National College of Ireland

Aspect Based Sentiment Analysis Using Data Mining Techniques Within Irish Airline Industry

MSc Research Project
Data Analytics

Aishwarya Mundalik
x15041450

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:	Aishwarya Mundalik
Student ID:	x15041450
Programme:	Data Analytics
Year:	2016
Module:	MSc Research Project
Lecturer:	Dr. Catherine Mulwa
Submission Due Date:	23/12/201
Project Title:	Aspect Based Sentiment Analysis Using Data Mining Techniques Within Irish Airline Industry
Word Count:	

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:	23rd April 2018

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 Using Data Mining Techniques Within Irish Airline Industry

Aishwarya Mundalik
x15041450
MSc Research Project in Data Analytics

23rd April 2018

Abstract

Irish airline transport has created several distinct and significant types of economic benefit and income. Principal benefits have been created for clients; the passengers land the shipper, who uses air transport services. In addition, the airline industry has created connections between cities and markets globally that represent an important infrastructure asset which has lead to generating benefits such as foreign direct investment, business clusters, increased specialization and increase in the countrys economy productivity capacity. This project applies the aspect-based sentimental analysis approach and techniques to investigate whether these techniques can provide meaningful insights to enhance the performance of the Irish industry. In addition by analyzing twits and reviews, the project tackles the question of to what extend can analyzing of sentiments with respect to aspects become beneficial from customers perspective for them to be able to choose better airline service. The results of analyzed literature review are also presented. Based on identified gaps, several aspect based sentiment models are developed.

1 Introduction

Irish airline industry is one of the major contributor in Irelands economy and provides almost five thousand jobs. According to sources, aviation sector in Ireland responsible to contribute more than four billion euros to Irish GDP (Gross Domestic Product). For such a big industry it become significant to know their customers and provide better service. Especially, it is necessary to know where they are performing wrong. By customers viewpoint it is also important to choose more better and comfortable airline for their journey as there are many choices exists. In the era of sentiment analysis, aspect-based sentiment analysis has been booming from past few ages in industrial as well as academic area. Aspect based sentiment analysis also known as feature-based sentiment analysis. There are some demanding applications in this fields for instance, Social media analysis and online reviews analysis. Customer decision can be easily affect by analyzing their reviews online Sidorov et al. (2012). With the good result this method can become one of the great business strategy. In this research project two data sets are being used.First data set is consist of online reviews of Irish airline customers. Reviews are fetched from airlinequality.com where each review is from authorized customer.For the second data

set social media micro blogging website Twitter is used. Twitter is one of the precious online source for reviews Khan et al. (2015).Every second approximately six thousands of tweets are posted by Twitter users worldwide.Both the sources gives authenticate and valuable reviews to analyze.The goal of this research project is to consolidate all the online available reviews about Irish airlines and analyze these reviews by applying aspect-based sentiment analysis method

1.1 Motivation and Background

Airline transport network industry has been flourishing in past few years and more than 30 million airplane take-off every day. In this technological era, social media addiction is booming rapidly. People post reviews about amenities they get in everyday life, and that becomes helpful for others. More often people look the reviews about product or services before choosing them. These reviews can be analysed using sentiment analysis method, but aspect-based sentiment analysis will give deeper and exact result about that review. For instance, if someone gave review about Ryanair "cheap rates, decent food but with staff service sucks" then typically this review will go under the negative segment. But here, there are three different reviews which are about cost (positive), food(neutral) and staff service (negative). Aspect based sentiment analysis used to mine text with respect to definite object of aspects Pontiki et al. (2016). However, not much research has been done in airline domain in sentiment analysis field. In Adeborna and Siau (2014) author identifies the polarity of reviews with the help of sentiment analysis of the classified text. Wan and Gao (2015) tried several sentiment analysis classification approaches and then analyzed each of them to find best approach. Previously not much research has been done in aspect-based sentiment analysis for airline domain. The main goal is to collect online reviews from different websites and analyze these reviews then perform not only sentiment analysis but also aspect-based sentiment analysis to analyze every review wisely. By creating model choose best fitted model depend on accuracy, precious and recall. At the end task is visualization of the results along with distinct aspects, it will help airline to know where they are performing wrong so that they will improve their services. It will also help customers to choose more beneficial airline services for their healthier journey.

1.2 Project Requirement Specifications

This research project specification phase emphasize on to developed aspect based sentiment analysis model for Irish airlines using various data mining techniques. This research will contribute to the Irish airline industry to know their customers as well as to customers to choose the best airline according their needs. The research question and research objectives are as follows.

1.2.1 Research Question

RQ: " To what extent can aspect based sentiment analysis of Irish airlines (AerLingus, Cityjet, Norwegian, Ryanair) conducted using supervised machine learning techniques(Gradient boosting, KNN,Naive Bayes,SVM,Linear Regression) be used to deliver meaningful insights to enhance the performance of the Irish airline industry?"

Sub-RQ: Can analysis of sentiments(tweets and reviews) within the four Irish airlines be used to support customers when deciding which airline service to choose?

1.3 Research Project Objectives

Following Objectives are specified as a solution to aforementioned research question: **Objective A** was to gather online reviews from different online sources through several tools and programming techniques and preprossessing of these reviews. After getting desire data **Objective C** was to calculate sentiment score for reviews using natural language processing methods. **Objective C** to find aspects from the available reviews (AerLingus,Cityjet,Norwegian,Ryanair) using NLP methods. After finding the aspects **Objective D** was to evaluate, deploy and inspect outcome of prediction model using supervised machine learning techniques. Following objectives are further divided into sub-objectives as shown in Table.2.

Objective DI	Implementation, evaluation and results of gradient boosting
Objective DII	Implementation, evaluation and results of k-nearest neighbour
Objective DIII	Implementation, evaluation and results of linear regression
Objective DIV	Implementation, evaluation and results of naive bayes
Objective DV	Implementation, evaluation and results of support vector machine

Table 1: Reserach Sub-Objectives

Hypothesis: Consider, hypothesis represented as H and X is presented as airline names where, A is input and B is output here and, S ¡= Sentiments. H ¡= Aspects Then, Ho: A gives B where $B¡=X,S,A$

1.4 Research Project Contribution

Fundamental contribution of this research project is entirely developed aspect based sentiment analysis model for Irish airlines using supervised machine learning models as shown in Table.2

Developed model using gradient boosting algorithm with results
Developed model using k-nearest neighbour algorithm with results
Developed model using linear regression algorithm with results
Developed model using naive bayes algorithm with results
Developed model using support vector machine algorithm with results

Table 2: Research contributions

Furthermore, data gathering and cleaning using different technical methods has been done. Aspect extraction using various NLP techniques has been conducted. Finally, results are presented using different measures like, accuracy, precision and recall.

1.5 Conclusion

The rest of the technical report as framed as follows: Chapter 2 represent related work of aspect based sentiment analysis within Irish airline industry from 2004 to 2017. Furthermore chapter 3 introduced the modified CRISP-DM methodology approach. Chapter 4 represents the implementation, evaluation and results of five supervised machine learning algorithms. Finally chapter 5 represents the final conclusion of conducted results and recommend future work.

2 Literature Review of Aspect Based Sentiment Analysis within Irish Airline Industry (2004-2017)

2.1 Introduction

In past three decades Irish airline industry has changed enormously. Airline industry is one of the biggest industry which has major influence on Irelands economy. Since now not much research has been done in airline domain with sentiment analysis Wan and Gao (2015). According to Misopoulos et al. (2014) from last 10 years, use of social media and web technologies mounting rapidly. Therefore, consequently people tends to post their reviews on internet in day to day life. As discussed by Wan and Gao (2015), analysing customers reaction through their feedbacks is important for airline companies. These reviews will help them to gain business knowledge by using aspect-based sentiment analysis and machine learning techniques. These methods help them to provide better services to their customers. Aspect level analysis focuses on each feature of service or review not only on view of people opinion Wan and Gao (2015). Several works have been done on sentiment categorisation methods and primarily focuses on three classification group. Groups that are involved are positive neutral and negative Adeborna and Siau (2014). In the next section candidate tried to discuss all the associated topics which are appropriate to this literature and paper.

2.2 Concept of Aspect Based Sentiment Analysis and its working

Sentiment analysis also called as opinion mining. It is a technique to identify writers attitude towards product or service. Collomb et al. (2014) says that, sentiment analysis is another method of text analysis and text analysis is used to demonstrate opinions or thoughts of the customers. However, another explanation described by Singh et al. (2013) that, sentiment analysis is the process which is done by computational method to distinguish the opinions of customers and classify them into positive, neutral and negative. As stated by Medhat et al. (2014) process of mining the sentiments chiefly involves three steps. These steps are identification of sentiments, feature or aspect extraction and classification of sentiments respectively. In Abirami and Gayathri (2017) they have discussed that, sentiment analysis is the approach which is primarily distributed in three types of analysis e.g. document level, sentence level and feature level classification. In document level classification purpose is to determine an entire document to identify the polarity of sentiment. Whereas, in sentence level and feature level classification the main goal is to determine sentiment polarity of each sentence or each word as every sentence and word has its own meaning. Much research has been done on aspect-based sentiment analysis in recent years as this analysis not only focuses on sentence or review but each feature of that sentence Collomb et al. (2014).

2.3 Impact of Sentiment Mining Practices

Since revolution of sentiment analysis several methods have been used to determine sentiments Pang et al. (2008). These methods are mainly categorized into two types e.g. by using lexicon-based approach and machine learning approach. As stated by Sidorov

et al. (2012), sentiment mining is a task done by using machine learning to classify sentiments and this task also known as classification. Polanyi and Zaenen (2006) discussed that, in lexicon-based approach there are generally manually set rules which is also called as lexicons. As stated by Hu and Liu (2004) in their paper, to make use of specific word of opinion they have tried to use lexicon-based method. These targeted words are most occurred words in a document like, bad/poor, Awesome, amazing, strong. By Ding et al. (2008) these words can be obtained by bootstrapping procedure by using WordNet. (English language lexical database). By using wordnet it is possible to count the positive and negative words for product within document. If number of positive words are greater than number of negative words, then document will be classified as a positive sentiment. According to Ding et al. (2008) Lexicon-based approach is one of the easy method to developed as this method has its own disadvantages.

2.4 Comparision of Techniques used in Aspect Based Sentiment Analysis

Supervised machine learning is the approach where primary goal is to use labelled data to classify textual data. By Pannala et al. (2016) machine learning algorithms used trained dataset, further these algorithms studies trained dataset to produced new outcomes with respect to previous trained data. Author also stated that, supervised and unsupervised machine learning techniques now become domain oriented. Sentiment analysis is the task which is one of the classification method of machine learning algorithms Sidorov et al. (2012). In aspect-based sentiment analysis more research has been done with the help of SVM, Naive Bayes, MaxEnt algorithms. Not much work has been done in airline domain with above algorithms by using aspect-based sentiment analysis. Machine learning algorithms plays significant role in sentiment analysis and it is responsible to give outcomes. Results will generate in terms of accuracy, recall and precision

2.4.1 Critiques of Supervised Machine Learning Models

Maximum entropy is one of the known and widely used method in supervised machine learning approach. This method identifies the data with great randomness to achieve high entropy. Therefore, maximum random variable in data means high entropy. Gupte et al. (2014) say that, the main concept behind this model was it should consider most uniform models and it should handle any type of problem. Maximum entropy is the statistical model with information, which is encoded. As stated by El-Halees (2015) this technique uses probability principle and consider probabilities as many as possible. Since now, several approaches carried out using maximum entropy method in aspect-based sentiment analysis to get some accurate results. El-Halees (2015) and Ronglu et al. (2005) both used maximum entropy model for different languages and did text classification and they got an accurate expected outcome. This method is more desirable for text classification because it works well with large datasets.

Nave Bayes algorithm is the combination of algorithm instead of one algorithm and its responsible for aspect selection. Nave Bayes classification algorithm follows Bayes theorem. This algorithm assumes that all classes are already independent which are going to use in classification Gupte et al. (2014). selection of the aspect is independent on value of the aspects. However, this algorithm is commonly used in document level classifications as well as in sentiment classifications. As stated by Leung (2007), Nave

Bayes classifiers predict the probabilities of their belonging class. For instance, this is the given dataset and its belongs to this class. Qiang (2010) have done research with consideration paragraph position and sentence. They have classified the trained dataset by using Nave Bayes algorithm.

Support vector machine is broadly used in supervised machine learning method. Analysing of data one of the major task in SVM. This supervised machine learning algorithm is responsible for both method e.g. regression and classification. According to Pannala et al. (2016) this method is useful to find patterns in data. These patterns can be further used for classification as well as for regression tasks. SVM analyse the borders of decision and follows decision plane principle. Key task which is controlled by SVM is to demonstrate linear separates. Pannala et al. (2016) stated that, SVM is responsible to classify two classes with maximum possible gap or hyper lane. When number of samples are less than number of dimensions SVM works well as discussed by Pang et al. (2008), they have used SVM by using bag-of-unigrams to analyse the sentiments of the movie reviews and ended up having 82.09% accuracy. SVM is one of the well-known and widely used method because of its high accuracy results.

Random forest algorithm also known as random decision forest and mainly used for classification and regression task. It is one of the most used algorithm for sentiment analysis. Gupte et al. (2014) stated that, random forest is working is like ensemble approach because we must generate multiple trees while training of dataset and use these to deploy this method. Gupte et al. (2014) also says that, because of selections of trees are random the correlation between them reduces and it influence to the high prediction power and efficiency. Fang and Zhan (2015) used random forest in product review sentiment analysing because of its excellent performance and result. However, they get their expected accuracy. They further, used random forest and performed aspect-based sentiment analysis and got 75% accuracy and recall. Many researchers preferred random forest with another algorithm to perform ensemble approach.

Gradient Boosting algorithm is one of the most relevant method to enhance the accuracy of machine learning. Fang and Zhan (2015) Says that, boosting reduces the presumptions in supervised machine learning. It is also called as machine meta learning. The primary concept of boosting is gather all weak learners by re-weighting trained dataset and forms ensemble method so that by combining all weak-learner and their weak strength, it will form new high accuracy boost the performance. Therefore, one can say that, this algorithm converts weak-learners into strong-learners. Many researchers used boosting method to enhance the complete accuracy. Xu and Li (2007) proposed boosting algorithm to retrieve the information called as AdaRank where, AdaRank is influence by Adaboost

2.5 Identified Gaps in Aspect Based Sentiment Analysis

As stated by Baharudin et al. (2010) they have worked with Naive Bayes for text classification to check two measures which are accuracy and recall. Moreover, they have used SVM with bag of sentence and bag of word approach to classify polarity. As obtained result showed that both algorithm works well for them for classification of sentence. Bag of word approach worked well than bag of sentence approach and end of the process they got more than 80 % of accuracy as a result. At the other side Blair-Goldensohn et al. (2008) discussed that, initially they have detected raw data and some derivatives as an extra parameter and got rough results. Maximum entropy algorithm used for detecting

aspects of products. Aspects were detected by star ratings of product. When the results are derived by using two or more approaches then this approach ordinarily called as ensemble approach. Ensemble classification method has been used in these days. In aspect-based sentiment analysis it is one of the new research. When one approach is not enough to get an expected result then some researchers use ensemble classification method. here Perikos and Hatzilygeroudis (2017) used ensemble classifiers by using SVM, Naive Bayes and maximum entropy to detect the sentiments of users comments. Consequently, they have found out that, ensemble approach is better than using individual learning algorithm. They stated that this approach could be better for aspect-based sentiment analysis problem. Even though there is not much research has been done in airline domain Wan and Gao (2015) used ensemble approach for sentiment analysis in airline domain.

2.6 Conclusion

From the above related work, to the best of the candidate knowledge it is clear that not enough research has been done in airline industry with aspect based sentiment analysis. Most of the researchers has done research with polarity score or sentiment analysis with product reviews. Most of them focused on airline delay analysis.To get complete insight from customers about airline services two three months is inadequate to collect the data. Large amount of data is needed to get an exact insights from the customers.

3 Scientific Methodology used

3.1 Introduction

This chapter represent the Scientist methodology used in this project. The min approach approach applies the same CRISP-DM which has been used by lot of researchers in this field.

3.2 Modified Methodology Approach Used

Project planning is essential stage for every project.In this research project candidate has adopted modified CRISP-DM(Cross-industry Standard Process for Data Mining) approach for the research planning throughout the research. Following (Fig.1) describes the modified CRISP-DM model for aspect based sentiment analysis within Irish airline industry.

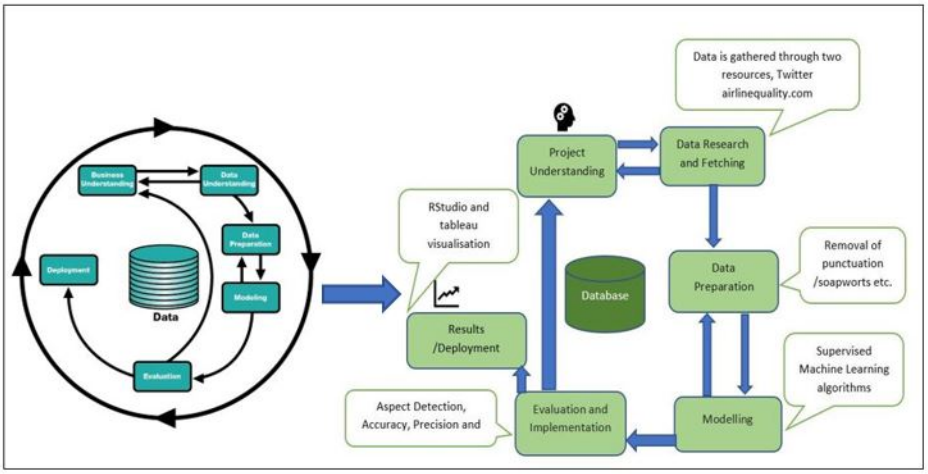

Figure 1: Modified Scientific Methodology used

1. **Project Understanding:** In the research project under-sanding stage addressed project objectives. Project objectives are related to developed a data mining model for aspect based sentiment analysis for Irish airline industry.

2. **Data Research and Data Fetching:** In this project two different data sources are used. One data source is from airlinequality.com and other is from Twitter. Data extracted using data miner tool and python programming language.

3. **Data Preparation:** In this phase of research data cleaning has been done in R studio. Cleaning refers to removal of extra spaces, convert text into lower, remove stop words etc.

4. **Modeling:** In this phase, different supervised machine learning models for classification techniques are used. Overall five models are deployed in research project namely as naive bayes, support vector machine, linear regression, k- nearest neighbour and gradient boosting.

5. **Results ans Deployment:** Aster selection of models for the research project. In this phase models are deployed and gained results. Candidate has checked that, whether this project objectives meets their desire outcomes or not.

6. **Results and Deployment:** In this last phase of research project, candidate has checked the project goals with the results. This phase is the last stage of the research project.

3.3 The Process Applied During Extraction of Tweets and Reviews

In this research project the candidate has used two data sources collected from online source. First data-set consist of online airline customer reviews. Second data-set consist of tweets posted on Twitter about Irish airlines. Following chapter represents the processed methodology used to extracting the tweets and reviews. Following (Fig.2) shows the

process of extracting data from airlinequality.com. Data miner tool were used for the extraction of data. Second data source contains tweets from Twitter. As shown in (Fig.3) initial process is to create Twitter authentication keys followed by script in python to save fetched tweets as a csv file.

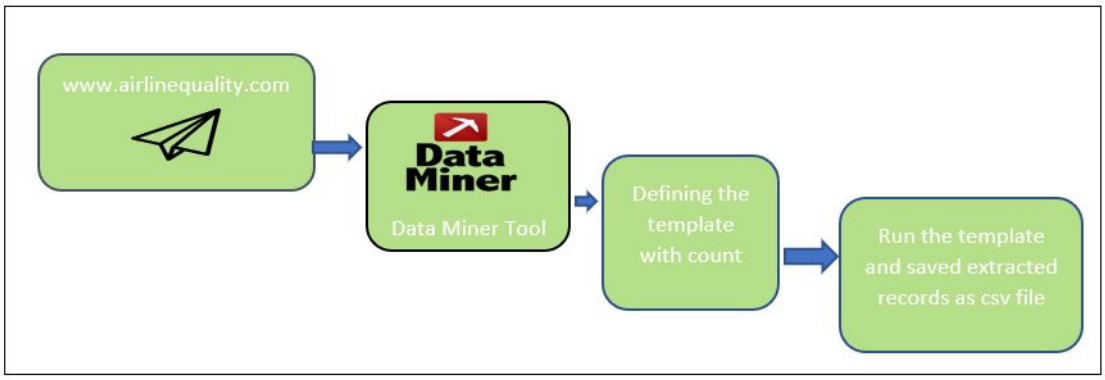

Figure 2: Fetching online airline reviews from airlinequality.com

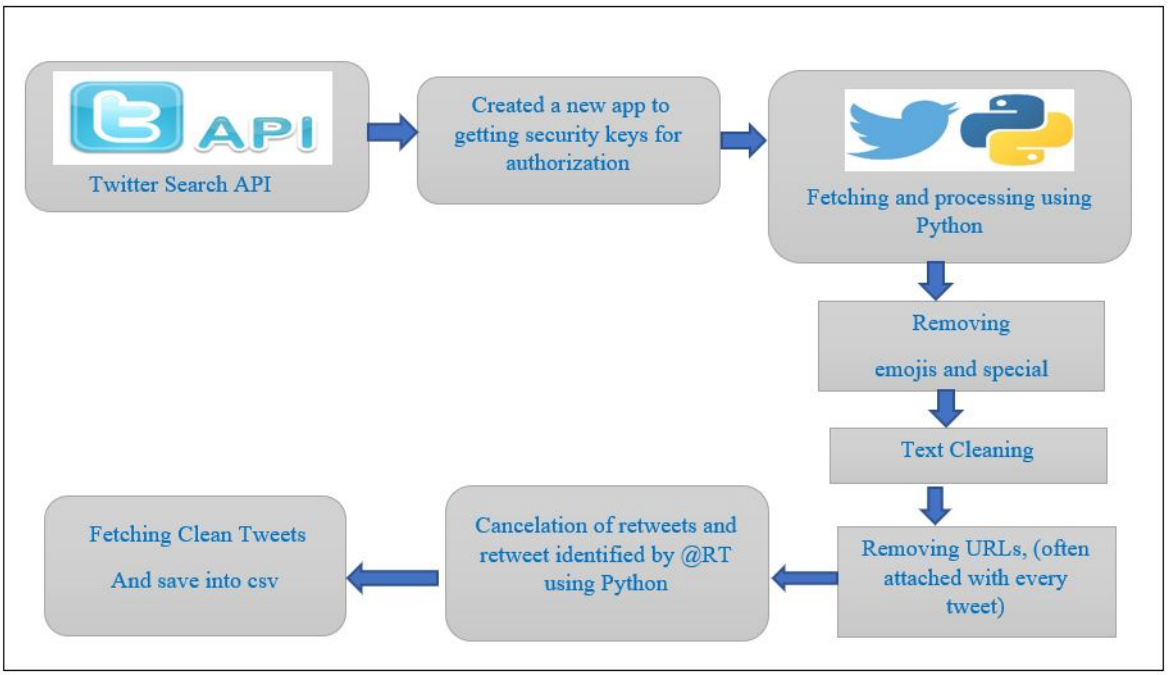

Figure 3: Fetching of tweets about airline from Twitter using python

3.4 Architectural and Technical Design

In this research project of aspect based sentiment analysis and implementation of sentiment models is designed using three tier architecture . Following (Fig.4) shows the entire component of the implementation in the project.This architecture consist of 3 layers.

First tier is the client layer which consist of user interface and visualizing the results to the Irish airline customers. The Second tier is business logic layers consist up of the main sentimental modules different machine learning algorithm and techniques. Finally

third tier is the data persistence that form the back end back. It consist of programming language used that are python and R. The data gathered is mapped in R.

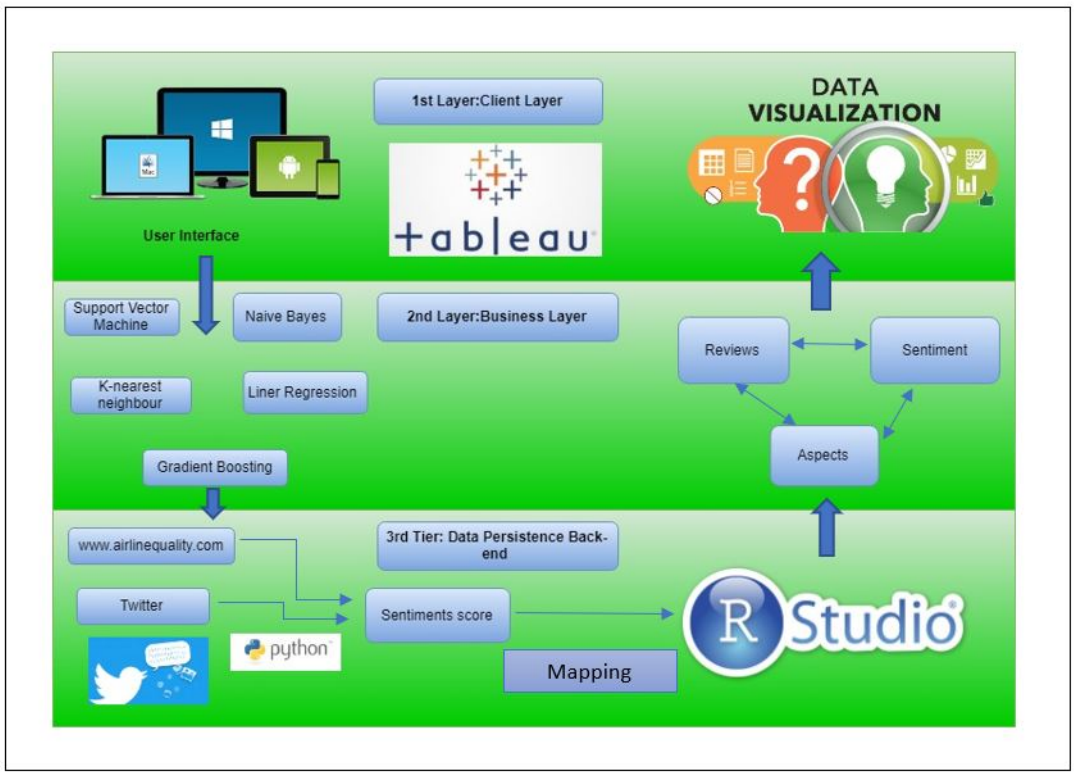

Figure 4: Architecture and Design of 3 tier Architecture

3.5 Conclusion

Candidate concludes that, a scientific methodology has been used to gather online data which is main part of this research. modified CRISP-DM methodology approach is used in research project. This methodology has been adopted for a project planning. Process flow defines the flow of this research from start to stop. Three tier architecture is been used for the research project architecture.

4 Implementation,Evaluation and Result of Aspect Based Sentiment Modeling: Irish Airline Industry

4.1 Introduction

This chapter presents the implementations and evaluation of sentiment models and the aspect based sentiment analysis is also conducted. In order to evaluate the data sets and the developed models several evaluation techniques and matrix are used. The main matrix for this project is accuracy which is very significant and its used when solving the research question.The computation and collection of the data, data pre-processing and aspect extraction is also conducted. This chapter later presents the implementation of machine learning algorithms and finally represents the overall outcomes.

4.2 Aspect Based Sentiment Analysis Process Flow Diagram

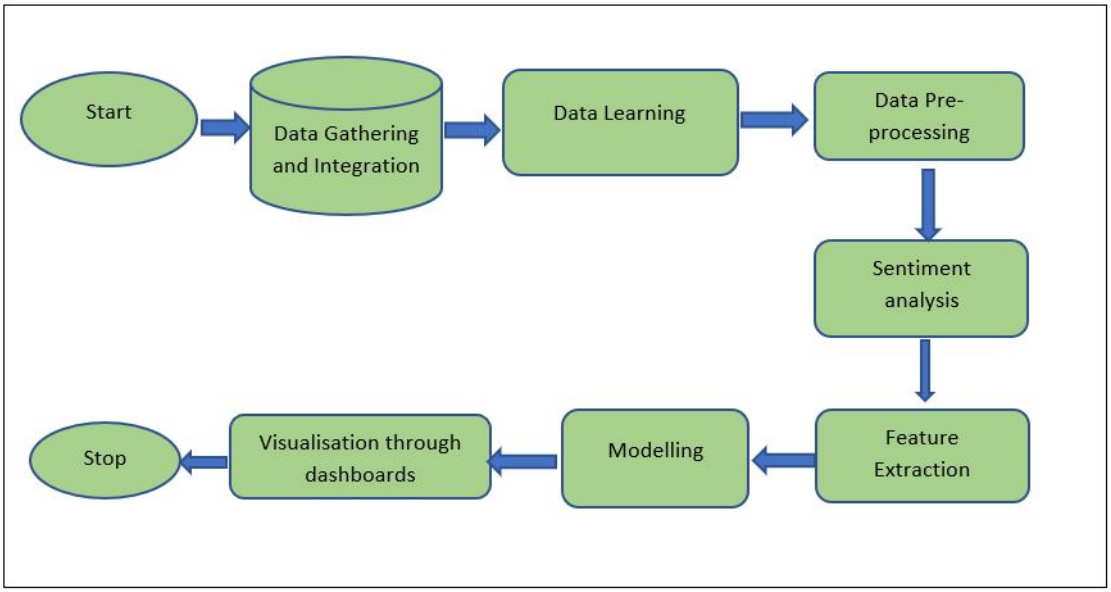

Figure 5: Aspect Based Sentiment Analysis Process Flow Diagram

The process flow diagram for the aspect based sentiment analysis of Irish airline industry is shown in (Fig.5). Process starts from gathering of data from different sources. After data gathering and data understanding data pre-processing is conducted. Pre-processing of data includes the data cleaning steps like removal of numbers, removal of stop words, removal of white spaces etc. Once the data is clean python code is applied on data to fetch polarity / sentiment score from reviews. Feature extraction is done by qdap package in R. After getting final data set with aspects and sentiment score five different supervised machine learning algorithms were tested and and highest accurate model is selected. Finally, visualization is done with tableau an R.

4.3 Data Extraction and Prepossessing

In this research project data gathering and pre-processing are fundamental objectives and therefore, objective A is derived in following section. For research project two data sets are used for conducting the aspect level analysis. First Data set is made up of customers reviews fetched from airlinequality.com. All of these reviews are from authorized customers therefore, there is no burden of spam review analysis here.For the process of reviews fetching data miner tool is used here and reviews scrapped into .csv file. Second data source is from social networking micro blogging website which is Twitter.com. Twitter data is being fetched with python programming. Tweepy package provided by python is used to connect with Twitter API. Tweets are identified with hash tag of the airline names in it. Later, tweets are fetched with the re-tweets eliminations. function clean-text called to get clean tweets and saved tweets in .csv file with clean text.

4.3.1 Sentiment Polarity Analysis

After Collecting the two files from two resources files are consolidate and save as one file. airline-id is manually assigned for each reviews to identify the belonging airline. Text-blob library provided by python is used to assign the polarity to each review. Text-blob library provides the subjectivity and polarity with NLP built in python. After running the python code another file generated with polarity score and polarity in it. For negative polarity flag is set to 0 where for positive it is 1 an therefore research project objective C has been presented and performed in this section.

4.3.2 Data Prepossessing

For cleaning the text and R tool is used. R is the open source free tool which provides vast functionality to the text analytics. In this phase text cleaning is done with the R tm package. After getting data with sentiment score from previous stage data set is converted into corpus to apply cleaning functions. All the text is converted into the lowercase then removal of special symbol, words, English stop words are done.In addition, some extra stop words are removed manually for text data validation. White space removal is performed with tm-map function in R. After getting clean text corpus is saved into data-frame for further process. R provides tokenizer library where each word can splits as a characters and each sentence can splits in words. Later, A word cloud is created in R to get initial insights from data as shown below in (Fig.6)

Figure 6: Word Cloud

4.3.3 Tagging the Positions

In this phase NLP and openNLP libraries are used in R to tag the positions. Tagging positions simply means tagging the English grammar positions of the words in sentence. tagPOS and extractPOS functions are written here to tag the positions. Data need to be convert in string here because both of them accepts string data type.

4.3.4 Extraction and Storage of Chunks :Noun and Adjectives

Chunks are simply means a group of words that tends to found together in English language. Nouns and adjectives are more interested in text-mining as it delivers much meaning than conjunctions and verbs. In this phase nouns and adjectives are extracted from the tagged string. extract-Chunks function are used here to extract the key phrase from the sentence.

4.4 Aspect Based Sentiment Analysis and feature selection

After tagging and extracting the nouns in this stage aspects are define in the array to map with available data and counted frequencies of each aspect. R qdap and tm packages are used here to count the frequencies with defined aspect. Aspects mainly are seat comfort,cabin staff,food,entertainment,ground service and wifi. At the end of this phase micro-soft file getting generated with 6 aspects distribution and counted frequencies with text. in sentiment polarity analysis phase polarity are define for the each review. At the last this Microsoft and aspect polarity file are consolidate by id column to map respective polarity for each review and its aspect.Following are some case studies where aspects are deteced and calculated for Irish airlines namely as Aerlingus, cityjet, Norwegian and Ryanair. As shown in (Fig.7) aspect is distributed and most of the aspects are for Aerlingus and least for Ryanair are detected. People mentioned much about cabin staff in their reviews for every airline except Norwegian. Seat comfort is second most appeared aspect for every airline. (Fig.8) described that polarity score appeared high for Norwegian and Aerlingus that means most of their reviews are positive reviews about aspects. Whereas, Ryanair shows less polarity score which means most the reviews are negative for Ryanair. cityjet remains stable with medium polarity score.

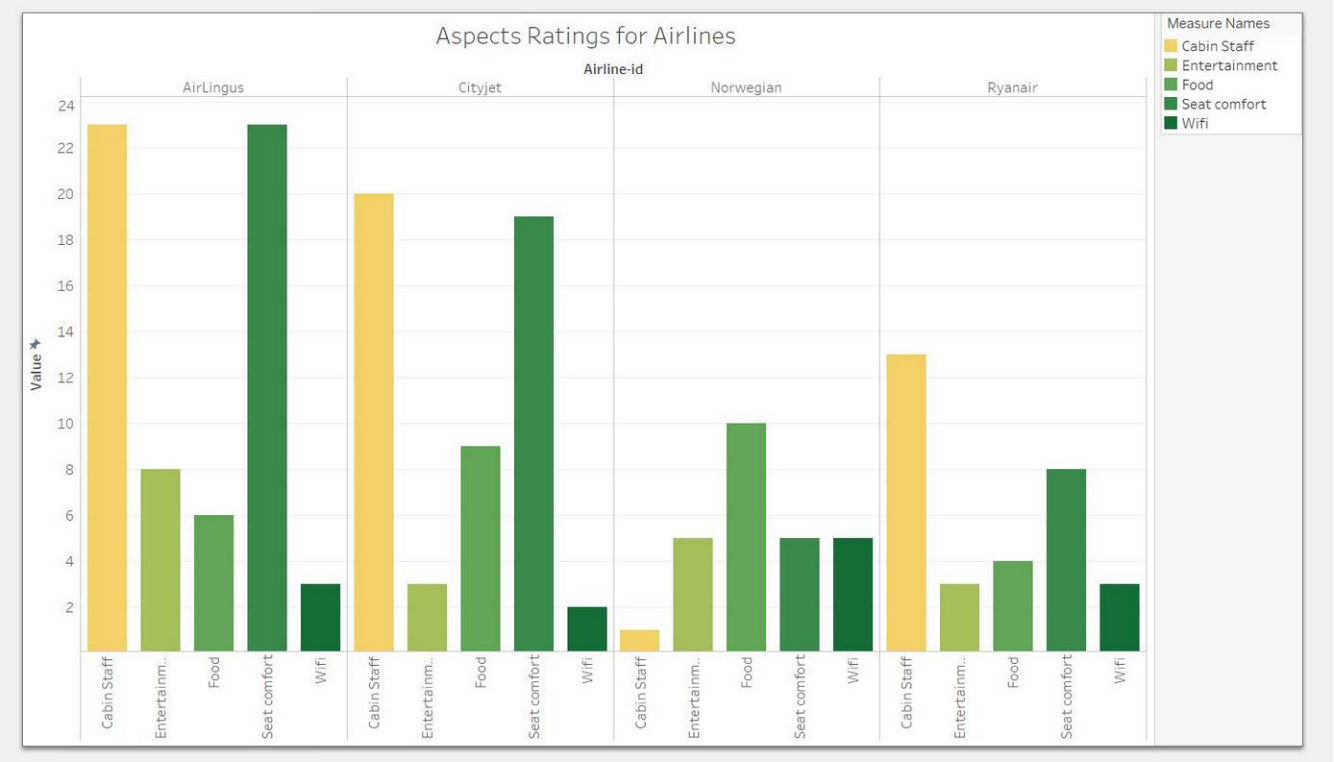

Figure 7: Aspect Distributions

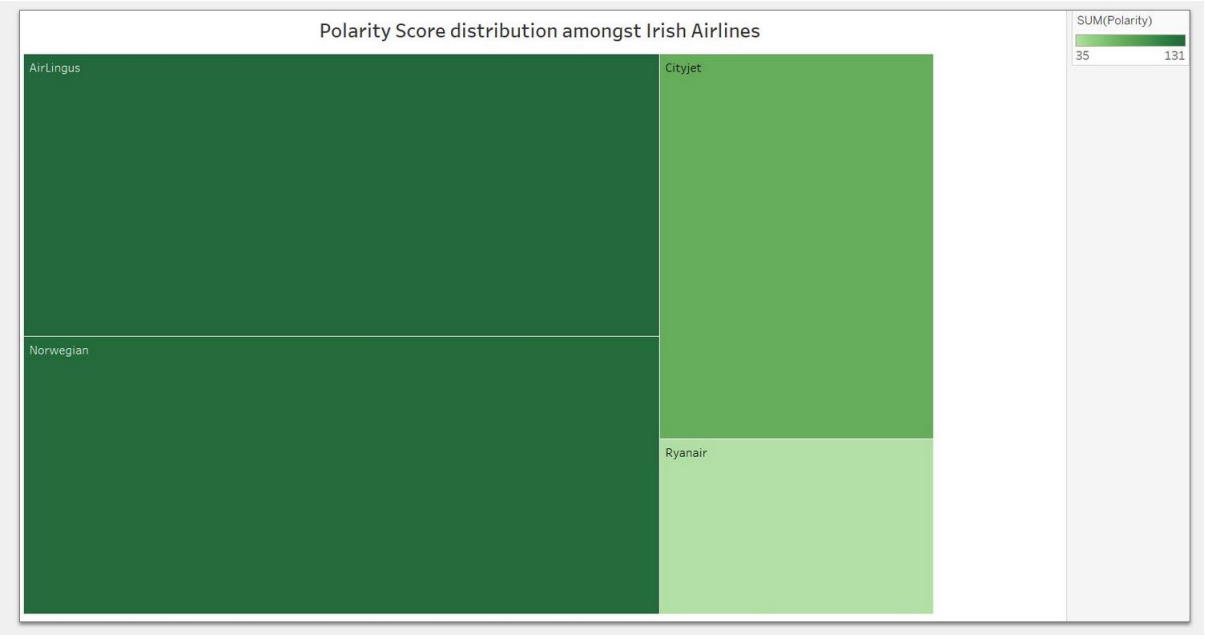

Figure 8: Polarity Score Distributions

4.5 Implementation , Evaluation and Results of Aspect Based Sentiment Modeling

In this chapter supervised machine learning techniques are implemented to tackle the research question. All of the techniques are evaluated in R using different machine learn-

ing packages provided by R. Research project objective D and each of sub-objectives are performed in here. Different machine learning algorithms were tested to derived expected results. For the research project solution candidate particularly have chosen on depend on previous research as described in chapter 2. Mainly five machine learning algorithms such as gradient boosting, k-nearest neighbour, linear regression, naive bayes and support vector machine are implemented in this chapter.Results are compared and analyzed with three measures that are accuracy,precision and recall.

As stated by Forman (2003) each of them are good in different solutions and each of them derived for different reasons. McLaughlin and Herlocker (2004) stated that machine learning algorithms can be performed and measured by using several evaluation functions. Accuracy, Precision and recall are one of the well known evaluations. Accuracy also known as the closeness of value that are derived to the standard known value. Following formula gives more clear view to calculate accuracy.

Assume TP is true positive,TN is true negative, FP is false positive and FN is false negative. then,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

The precision is the anticipated positive value.Ideal model will only predict positive classes for the situation Lantz (2013). In other words if the model gained high precision then it has high chances to be true. Following is the formula to calculate precision,

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall is the number of all true positive values upon the total true positive valuesLantz (2013).Recall is the sensitivity derived from the confusion matrix.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

In Following sections Implementation, evaluation and results are presented for five machine learning models and finally results are derived by analyzing confusion matrix with accuracy,precision and recall.

4.5.1 Implementation, Evaluation and Results of Gradient boosting Model

Gradient boost construct additive models for clasification as well as for regression techniquesFriedman (2002) Gradient boosting works in three phases. First phase of this algorithm is to optimized and analyzed loss function and loss function is user defined function. Second stage is to use weak learner for prediction, often decision tree are used as a weak learner in gradient boosting. Last task is to add weak learners by creating additive model for minimizing function loss. In other words gradient boosting takes weak learner in assemble approach and use them all together to predict high accuracy and in many cases those weak learners are decision trees.

Implementation :In this research to implementation of gradient boosting model is done using "gbm" library provided by R. Polarity score of each review is taken as independent variable. " Gaussian" method is used for gradient boost distibution and number of trees are 10000. accuracy,precision and recall is derived by accuracy.mean function.

Evaluation and Results :Overall accuracy gained by gradient boosting model is 54.78% which is well enough to predict the polarity.Precision gained is 0.99% which tells that positive predictions for this model is quite high. Recall achieved is 0.99% which also contributes this model to achive high positive predictions.All of three measures is calculated in this section hence research project objective DI (Chapter 1, Section 1.3) is successfully performed.Results are shown in following (Figure.9)

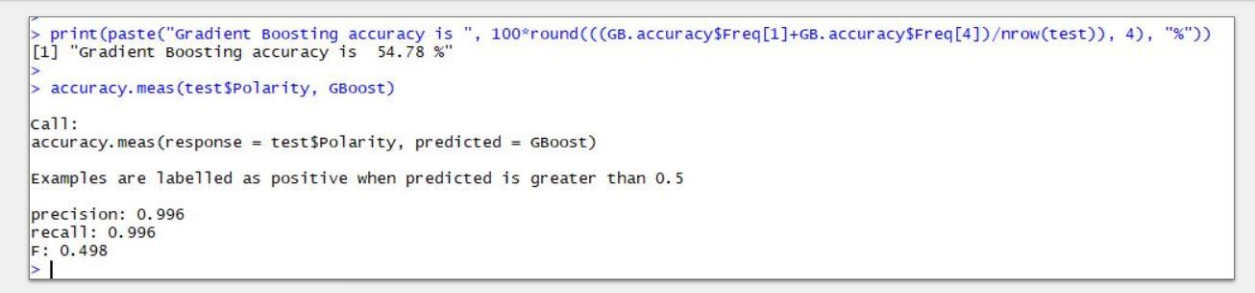

Figure 9: Gradient Boosting Results

4.5.2 Implementation, Evaluation and Results of K-Nearest Neighbour Model

KNN is the non-parametric method in pattern recognition which is mainly used for classification and regression. KNN is the simple model which stores all the causes and classify new cases in the basis of distance. By Jiang et al. (2012)KNN is often used for text classification and it is one of the best algorithm when problem addresses text classification. KNN works on the minimum distance to the training samples to calculate the k-nearest neighbour. KNN algorithm is chosen based on previous related work as mentioned in chapter 2. Many researchers used KNN and achieved high measures although model performance varies because its depends upon the data set.For the classification problems KNN uses hamming distance to achieve k nearest neighbour Formula for k-nearest neighbour to find hamming distance is given below,

$$HammingDistance(D) = \sum_{i=1}^{k} |x_i - y_i| \tag{4}$$

Implementation: In this research for the implementation of the K-nearest neighbour package "caret" is used which is provided by R. Independent variable taken is polarity score of each reviews. Initially k is set to 5 for the KNN. confusion matrix are calculated here to get accuracy, precious and recall for the KNN model.

Evaluation and Results: Overall accuracy gained by K-nearest neighbour is 57.5% which is quite good and bit better than last model. In confusion matrix sensitivity refers to recall and pos pred values refers to precision. KNN gained 0.53 of recall and 0.51 precision which directly contributes to the positive prediction of the model. All of three measures is calculated in this section hence, objective DII (Chapter 1, Section 1.3) is successfully performed.Results are shown in following (Figure.10)

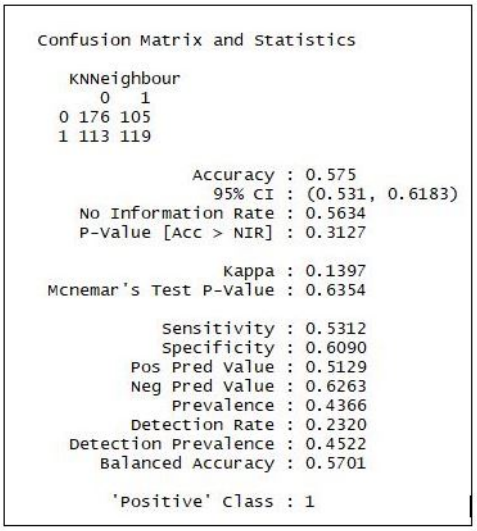

Figure 10: K-Nearest Neighbour Results

4.5.3 Implementation, Evaluation and Results of Linear Regression Model

Linear regression is the technique in supervised machine learning which used to model relationship between dependent and independent variable.By Andrews (1974) it is one of the most used algorithm in every discipline linear regression makes the relationship between dependent and independent variables by using best fit line and this best fit line is calculated using above formula,

$$Y = a * X + B$$

where, Y= dependent variable, a=Slope,x=Independent variable and b=intercept (5)

Implementation: In this research linear regression model is developed by method lm() which is provided by R. First method applied on train data set with the independent variable Polarity score for the each review. Then, using predict() function accuracy, recall and precision are measured by the accuracy.meas function. Linear regression results always depends upon the dataset.

Evaluation and Results: Overall accuracy gained by linear regression is 0.19% which is not ideal result because this accuracy define the model performance. Precision gained by linear regression is 0.45 which is not worst and directly contributes to the positive prediction for model. Recall achieved here is 1.00 which is almost high and also sign of the positive prediction. All of three measures is calculated in this section hence, objective DIII (Chapter 1, Section 1.3) is successfully performed.Results are shown in following (Figure.11)

```
> linear= predict(linear,test)
> LR<-table(test$Polarity, linear)
> linearReg.accuracy.table <- as.data.frame(table(test$Polarity, linear))
> print(paste("Linear Regression accuracy is ", 100*round(((linearReg.accuracy.table$Freq[1]+linearReg.accuracy.table$Fr
eq[4])/nrow(test)), 4), "%"))
[1] "Linear Regression accuracy is  0.19 %"
> accuracy.meas(test$Polarity, linear)

Call:
accuracy.meas(response = test$Polarity, predicted = linear)

Examples are labelled as positive when predicted is greater than 0.5

precision: 0.459
recall: 1.000
F: 0.315
>
```

Figure 11: Linear Regression Results

4.5.4 Implementation, Evaluation and Results of Naive Bayes Model

Naive bayes algorithm follows the technique of classification which refers the "Bayes theorem " for classification. It assumes the Independence within variables. As stated by Leung (2007) naive bayes find the probabilities within class with the other classes. In simple terms, naive bayes assumes that particular aspect within class is not related to any other aspects within class, it is independent on its own. P(c—x)=P(x—c)P(c)/P(x), where P(c—x) is probability of class P(x—c) is probability of predictor class, P(c) is the initial probability and P(x) is the predictor probability.

Implementation: In this research for implementation of naive bayes model package "e1071" is used which is provided by R. Polarity score is the independent variable here to gain the model performance. Initially train data is used to train the model.Later, model is implemented using test data set. Accuracy, precision and recall are calculated using confusion matrix.

Evaluation and Results: Overall accuracy gained by naive bayes is 79.14% which is highest accuracy gained till the execution of this algorithm. Accuracy defines the high performance of model. because this accuracy define the model performance. Precision gained by naive bayes is 0.62 which is directly contributes to the positive prediction for model. Recall achieved here is 0.88 which is very high and supports model to be positive prediction. All of three measures is calculated in this section hence, objective DIII (Chapter 1, Section 1.3) is successfully performed.Results are shown in following (Figure.12)

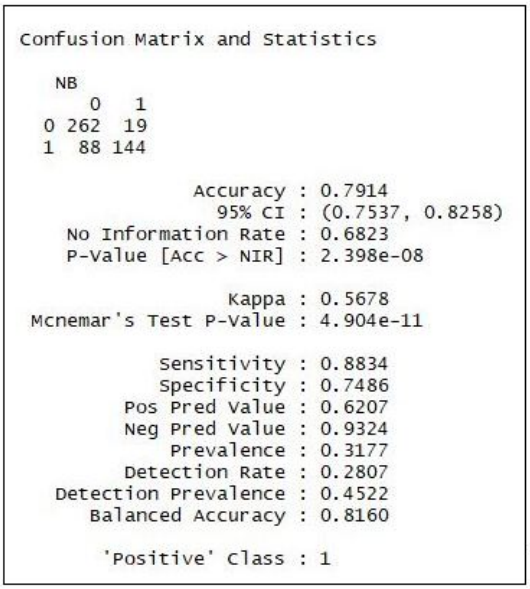

Figure 12: Naive Bayes Results

4.5.5 Implementation, Evaluation and Results of Support Vector Machine

Support vector machine is one widely used in supervised machine learning algorithm in regression as well as in classification. Pannala et al. (2016) this method is used to find different patterns in a data set. SVM working is simple that it analyze the border of decision and follows decision plane principle.It classifies classes with maximum gap.This method works well with less samplesPannala et al. (2016)

Implementation: In this research for the development of support vector machine library rose and library caret is used. First train is used for training the model with independent variable polarity using svm() function provided by R. Results are derived for the test data set after training the model. accuracy, recall and precision are measured by the confusion matrix function.

Evaluation and Results: Overall accuracy gained by support vector machine is 95.13% which very high and it describes that the performance of the model predicts as high because accuracy defines the high performance of model because this accuracy define the model performance. Precision gained by support vector machine is 0.90 which high enough to contributes to the positive prediction for model. Recall achieved here is 0.99 which is very high and supports model to be positive prediction. All of three measures is calculated in this section hence, objective DIII (Chapter 1, Section 1.3) is successfully performed.Results are shown in following (Figure.13)

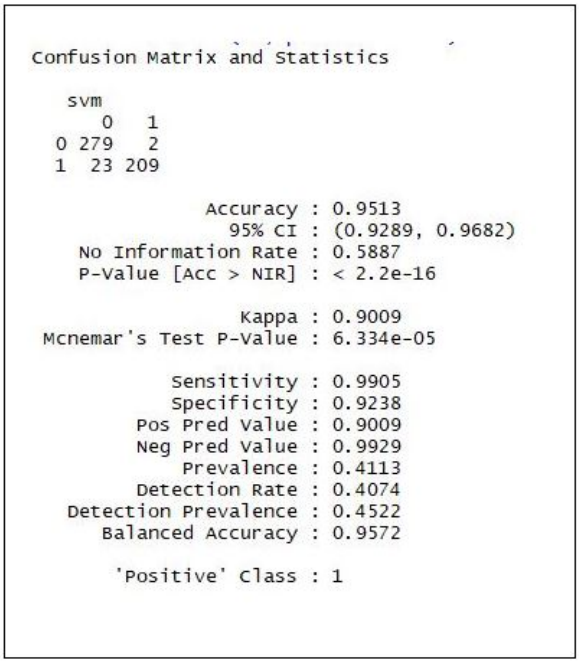

Figure 13: Support Vector Machine Results

4.5.6 Comparison of Developed Models

Algorithms	Accuracy (%)	Precision	Recall
Gradient Boosting	54.78	0.99	0.99
K-Nearest Neighbour	57.5	0.51	0.53
Linear Regression	0.19	0.45	1
Naïve Bayes	79.14	0.62	0.88
Support Vector Machine	95.13	0.9	0.99

Figure 14: Support Vector Machine Results

From the above table (Fig.14) its is noticed that support vector machine perform well along with the naive bayes and k-nearest neighbour. Support vector machine achieved highest accuracy as 95.13% with the highest precision and recall as 0.99 and 0.99 respectively.This is high enough to call it as a best fit model. The Naive Bayes also performed well by giving 79.14% of accuracy. Naive bayes achieved 0.62 precision and 0.88 recall which is second higher well performed model. Gradient boosting performance was stable for accuracy and received 54.78% of accuracy but high for the precision and recall. Linear regression is not up for the best fit as it performed low among all models. K-nearest neighbour is in the range of 50 for three measures. Hence it is proved that Support vector machine and naive bayes are the best-fit algorithms for the research project. This section also answers the research objectives.

5 Conclusion and Future Work

In research project aspect level sentiment analysis on Irish airlines has been carried out using various natural language processing libraries in R and python. Different supervised machine learning models has been tested on data .However, support vector machine gives highest accuracy of 95.13%. This research is mainly divided into three parts etc. to gather online reviews, perform aspect level sentiment analysis and supervised developed machine learning models.

To the best of the candidate knowledge, in past few decades inadequate research has been done in aspect based sentiment analysis with airline domain. This research is useful to bridge the gaps. This research contributes to the customers of Irish airlines and helps them to choose better airline according to their needs. Also , it significantly contributes to the Irish airline industry to know their customers and tackle their competitors. As stated in (section 1.2) all the research question objectives have been accomplished.

Future Work: This research has been carried out for the English language only. Therefore, this research has wide scope for future work in other language as each language has its own structure of grammar rules. While deploying the supervised machine learning models only liner regression gave least accuracy therefore, this part could be investigate in future work. Also, research been carried out within three months of time span. thousands of GB data could be collected within large time span and again this research could be preform to get more accurate insights. Supervised machine learning techniques are used in research therefore for unsupervised machine learning techniques could be used as a future work.

Acknowledgment I would specially like to thank my Supervisor Dr.Catherine Mulwa for her continuous guidance and supporting me through out the semester. I would like to thank my Mom and Dad for their support and trust towards me. I would also like to acknowledge my dearest friend Mahesh Talekar who supported me and was with me in my hard times.

References

Abirami, A. and Gayathri, V. (2017). A survey on sentiment analysis methods and approach, *Advanced Computing (ICoAC), 2016 Eighth International Conference on*, IEEE, pp. 72–76.

Adeborna, E. and Siau, K. (2014). An approach to sentiment analysis-the case of airline quality rating., *PACIS*, p. 363.

Andrews, D. F. (1974). A robust method for multiple linear regression, *Technometrics* **16**(4): 523–531.

Baharudin, B. et al. (2010). Sentence based sentiment classification from online customer reviews, *Proceedings of the 8th International Conference on Frontiers of Information Technology*, ACM, p. 25.

Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G. A. and Reynar, J. (2008). Building a sentiment summarizer for local service reviews, *WWW workshop on NLP in the information explosion era*, Vol. 14, pp. 339–348.

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* .

Ding, X., Liu, B. and Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining, *Proceedings of the 2008 international conference on web search and data mining*, ACM, pp. 231–240.

El-Halees, A. M. (2015). Arabic text classification using maximum entropy, *IUG Journal of Natural Studies* **15**(1).

Fang, X. and Zhan, J. (2015). Sentiment analysis using product review data, *Journal of Big Data* **2**(1): 5.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification, *Journal of machine learning research* **3**(Mar): 1289–1305.

Friedman, J. H. (2002). Stochastic gradient boosting, *Computational Statistics & Data Analysis* **38**(4): 367–378.

Gupte, A., Joshi, S., Gadgul, P., Kadam, A. and Gupte, A. (2014). Comparative study of classification algorithms used in sentiment analysis, *International Journal of Computer Science and Information Technologies* **5**(5): 6261–6264.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews, *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 168–177.

Jiang, S., Pang, G., Wu, M. and Kuang, L. (2012). An improved k-nearest-neighbor algorithm for text categorization, *Expert Systems with Applications* **39**(1): 1503–1509.

Khan, A. Z., Atique, M. and Thakare, V. (2015). Combining lexicon-based and learning-based methods for twitter sentiment analysis, *International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE)* p. 89.

Lantz, B. (2013). *Machine learning with R*, Packt Publishing Ltd.

Leung, K. M. (2007). Naive bayesian classifier, *Polytechnic University Department of Computer Science/Finance and Risk Engineering* .

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.

Medhat, W., Hassan, A. and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey, *Ain Shams Engineering Journal* **5**(4): 1093–1113.

Misopoulos, F., Mitic, M., Kapoulas, A. and Karapiperis, C. (2014). Uncovering customer service experiences with twitter: the case of airline industry, *Management Decision* **52**(4): 705–723.

Pang, B., Lee, L. et al. (2008). Opinion mining and sentiment analysis, *Foundations and Trends® in Information Retrieval* **2**(1–2): 1–135.

Pannala, N. U., Nawarathna, C. P., Jayakody, J., Rupasinghe, L. and Krishnadeva, K. (2016). Supervised learning based approach to aspect based sentiment analysis, *Computer and Information Technology (CIT), 2016 IEEE International Conference on*, IEEE, pp. 662–666.

Perikos, I. and Hatzilygeroudis, I. (2017). Aspect based sentiment analysis in social media with classifier ensembles, *Computer and Information Science (ICIS), 2017 IEEE/ACIS 16th International Conference on*, IEEE, pp. 273–278.

Polanyi, L. and Zaenen, A. (2006). Contextual valence shifters., *Computing Attitude and Affect in Text* **20**: 1–10.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O. et al. (2016). Semeval-2016 task 5: Aspect based sentiment analysis, *ProWorkshop on Semantic Evaluation (SemEval-2016)*, Association for Computational Linguistics, pp. 19–30.

Qiang, G. (2010). Research and improvement for feature selection on naive bayes text classifier, *Future Computer and Communication (ICFCC), 2010 2nd International Conference on*, Vol. 2, IEEE, pp. V2–156.

Ronglu, L., Jianhui, W., Xiaoyun, C., Xiaopeng, T. and Yunfa, H. (2005). Using maximum entropy model for chinese text categorization [j], *Journal of Computer Research and Development* **1**: 22–29.

Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Díaz-Rangel, I., Suárez-Guerra, S., Treviño, A. and Gordon, J. (2012). Empirical study of machine learning based approach for opinion mining in tweets, *Mexican international conference on Artificial intelligence*, Springer, pp. 1–14.

Singh, V. K., Piryani, R., Uddin, A. and Waila, P. (2013). Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification, *Automation, computing, communication, control and compressed sensing (iMac4s), 2013 international multi-conference on*, IEEE, pp. 712–717.

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.

Xu, J. and Li, H. (2007). Adarank: a boosting algorithm for information retrieval, *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 391–398.