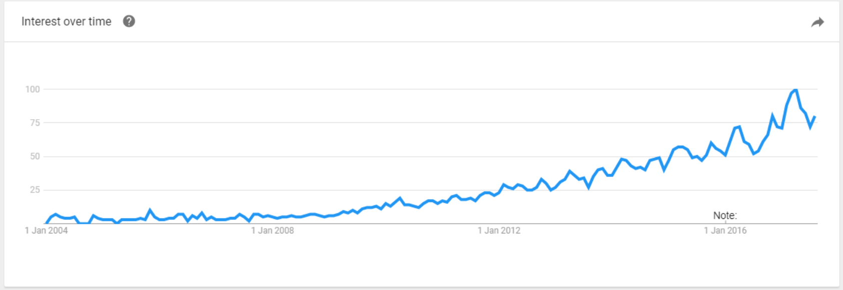


Figure 1: Sentiment Analysis trend over the past decade

Previously, in order to get customer reviews or opinions, people used to prepare questionnaires and surveys and based on the answers, they got the insights about their services from a customer's perspective. This process used to be very time-consuming and at times the customers did not answer the questioner's questions seriously and sometimes left them blank leading to very misleading information. Furthermore, the questioner questions were designed especially to their requirements and were not available to the public. In today's world with the help of social media and online review platforms, it become easy for customers to give their reviews as well as for companies to analyze them. In this research, microblogging site Twitter is used, since Twitter is one of the most popular social media sites across the globe. In Twitter, the user delivers his/her opinion in just 140 words and there are nearly 1 million Twitter users. In general, approximately 250 million tweets are posted on Twitter (Wan and Gao; 2015). The second dataset for the research is the online review site Skytrax, here customers write their experience about their flights and it also has a section of verified users, where only verified users provide their reviews. It removes the barrier of spam reviews which makes it more valuable for our research.

1.1 Background and Motivations

The airline industry is one of the leading industries in the world, providing services to thousands of customers in a single day. In the reports of Federal Aviation Administrations Air-Traffic (FAA), it is found that every day approximately 2,246,000 passengers take flights in the United States of America (USA) (Anitsal et al.; 2017). The research for this project is focused on the top ten US-based airline carriers namely Alaska Airlines, America Airlines, Delta Airlines, Hawaiian Airlines, Jetblue Airlines, Skywest Airlines, Southwest Airlines, Spirit Airlines, United Airlines and US Airways. The primary reason for choosing these airlines is that these airlines fly across the same geographical area i.e.

the USA. Also, these are the lost cost carriers present in the USA and have a similar flight fare. Furthermore, there is a great competition among them and each airline wants to have a good competitive edge from their competitors.

To the best of the candidate's knowledge, there has not been much research done on the airline industry in the context of aspect based sentiment analysis. This research will attempt to bridge the gaps between the customers' views and airlines carriers.. Besides, the proposed research can further be implemented in other domains such as education, automobiles, entertainment etc.

1.2 Project Specification

The following project specification has been suggested in order to get the aspect based sentiment analysis for the US-based airline. This proposed model will be an asset for the customers in selecting the desired airline in the United States. Also, it will provide a check for the airline carriers. The research question and research objective are shown below:

1.2.1 Research Question

"Can aspectbased sentiment analysis of 10 US Airlines namely Alaska Airlines, America Airlines, Delta Airlines, Hawaiian Airlines, Jetblue Airlines, Skywest Airlines, Southwest Airlines, Spirit Airlines, United Airlines and US Airways, using supervised machine learning techniques provide insightful information, which can be used to enhance/or improve the US Airlines industry?"

Sub Research Question: *"An critical investigation of aspect-based sentimental analysis in American airline industry"*

1.2.2 Research Objectives

In order to answer the research question, the following research objectives needs to be addressed.

- Objective 1: Perform extraction and transformation of reviews which will involve processing and cleaning.
- Objective 2: Finding aspects and factors from the online reviews for the 10 airlines (Alaska Airlines, America Airlines, Delta Airlines, Hawaiian Airlines, Jetblue Airlines, Skywest Airlines, Southwest Airlines, Spirit Airlines, United Airlines and US Airways) using aspect based sentiment analysis.
- Objective 3: statistical operation to find relationship among the aspects and derive the ranking of airlines on that.
- Objective 4: Comparison of multiple machine learning techniques (i.e. SVM, Decision tree, Random Forest, Bagging and Boosting, SLDA and Maximum Entropy). The aim is to get Precision Recall and F-score to find which algorithm produces

best result to support aspect based sentiment analysis.

In addition results are analyzed on the basis of different visualization methods like d3.js and Tableau to find What factors effect the most and attracts the least for an established airlines.

Hypothesis: Assume,Hypothesis is represent by H and AN is represented as Airlines name where I is the input and O is output.
A <= Aspects
S <= Sentiments
Therefore, Ho: I gives O where O is the Subset of {AN,A,S}

1.3 Research Contribution

After the critical review of the related field, gaps are recognized which results in the following contribution to the body of knowledge.

- A fully developed aspect level sentiment analysis for the top ten US based airlines.
- In determining the relationship between the aspects of the reviews.
- A critical comparison between the evaluated results of the different models(Random forest. decision tree, SVM).
- The configuration manual utilized in the development and implementation of the project.

In order to answer the proposed research question, the project follows a modified KDD methodology. The methodology stages are modified according to the need of the project and it fits best for this research.

1.4 Conclusion

The result of the reviewed work will provide a robust decision support for the customers which will not only help them in making an effective decision for choosing their airlines but also help the airline companies to look after the areas of improvement. This will also help them to get an competitive edge over their other rivals/competitors in the airline industry.

Further, the remaining research is organized into the following sections. Section 2 outlines the related work done under the field of aspect-based sentiment analysis and research done in the airline industry; Section 3 consists of the methodology used in the research and the process flow of the research. Section 4; will have the implementation architecture of the research and the implementation process. Section 5; describes the results of the evaluation model in the research, followed by a summary of the research and discussion for the potential areas of future work in the area.

2 Literature Review of Aspect Based Sentiment Analysis for US Airlines (2010-2017)

2.1 Introduction

In recent years, marketing styles and approaches have significantly changed (Mangold and Faulds; 2009). A lot of organizations analyze and rely on customers' opinions in order to get valuable insights about their services and goods from different online platforms like social media, blogs or online review portals. Consumers also utilize these channels not only to express their opinions but also to get the views of other consumers about the commodity or services. Hence, customers' opinions play a very important aspect in this buyer-seller relationship. With the increase in digital marketing and opinion sharing on virtual networks, many researchers have been attracted towards opinion mining and text analysis in the last decade. Opinion mining is not only profitable for the reviewer but also it is important for a service provider to find the major strengths and drawbacks in their business models by having a direct view from their customers.

The work of Jindal and Liu (2006) defines opinions in majorly two forms, Regular opinion and Comparative opinion. In a Regular opinion, the person conveys their perspective in the form of a positive, negative or neutral comment about a product. Whereas, a Comparative opinion can be defined as when a person conveys a comparison between products while expressing his/her opinion.

For example- Alaska Airline customers satisfaction is better than that of Jetblue Airlines. Here there is a comparison of customer satisfaction between the airline carriers.

Further sentiment analysis can be derived from a deeper analysis of other areas. The next section is reflecting one of the most important fields of sentiment analysis, i.e. aspect based analysis. Aspect based sentiment analysis is important to find the polarity of text in the sentiment analysis.

2.2 A critical review of Aspect-based sentiment analysis

Understanding opinions is a vital part in sentiment analysis but an opinion only exposes the polarity of a text in a sentiment analysis. Collomb et al. (2014) classify sentiment analysis into three levels which are document level, sentence level and word/feature level or can be said aspect level.

In the document Level, the document is consider as a whole and taken as either positive or negative but the only drawback of this is that we don't know which part of the document is positive or negative. Whereas, in the sentiment and Aspect level, the opinion is retrieved at a finite level and we can analyze which part or aspect the customer liked or disliked.

This research will be carried out on aspect level of sentiment analysis for both regular and comparative opinions.

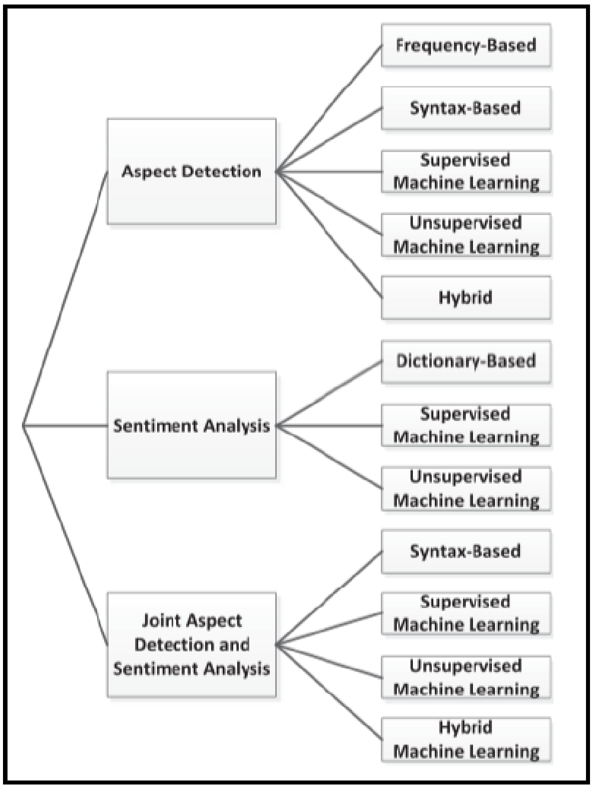

Aspect Detection
Frequency-Based
Syntax-Based
Supervised Machine Learning
Unsupervised Machine Learning
Hybrid
Sentiment Analysis
Dictionary-Based
Supervised Machine Learning
Unsupervised Machine Learning
Joint Aspect Detection and Sentiment Analysis
Syntax-Based
Supervised Machine Learning
Unsupervised Machine Learning
Hybrid Machine Learning

Figure 2: Aspect based sentiment analysis approaches (Schouten and Frasincar; 2016)

2.3 Identified Research Gaps in Airlines Industry

In recent years, sentiment analysis has been performed in various domains from analyzing movie reviews to stock marketing analysis but little research has been done in the airline industry. Wu and Liao (2014), looked at the leading and lagging parameters for 38 airline companies from the annual and business reports generated from data envelopment analysis(DEA). Similar to that, Hannigan et al. (2015) analyzed the relationship between the different performance factors in the United States of America based airlines from 1996-2011 from the annual reports provided by DEA. In the results, they found that there is a positive relationship between the price and performance (in terms of stock market shares) over the years using regression models and found a negative relationship with the quality of the services. The authors discuss future work in terms of analyzing other data resources that could bring different insights to the research. Sultan and Simpson Jr (2000) gathered surveys from US and European travellers and built a SERVQUAL model from customer expectations and perceptions towards an airline's performance. In the model, the reviews were segregated into five factors i.e Reliability, Assurance, Tangibles, Empathy and Responsiveness. Similarly, Min and Min (2015) noticed 18 factors that can provide a benchmark in the airline services. Extending the work of Min and Min (2015), Anitsal et al. (2017) analyze the sentiment of passengers for top ten U.S airlines from Skytrax. The researchers generated a sentiment score and word cloud of the reviews for each airline using the tool Semantria. Skytrax is the leading global airline consultancy firm which conducts annual surveys and conducts star based analysis under a global airline rating program (1-5) (Yakut et al.; 2015). The program gives feedback for more than 670 airlines performance globally (Pérezgonzález and Gilbey; 2011). Out of many researchers Jansen et al. (2009) mention free accessible, experienced auditing, the unique ranking system and competent textual reviews as a major backbone of Skytrax rating system. In 2009, Skytrax faced controversies as mentioned by Jansen et al. (2009), out of all the airline open for review only 25 percent were scored that do not portray a real picture of the Airline industry. Hence, for a better textual analysis, heterogeneous sources are needed to cover the full range and all aspects of a particular industry. Safko (2010) mentions that social media is a reliable source of data for sentiment analysis as the data on such media keeps on updating. However, Ivanscenko (2016) points data collection and textual analysis for the random data makes the process more difficult, but Yee Liau and Pei Tan (2014a) state text mining as a major contributor in the findings of brand awareness, loyalty and recognition in the aviation industry.

Li (2017) has focused on the gaps in sentiment analysis in order to fulfil criteria like reliability, discriminate validity, and external validity for airline review site SKYTRAX. The study reveals how research on SKYTRAX using numerical rating is non-reliable to judge the performance of any airline. According to Kaur and Duhan (2015), converting textual information into a meaningful analysis faces six critical issues: negation handling, domain generalization, pronoun resolution, language generalization, related knowledge and mapping slangs. Li (2017) has done an exceptional research in handling these hurdles to an acceptable degree, but there is still some room for errors while dealing fake or spam reviews.This research is helpful for proceeding our research towards the aspect based analysis as Document-based analysis lacks for extracting features from the content.

As Li (2017) has researched strictly around airline industry deriving data from Skytrax and twitter, there are some points which lead this research under few doubts. External validity is an important part of the Skytrax global rating system and this feature has not been tested by the researchers. According to Cachia et al. (2007) organising unstructured data in a systematical way can be an issue as data is derived from many sources like newspapers, blogs, social media etc., which can lead to ambiguity and repetition of data. Text analysis faces the problem of normalisation which should be handled carefully, and this feature is very rare to come across in the existing research. The airline industry is being researched and analysis is done by many researchers, out of those Li (2017) performed most efficient manner of sentiment analysis. But author follows document based analysis which can further be improved by aspect based analysis on heterogeneous data source.

Sankaranarayanan and Rathod (2017) classify the airline ratings for low-cost Indian airlines. The research was done on three low-cost Indian airlines from the review ratings generated from the Trip Advisor website using Expectation Maximization clustering techniques. Further, for prediction accuracy, a comparison was drawn among various machine learning techniques like Kmean, Naive Bayes and logistic model trees. The authors suggest a future work in considering text mining for the reviews in order to find the reason behind the rating of the reviews. There were other limitations in the research like missing data values and not considering other important aspects like baggage issues and seat comfort in the dataset that should be addressed. With the rising competitive conditions in the low-cost carrier (LCC) airline industry, a notable work has been accomplished in the past. O'Connell and Williams (2005) conducted a passenger survey for two Malaysian Airlines (Malaysia Airline and AirAsia) and found very interesting insights about the passengers thinking for choosing an airline. In the research, it is observed that the young age group of passengers was highly attracted towards the LCC airlines because of their ticket cost whereas other segments of passengers chose their airlines on the basis of the staff services. It is found that the Air Asia airline staff have a very high productivity level which results in excellent customer satisfaction. Similarly, in the research of Saha and Theingi (2009) on Thailand Airlines, they found customer satisfaction as the most important factor in the buying behaviour of the customers. Dobruszkes (2006)researched the European LCC and noted a significant decrease in the cost of airline fares and an increase in air traffic. The researcher suggested that in order to be competitive in the market of LCCs, organizations should pay significant importance to their customers welfare.This proves that Airline industry need features to be focused on which can be derived from the reviews of the customer.

There has been ample study conducted towards the LCCs but solely through employing passenger surveys. In contrast, Yee Liau and Pei Tan (2014b) analysed customer opinions through the microblogging site Twitter for LCCs in Malaysia; Twitter tweets were collected for two and a half months for the Airlines. Since most of the tweets were in the local Malaya Language, hence, a Malay lexicon was built for sentiment analysis and further the tweets were segregated into four different clusters, i.e. ticket promotions, flight cancellations, delays and customer satisfaction. Clustering is performed by using Kmeans and spherical Kmean algorithms. In the results, there was an overall satisfaction with the services provided by Malaysian LCCs. Apart from the research in LCC airlines, there has been some other research done toward the airline industry. Dharmavaram Sreenivasan et al. (2012) analysed sentiment polarity for three airline groups from the tweets extracted for a period of four months using lexicon analysis. Miner (2012) conducted research on predicting the airline ranking for US based airlines. The research follows the Variational Expectation Maximization (VEM) technique to determine the airline ranking and the Correlated Topics Models (CTM) for extracting the sentiment topics. Also, Baumgarten et al. (2014) predicted the delay in flights by analysing the departure and arrival times of the US-based flights.

2.4 Conclusion

To the best of the candidates knowledge, it is clear based on the reviewed literature that there is an inadequate amount of research done in determining the Aspect level of consumer sentiment in the airline industry. Most of the research focuses either on finding the sentiment polarity of the consumers experience with airline services or focuses on determining the rating or delay time in the airlines. Also, in order to get proper insights into customer behaviour towards airlines, utilizing surveys or Twitter feeds for two to three months is not adequate; hence, there is a need for heterogeneous data sources. Based on the identified gap there is a need to develop an Aspect-based decision support system for the US Airline Industry.

3 Methodology

3.1 Introduction

The Methodology section comprises, the process flow of the research along with the discussion of the methodology used in the research and a brief explanation of each part of the methodology. Lastly, there is a conclusion of this section.

3.2 Modified KDD Methodology

The process of extracting essential insights from the data to provide a better decisions support is known as Data Mining (Berry and Linoff; 1997). Methodologies such as knowledge discovery and data mining (KDD) and Cross-Industry Standard Process for Data Mining (CRISP-DM) are widely used for data mining (Azevedo and Santos; 2008).

The CRISP-DM methodology was introduced in 1996 and follows a six-stage process flow (Chapman et al.; 2000). Since in this research, stages like business understanding and development cannot be utilised, hence CRISP-DM methodology does not fit into this research. Furthermore, This research will follow a modified KDD methodology where a few stages are modified e.g. implementation stage modified to Aspect and sentiment detection stage and additional attributes are added. Hence, the modified KDD methodology fits best for the research. The proposed methodology is shown in Figure 3.

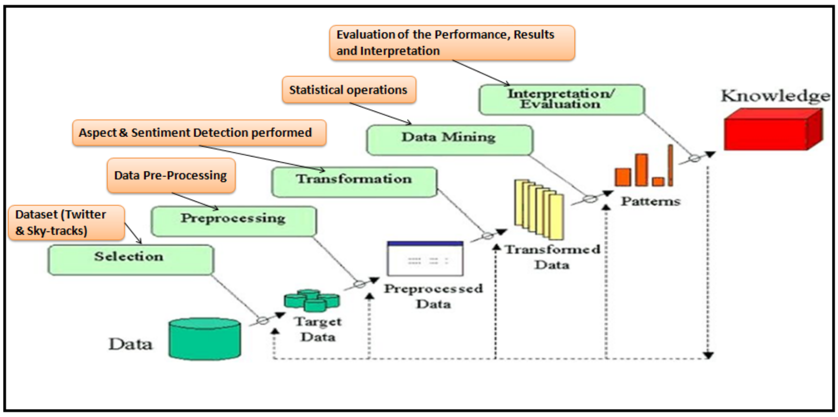


Figure 3: Modified Methodology

3.3 Data Collection

This is the first step of the methodology. The data is extracted from various sources like Twitter tweets and online reviews from Skytrax (2014-2017) for the top 10 US-based airlines. The main purpose of collecting the data is to find out the leading and lagging performance parameters among the airlines.

3.4 Data Pre-processing

After the data collection, the text data is cleaned and cleansed. This phase of the methodology is the most important phase; in order to build a robust, stable and efficient system, the data must be properly cleaned. Each review and tweet was converted to a sentence after the data pre-processing stage for further analysis.

3.5 Aspect and Sentiment Detection

Once the data is cleaned and the reviews and tweets were converted to cleaned sentences, the sentiment was extracted from each sentence and a polarity was assigned. Similarly, Aspects were detected for each sentence and a separate column was made for each aspect with its polarity score. Finally, the aspects and sentiment of the sentences were stored in a tabular form for each airline.

3.6 Data Mining

Statistical operations like correlation and linear regressions were performed on the aspects and polarity score of the reviews for the airlines. Hence, these statistical operations provide a better insight into the data and help by understanding the relationships among different factors in the dataset.

3.7 Interpretation, Evaluations and Visualization

The evaluation of the research is based on the Precision, recall and F-score values. Further, various machine learning algorithms were performed in order to find the best-fit algorithms for the system.

Last but not the least, the results were visualized using Tableau and a case study was formed in order to answer the proposed research questions.

3.8 Conclusion

Hence, for this research a modified KDD approach is implemented as it fits best for the research. Furthermore, a detail explanation of the steps are discussed in the following implementation section.

4 Implementation

4.1 Architecture Design

The proposed research follows the architecture design shown in Figure 4. It highlights the tools and technologies that are used in the research. The proposed design has a good inter connectivity with the stages which corresponds with the modified KDD methodology. First part of the design is the Data Persistent layer which deals with the data sources, Data processing, and Databases. This layer is connected to the Business project requirement layer which provides the solution to the Business problems and the final stage, Client presentation layer, where the client and data analyst interact and the reports and dashboards are presented.

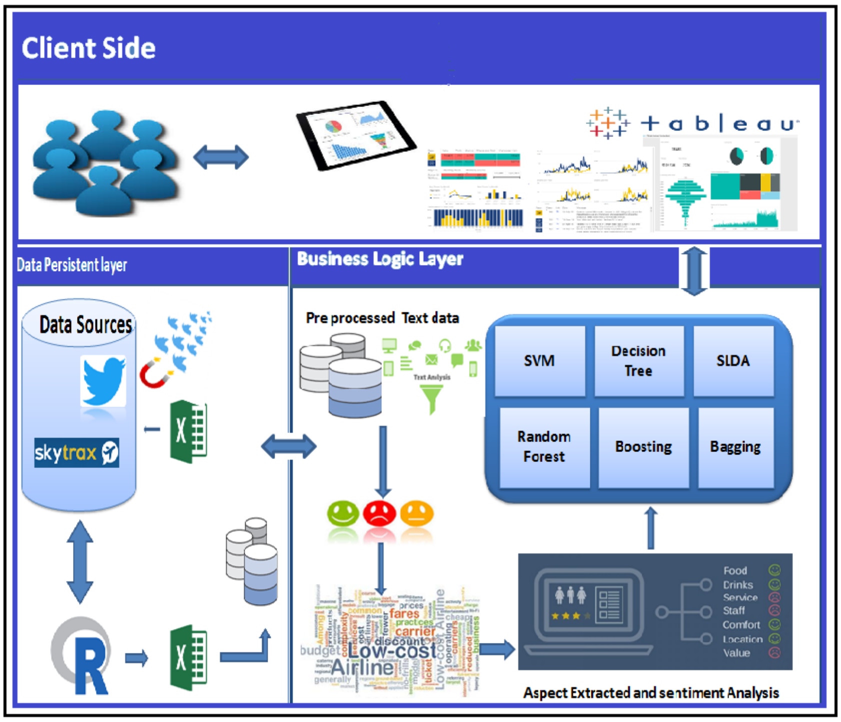


Figure 4: Architecture Diagram

4.2 Process flow

The process flow diagram for the research is shown in Figure 5. The data for the research was taken from two sources, Twitter and online review site Skytrax. For this task, the twitterR package in Rstudio was used (Gentry; 2012) to collect Twitter tweets for the Airlines. API keys and a secret key was generated for authentication access to Twitter. Once the configuration was setup in R, Twitter tweets for airlines were collected for 18 days and a total of 60,445 tweets were collected for the research. The data consists of tweets, airline name, and airline code. Once the data were gathered it was extracted to a comma separated file (CSV) file and stored in the system and external storage device.

For the Skytrax data source, the data was scrapped using Google Chrome Plugin "Data Miner" which extracts data automatically.The tool extracted the airline reviews, airline name, timing, location, reviewer name, and their review ID. The reviews were collected from 2014 to November 2017 and each airline review was collected separately and extracted to a CSV file. All the CSV files from Skytrax reviews were consolidated in a single CSV file. In the end, both the data sources from Twitter and Skytrax were consolidated into a single dataset.

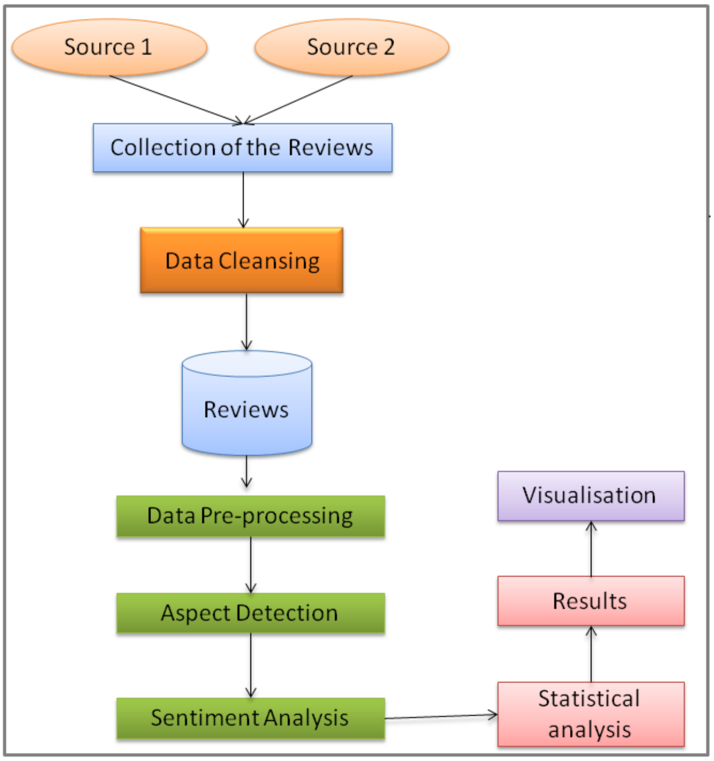


Figure 5: Implementation Process flow Diagram

After arranging the consolidated data-set, unnecessary observations like author names, locations, and review Ids were removed in order to maintain ethical privacy issues. Further, the data was moved to a data frame and was converted into a corpus in order to perform various text processing methods which includes removing the punctuation, converting the text into lower case, removing white-spaces, stemming and removing empty rows. Also, additional customized stop-words were added in order to remove the stop-words using 'tm' package in R. Figure 6 illustrates the data pre-processing.

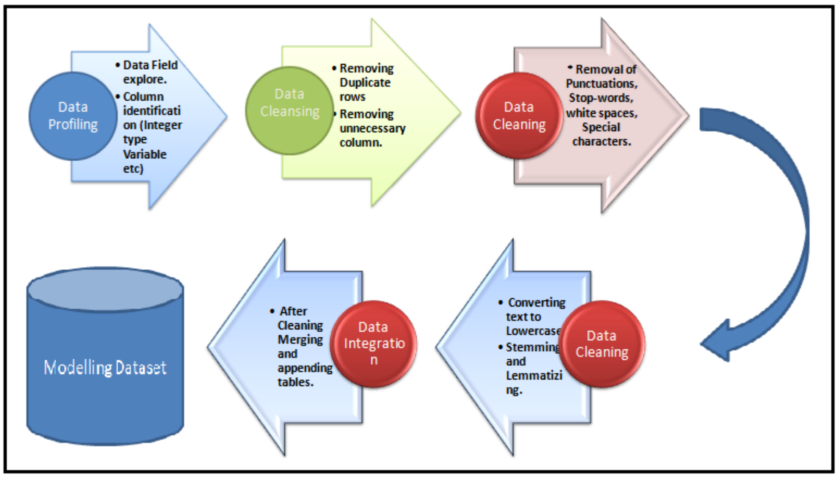

Data Profiling
• Data Field explore.
• Column identificati on (Integer type Variable etc)
Data Cleansing
• Removing Duplicate rows
• Removing unnecessary column.
Data Cleaning
• Removal of Punctuations, Stop-words, white spaces, Special characters.
Data Cleaning
• Converting text to Lowercase
• Stemming and Lemmatizi ng.
Data Integratio n
• After Cleaning Merging and appending tables.
Modelling Dataset

Figure 6: Data Pre-Processing

The gathered data was then interpreted as 1-gram, 2-gram and 3-gram manually by inspecting their frequency using Wordle, then passed to the part of speech (POS) tagging part using the openNLP package in R. Where each word in the sentence is tokenized and tokenized words were split into new sentences. POS tags were extracted as it provides noun, pronoun, adjective etc. for each word.

As there was no proper dataset available for training data, the labelling for the categories were done manually which took quite a long time. After which, reviews were further categorized into five aspects (Food and Beverages services, Staff services, Luggage, Punctuality and Seat Aspects). A bag of words were added for each aspect after analysis 1-gram and 2-gram. The polarity of each sentence was identified and a score was assigned as negative,neutral and positive as -1,0 and 1 respectively through the polarity function of the QDAP package in R. Similarly, words were checked in the sentence and if found in the Aspect lists that category was then flagged with 0 or 1.

Machine learning algorithms like support vector Machine (SVM) (Dimitriadou et al.; 2005), Glment , Maxent , SLDA , boosting (Tuszynski; 2012), Random forest (Liaw et al.; 2002),Neural networks and Decission Tree (Palkar et al.; 2016) were applied and tested using RTexttools package (Collingwood et al.; 2013) in R. It is an open source resource, very flexible and easy to use, which give F score,precision and Recall to make an ease of decision.The training and test data was separated in the ratio of 70:30. As there were grammatical errors, we used unigrams and bi-grams to tackle these (Cavnar et al.; 1994).

Finally, the file is extracted from R and saved to a CSV format in the system. The final dataset consists of the attributes like Cleaned sentence, Polarity score, Food and Beverages aspect score, Staff services score, Luggage score, Punctuality score and Seat Aspects score, Airline names and Airline code.

This data-set is further processed for statistical operations using R statistical packages. The purpose is to find out the correlations between the different aspects and polarity score of the airlines. Also, linear regression is performed in order to find the relationship between the attributes of the airlines.

At last, visualization is performed in the data set using Tableau, since Tableau is very interactive software and very easy to use. It has pick and drop features which are very convenient for the user. Also it supports many types of data sources like .CSV, .xls, .txt, .JSON and databases like MySQL, MongoDB etc.

4.3 conclusion

The results of this section have solved the research question asked in section 1. Furthermore, the results evaluation and interpretations is explained in the next section.

5 Evaluation and Results

According to McLaughlin and Herlocker (2004), algorithms can be measured by utilizing several evaluation methods. The commonly known and simple to utilize them are Precision, Recall, and F score. The Precision is the positive anticipated values i.e. the extent of positive cases that are really positive cases. In another term, when the model anticipated positive class, how frequently is it true? The accurate model will just predict the positive class for the situation prone to be positive (Lantz; 2013).

$$Precision = \frac{TP}{TP + FP}$$

Whereas, Recall is the number of true positive upon the total number of positive (Lantz; 2013). It is like sensitivity. In a corpus, a high percentage of recall implies obtaining a large portion of Positive.

$$Recall = \frac{TP}{TP + FN}$$

F-score is a measure of evaluating the model performance it joins the precision and the recall into a single value. The value of F-score lies from 0 to 1 which illustrates the weakness and strength of the system. If the F-score value is 1 it is considered as the strongest and 0 signifying the weakest.

$$Fscore = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Table 2 describes the evaluation measures for the applied algorithms.

Algorithm	Precision	Recall	F-score
SVM	0.64	0.66	0.59
Decision Tree	0.72	0.68	0.63
Random Forest	0.71	0.66	0.60
Bagging	0.68	0.67	0.65
Boosting	0.68	0.66	0.58
SLDA	0.46	0.38	0.28
Maximum Entropy	0.23	0.33	0.26

Table 1: Comparison between Model's performance

From the above table, it is observed that Random Forest along with decision tree and Bagging performed well with the F-score of 0.60.SVM performed well by scoring 0.59 F-score. The boosting algorithm also showed a strong F-score of 0.65. Although, other algorithms performance like SLDA, Maximum entropy was not up to the mark as the F-score is very low. Hence, It is evident that for the project, random forest and decision tree are the best-fit algorithms. This also answers the research objective of the project.

Practically, every model has its own advantages and disadvantages. When we have infinite dataset with minimum risk, outlier free and perfectly clean data ,then SVM proves to be a powerful approach but as we have neutral comments in the reviews which do not contribute much towards aspects which are treated as outlier by the Model then Random Forest and Decision Tree supports the research.

Both random forest and Decision tree are supervised learning technique follow divide and conquer method for classification for random variables as input.Random forest approaches data in three simple stages:

- Bootstrap reviews as sample.
- Random selection as value for polarity, where Polarity is less than the split variable.
- then breaking into different branches of threshold for decision making.

Further, statistical technique were performed to find the co-relation between the aspects and polarity, which shows Seats and Staff to be highly correlated with Polarity as compare to Luggage, Punctuality and Food. P value is less the 0.05 which implies we do not reject our Null Hypothesis. In addition to co-relation, Linear regression was perform which gives R square less than 0.05. R square value helps to find how tightly our value are bounded towards linear regression line. Low R square Value represent how people have different opinion in different aspects.

5.1 Experiment and Case Study Results 1

D3 js being an open framework for visualization is used to represent an Alluvial diagram which shows relation between three layer of our research as Airlines,Aspects and Sentiments. Every airline is defined by a color and width of the flow defines how strongly it is related to the aspects. Further, Flow from aspects towards sentiments represents sentiments associated with each aspects for each airline by the color and width (Figure 7).

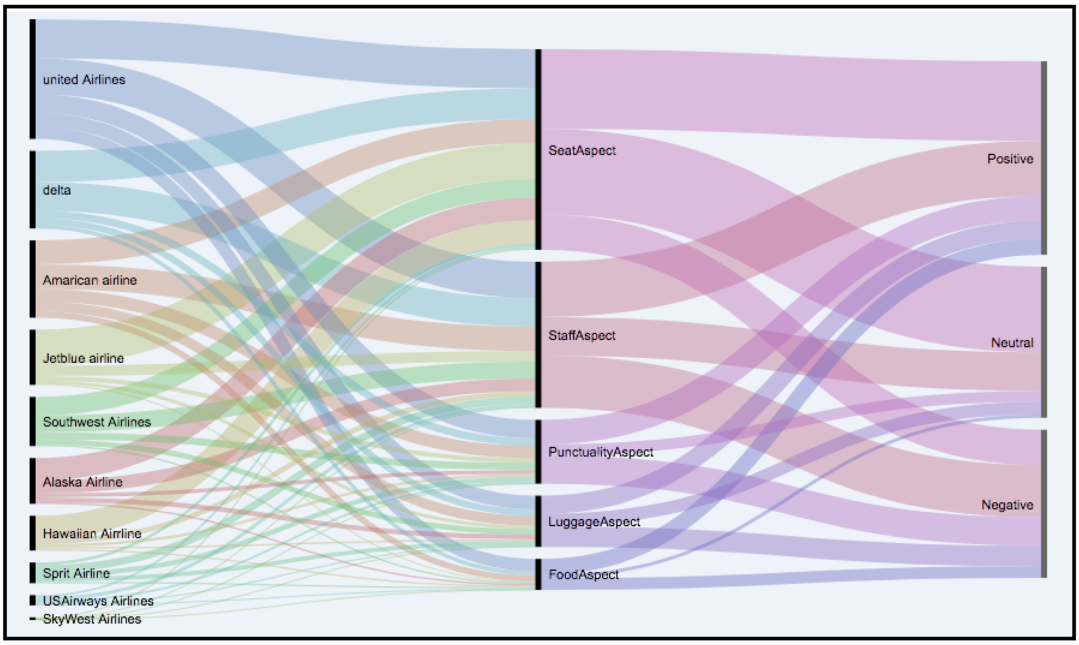


Figure 7: Tree-map graph

Tableau is used for implementing second detailed part of case study 1. The first case study derives the ranking of Airline, based on the polarity score and sentiments. When reviews were operated in R, we get polarity score based on the positive and negative words or phrase in the sentence from -1 to 1 as a continuous measure. 0(zero) is treated as the Median of the Polarity score, Positive sentiments are varied above 0 whereas Negative have a variance below 0.

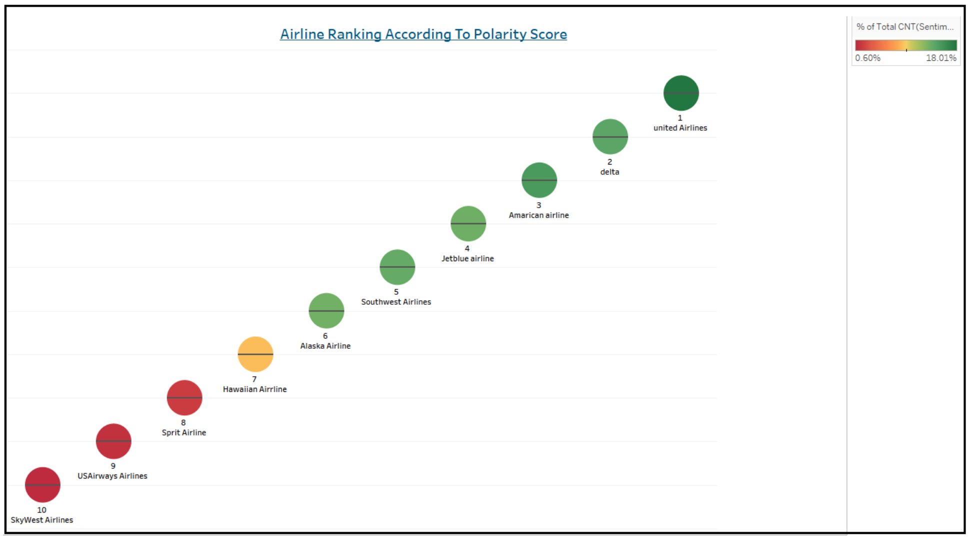


Figure 8: Tree-map graph

Colors in the case study represents the intensity of sentiments across the rank of the Airlines. United Airlines comes out to be 1st rank holder with dark green color showing high positive response whereas Skywest Airline bags 10th rank with dark red shade with highest negative sentiments (Figure 8)

Next part of the case study is a doughnut projecting division of sentiments i.e, positive, negative and neutral for an airline which can be selected from the drop down above. For example, On selecting United airlines and spirit airlines, it is evident that spirit airline suffers with more negative sentiments than united airlines as they have 57.61% and 31.4% negative sentiments respectively (Figure 9).

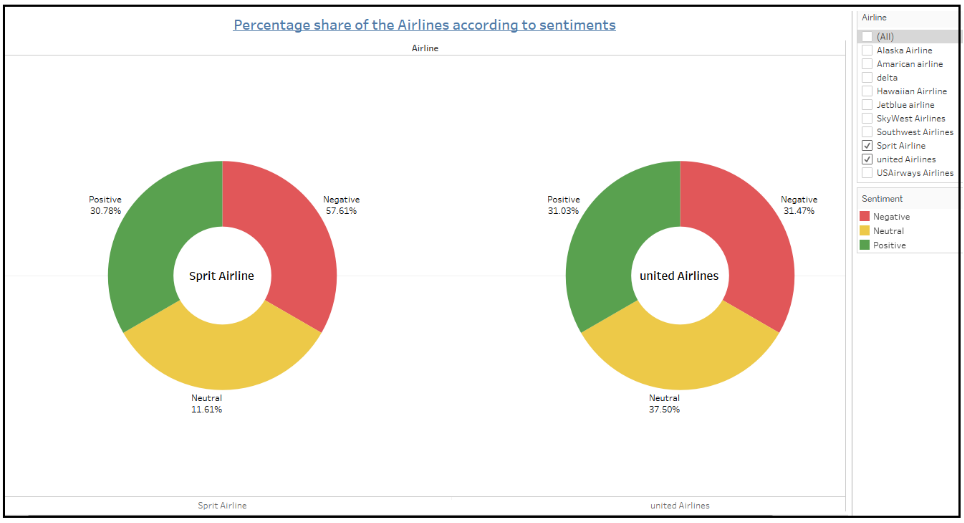


Figure 9: Tree-map graph

5.2 Experiment and Case Study Results 2

The second case study is an analysis of aspects contribution towards the reviewer sentiments. From the uni-gram and word cloud implementation as a word cloud, we found Seats, Staff, Punctuality,Luggage and Food are the major areas of customers sentiments.

The case study projects Seats Aspect as the most important aspect represented by the largest size of the pie, On the other hand food is considered to be lowest contributor in the aspects towards customers sentiments. Also, colors in the pie shows sentiments with percentage share in each aspects as a total of 100%. From the plot we can easily notice, Seat Aspect receiving 41.16% of positive sentiments whereas Staff Aspects receives 35.32% of negative aspects which conclude its important to maintain good factors but it is more important to take negative factors into more consideration (Figure 10).

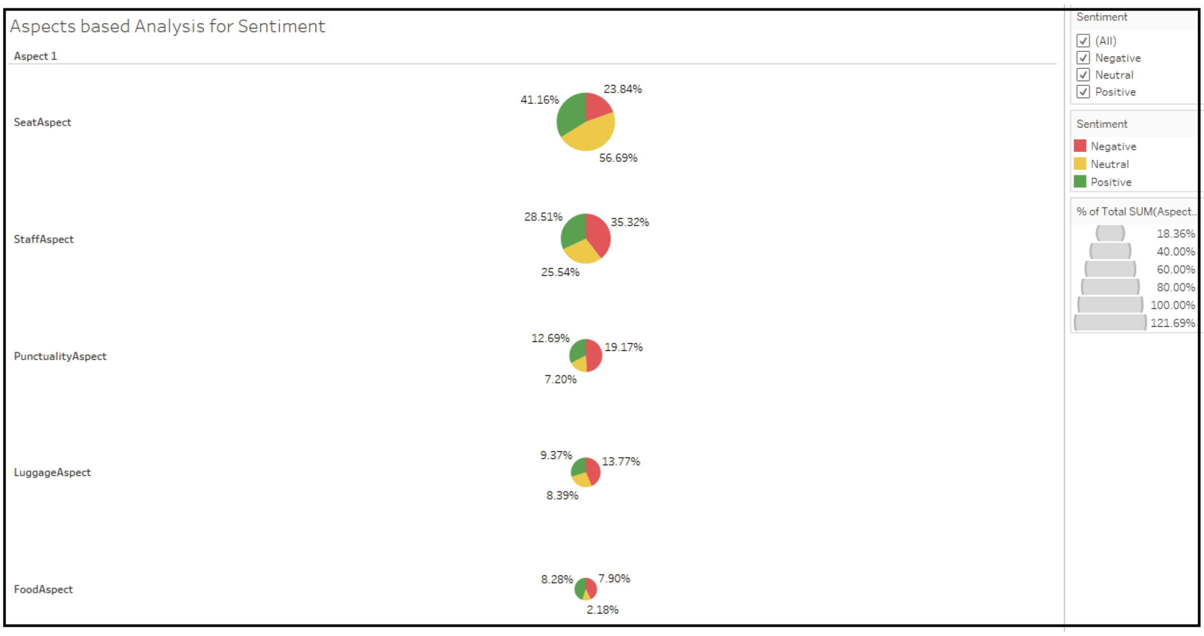


Figure 10: Tree-map graph

Next part of the case study is a bar graph representing the ranking of the airlines according to each aspect holding a percentage share of total 100%.Being one of the most popular airlines United bags first rank in every aspects but there is a competitive edge between flights bagging mediocre ranking. According to previous study Alaska Airline had a better position than Hawaiian Airline, but surprisingly Hawaiian Airline is at a better position with 15.96% than Alaska. But Alaska provide better services in terms of Luggage aspects with 11.89% whereas Hawaiin Airline has 4.20% in our research (Figure 11).

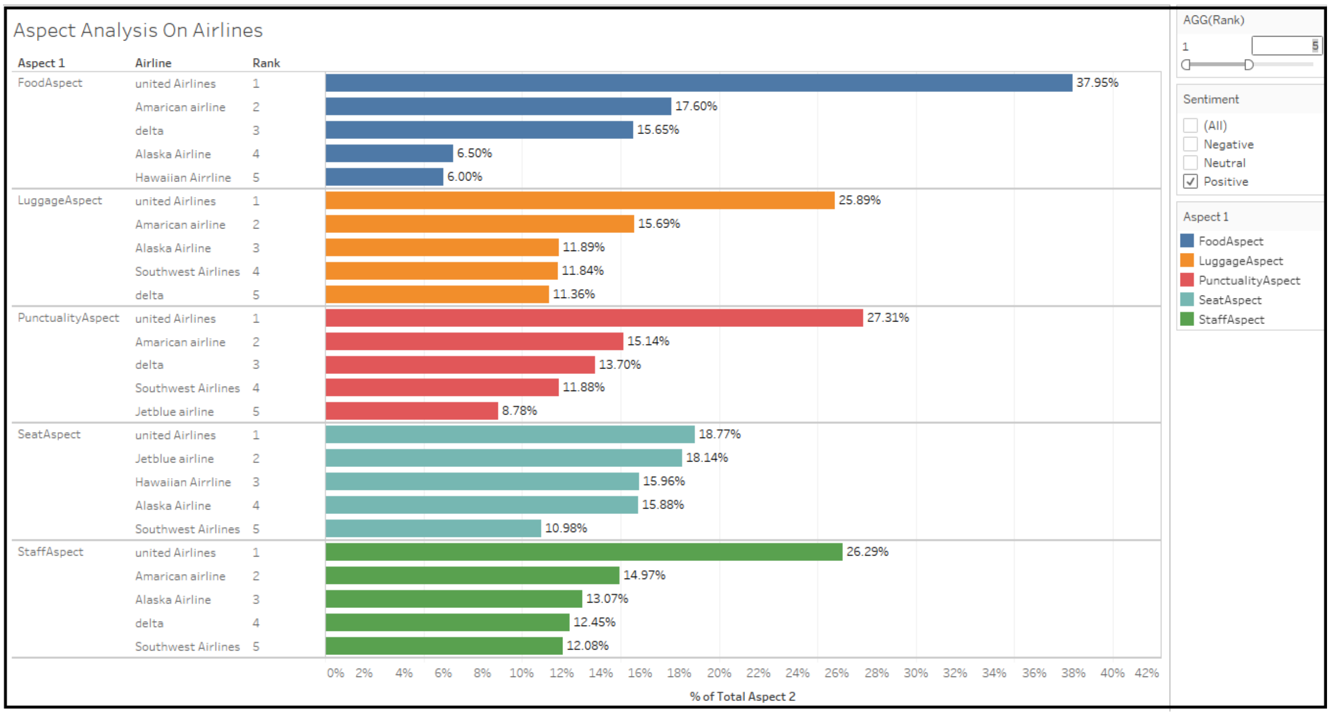
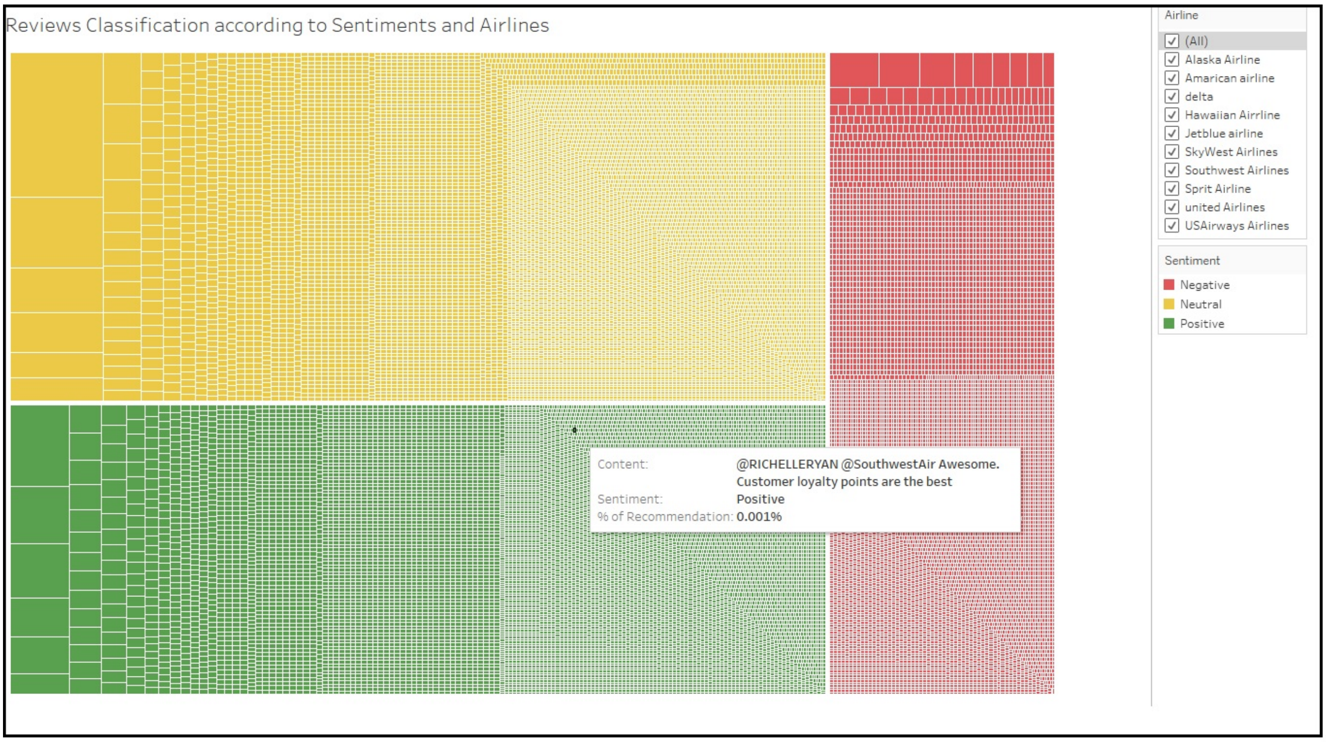


Figure 11: Tree-map graph

5.3 Experiment and Case Study Results 3

In our third case study, we represent a tree-map graph showing all the reviews on hovering cursor to the boxes. different color shows different mood of the reviews whereas the polarity score is define by the size of the box, the bigger the box more intense the reaction is. Figure shows the use case of Southwest Airline showing the comment and positive sentiment in the tool box (Figure 12).



Content: @RICHELLERYAN @SouthwestAir Awesome. Customer loyalty points are the best
Sentiment: Positive
% of Recommendation: 0.001%

Figure 12: Tree-map graph

5.4 Discussion

The proposed results are more efficient when we compare it with the work of the Sankaranarayanan and Rathod (2017). Also prior research in airline field is generally based on finding the leading and lagging perimeters or to find the airline rating. This proposed research is different from the other researches in the past it not only have heterogeneous data sources but also perform a supervised machine learning algorithms. There is significant minimal research in Aspect detection in airline industry. Thus, this research bridges the gap in the research.

6 Conclusion and Future Work

An Aspect-based sentiment analysis has been implemented in the research by using the openNLP library. Various supervised machine learning models were tested. However, Random forest gives the best F-score of over 60 percent.

This project is divided into two main segments. Firstly the main categories of the reviews were identified. Secondly, sentiment analysis is performed for the categories/aspect detected from the reviews.

To the best of candidate knowledge, In the past, research on Airline industry especially by using aspect-based sentiment analysis is minimal. This research will bridge the gaps and will provide a significant contribution to the customers in deciding their airlines. Also, it will help the US based airlines to look after the areas of improvement and they can compare their performances with their competitors in order to get a better competitive edge in the market.

The research questions in section 1.2 and the project objective have been achieved. Also, a significant amount of contribution to the body of knowledge is evident in the technical report.

Besides, there are shortcomings in this research. The model used for this research is only applied to the English language. Since every other language will have a different grammatical structure. Therefore, this model will not work for other languages. Furthermore, in order to add new features or to change the existing once it will require users input.

Future Work: This research has a lot of scope for future improvement. Since this project utilizes a supervised machine learning technique. Thus, in future unsupervised machine learning techniques can be implemented. Moreover, This project can further be implemented in other regional languages and on different fields as well for example education, automobiles etc.

7 Acknowledgement

I would like to thank my Supervisor (Dr. Catherine Mulwa) for guiding and supporting me in my thesis. I would like to thank my family for their love and support in my whole life. I would also like to acknowledge my dearest friend Priyadarshini Tiwari who supported me and was there for me during my hard times.

References

Anitsal, M., Anitsal, I. and Anitsal, S. (2017). A sentiment analysis of air passengers of top ten us based airlines, *Atlantic Marketing Association Proceedings,* pp. 37–50.

Azevedo, A. I. R. L. and Santos, M. F. (2008). Kdd, semma and crisp-dm: a parallel overview, *IADIS European Conference Data Mining(IADS-DM) 2008,* pp. 182–185.

Baumgarten, P., Malina, R. and Lange, A. (2014). The impact of hubbing concentration on flight delays within airline networks: An empirical analysis of the us domestic market, *Transportation Research Part E: Logistics and Transportation Review,* **66**: 103–114.

Berry, M. J. A. and Linoff, G. S. (1997). *Data mining techniques: for marketing, sales, and customer support*, John Wiley & Sons, Inc.

Cachia, R., Compañó, R. and Da Costa, O. (2007). Grasping the potential of online social networks for foresight, *Technological Forecasting and Social Change* **74**(8): 1179–1203.

Cavnar, W. B., Trenkle, J. M. et al. (1994). N-gram-based text categorization, *Ann Arbor MI* **48113**(2): 161–175.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000). Crisp-dm 1.0 step-by-step data mining guide.

Chen, H. and Zimbra, D. (2010). Ai and opinion mining, *IEEE Intelligent Systems* **25**(3): 74–80.

Collingwood, L., Jurka, T., Boydstun, A. E., Grossman, E., van Atteveldt, W. et al. (2013). Rtexttools: A supervised learning package for text classification.

Collomb, A., Costea, C., Joyeux, D., Hasan, O. and Brunie, L. (2014). A study and comparison of sentiment analysis methods for reputation evaluation, *Rapport de recherche RR-LIRIS-2014-002* .

Dharmavaram Sreenivasan, N., Sian Lee, C. and Hoe-Lian Goh, D. (2012). Tweeting the friendly skies: Investigating information exchange among twitter users about airlines, *Program* **46**(1): 21–42.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D. and Weingessel, A. (2005). Misc functions of the department of statistics (e1071), tu wien, *R package version* pp. 1–5.

Dobruszkes, F. (2006). An analysis of european low-cost airlines and their networks, *Journal of Transport Geography* **14**(4): 249–264.

Gentry, J. (2012). twitter: R based twitter client, *R package version 0.99* **19**.

Hannigan, T., Hamilton III, R. D. and Mudambi, R. (2015). Competition and competitiveness in the us airline industry, *Competitiveness Review* **25**(2): 134–155.

Ivanscenko, A. (2016). Topic and sentiment analysis of customers reviews via application of text mining.

Jansen, B. J., Zhang, M., Sobel, K. and Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth, *Journal of the Association for Information Science and Technology* **60**(11): 2169–2188.

Jindal, N. and Liu, B. (2006). Identifying comparative sentences in text documents, *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 244–251.

Kaur, A. and Duhan, N. (2015). A survey on sentiment analysis and opinion mining, *International Journal of Innovations & Advancement in Computer Science* **4**: 107–116.

Lantz, B. (2013). *Machine learning with R:Learn how to use R to apply powerful machine learning methods and gain an insight into real-wold applications*, Birmingham: Packt Publishing Ltd.

Li, G. (2017). *Application of sentiment analysis: assessing the reliability and validity of the global airlines rating program*, B.S. thesis, University of Twente.

Liaw, A., Wiener, M. et al. (2002). Classification and regression by randomforest, *R news* **2**(3): 18–22.

Mangold, W. G. and Faulds, D. J. (2009). Social media: The new hybrid element of the promotion mix, *Business horizons* **52**(4): 357–365.

McLaughlin, M. R. and Herlocker, J. L. (2004). A collaborative filtering algorithm and evaluation metric that accurately model the user experience, *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 329–336.

Min, H. and Min, H. (2015). Benchmarking the service quality of airlines in the united states: an exploratory analysis, *Benchmarking: An International Journal* **22**(5): 734–751.

Miner, G. (2012). *Practical text mining and statistical analysis for non-structured text data applications*, Oxford:Academic Press.

O'Connell, J. F. and Williams, G. (2005). Passengers perceptions of low cost airlines and full service carriers: A case study involving ryanair, aer lingus, air asia and malaysia airlines, *Journal of Air Transport Management* **11**(4): 259–272.

Palkar, R. K., Gala, K. D., Shah, M. M. and Shah, J. N. (2016). Comparative evaluation of supervised learning algorithms for sentiment analysis of movie reviews, *International Journal of Computer Applications* **142**(1).

Pérezgonzález, J. D. and Gilbey, A. (2011). Predicting skytrax's official world airline star ratings from customer reviews, *Aviation Education and Research Proceedings 2011* pp. 48–50.

Safko, L. (2010). *The social media bible: tactics, tools, and strategies for business success*, 2nd ed. Hoboken, New jersey: John Wiley & Sons.

Saha, G. C. and Theingi (2009). Service quality, satisfaction, and behavioural intentions: A study of low-cost airline carriers in thailand, *Managing Service Quality: An International Journal* **19**(3): 350–372.

Sankaranarayanan, H. B. and Rathod, V. (2017). A novel approach for predicting ancillaries ratings of indian low-cost airlines using clustering techniques, *International Conference on Information and Communication Technology for Intelligent Systems*, Springer, pp. 199–209.

Schouten, K. and Frasincar, F. (2016). Survey on aspect-level sentiment analysis, *IEEE Transactions on Knowledge and Data Engineering* **28**(3): 813–830.

Sultan, F. and Simpson Jr, M. C. (2000). International service variants: airline passenger expectations and perceptions of service quality, *Journal of services marketing* **14**(3): 188–216.

Tuszynski, J. (2012). catools: Tools: moving window statistics, gif, base64, roc auc, etc. r package version 116, *URL http://CRAN. R-project. org/package= caTools.[accessed 01 April 2014]* .

Wan, Y. and Gao, Q. (2015). An ensemble sentiment classification system of twitter data for airline services analysis, *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*, IEEE, pp. 1318–1325.

Wu, W.-Y. and Liao, Y.-K. (2014). A balanced scorecard envelopment approach to assess airlines' performance, *Industrial Management & Data Systems* **114**(1): 123–143.

Yakut, I., Turkoglu, T. and Yakut, F. (2015). Understanding customers' evaluations through mining airline reviews, *International Journal of Data Mining Knowledge Management Process (IJDKP).* pp. 5(6): 1–11.

Yee Liau, B. and Pei Tan, P. (2014a). Gaining customer knowledge in low cost airlines through text mining, *Industrial Management & Data Systems* **114**(9): 1344–1359.

Yee Liau, B. and Pei Tan, P. (2014b). Gaining customer knowledge in low cost airlines through text mining, *Industrial Management & Data Systems* **114**(9): 1344–1359.