National College of Ireland

Classification of Airline Customer Sentiment Expressed in Twitter Tweets using Lexicons, Decision Tree, and Naïve Bayes

MSCDATOP

Liam Higgins
Student ID: x21182523

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Liam Higgins
Student ID:	x21182523
Programme:	MSCDATOP **Year:** 2022
Module:	MSc Data Analytics
Supervisor:	Jorge Basilio
Submission Due Date:	15th August 2022
Project Title:	Classification of Airline Customer Sentiment Expressed in Twitter Tweets using Lexicons, Decision Tree and Naïve Bayes
Word Count:	7629..... **Page Count:** 22

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 required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Liam Higgins

Date: 13th August 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

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):	

Classification of Airline Customer Sentiment Expressed in Twitter Tweets using Lexicons, Decision Tree, and Naïve Bayes

Liam Higgins
x21182523

Abstract

This paper describes Natural Language Processing (NLP) and Machine Learning approaches to Sentiment Analysis of Twitter Tweets relating to commercial passenger airlines. By extracting and analysing textual data obtained in real-time from the social media platform, Twitter, the research proposes a methodology to collect, process, and interpret emotional responses contained within Tweets. Two main approaches to classifying sentiment are described. Firstly, lexicon-based approaches using three valence lexicons (Syuzhet, Afinn, and Bing) and one emotion lexicon (NRC) to determine the semantic orientation of words found within Tweet text are discussed. Secondly, two supervised machine learning classification algorithms (Naïve Bayes and Decision Tree) are used to perform sentiment classification. The goal of the research is to provide a diverse and commercially useful method for airlines to monitor customer sentiment relating to their experiences of airline services. The importance and commercial application of obtaining customer insights from Tweets which have been posted online and describe customer experiences and attitudes is discussed. The paper aims to provide airlines with a means to improve service offerings, differentiate from competitors, and gain competitive advantage based on analysing customer sentiment to their services. A maximum accuracy of 71% was achieved using a Naïve Bayes classifier algorithm.

1 Introduction

Twitter tweets are a form of unstructured data primarily consisting of text but may also include unique identifiers, dates, timestamps, geographical data, images, and Uniform Resource Locators (URLs). In contrast to structured data kept in fielded form in relational databases, unstructured textual data often contains inconsistencies and ambiguities that make it difficult to analyse using conventional methods. According to several estimates (Grimes 2008) unstructured data accounts for approximately 80% of all enterprise data globally. In contrast to structured data, unstructured data isn't as easily analysed using traditional machine learning techniques which is arranged in a searchable tabular fashion, such as a database. However, textual data can provide more nuanced insight into a topic domain as human emotional responses can be determined.

To analyse consumer social media discussions and interactions and to acquire insights into general customer behaviour and preferences, the extraction of Tweets and subsequent application of Natural Language Processing (NLP) can be highly commercially valuable. Due to its effectiveness in language modelling, developments in NLP have continued to gain

popularity in textual analysis. Language models are used in many different industries because they offer reliable, timely, and cost-effective methods for text analysis. One of the most common forms of NLP is sentiment analysis. Sentiment analysis is the technique of computationally finding and classifying opinions conveyed in a text, primarily to ascertain if the writer has a positive, negative, or neutral attitude towards a given topic, product, or service. The ability to translate emotions and attitudes into useful information through sentiment analysis is becoming increasingly important to understand customer attitudes and satisfaction.

Sentiment analysis automates the extraction of attitudes, opinions, viewpoints, and emotions from textual data, such as Twitter Tweets, and can provide data visualisations and statistical outputs. One of the most significant functions in text mining is visually portraying the content of a text document and these can take the form of graphs, plots, and word clouds.

Sentiment analysis relates to the polarity strength of words and phrases contained in text, whereas semantic analysis refers to the process of extracting subjectivity from the text. If a sentence includes non-factual information like individual forecasts, judgments, and personal opinions, it is said to be subjective. If a sentence is factual rather than subjective, it is considered objective. A decimal value in the range of [-1,1] is used to indicate the polarity of a text and indicates the sentence's tone is positive or negative.

To compare and analyse airline service, gain a competitive edge, and generate commercial value, this paper describes NLP techniques for analysing and forecasting consumer sentiment in the airline industry by analysis of Twitter Tweets. The goal is to provide companies in the airline industry with accurate models so they can monitor customer satisfaction levels with their services. The paper outlines how passenger experiences which have been posted online via Twitter can be used to analyse whether consumers have a positive or negative attitude towards an airline’s product or service. Sentiment analysis is useful for social media monitoring since it gives a broad picture of how the public feels about their airline experiences.

2 Background & Motivation

Commercial passenger airlines were negatively impacted by a drop in passenger numbers and earnings during the COVID-19 pandemic. Traveller numbers decreased because of a sharp drop in demand and government-imposed restrictions on cross-border travel. As the industry continues to recover from over two years of contraction, airlines are competing for customers to return to profitability. In this context, a research project has been undertaken to explore and analyse data from the airline industry in the form of sentiment analysis. By exploring different approaches to sentiment analysis, it is suggested that a relatively low-cost method of gaining customer insight can be provided to airlines. Gaining recent and relevant consumer insight can lead to better product and service differentiation from competitors, and ultimately provide a competitive advantage in an industry that is still recovering from severe contraction due to COVID-19.

Twitter is the most popular microblogging platform globally and allows users to publish images, videos, and text, in the form of tweets, up 140 characters in size. User tweets allow researchers and companies to obtain meaningful data relating to users, their likes,

dislikes, and preferences. Airport facilities, check-in speed, interactions with airline personnel, travel class, onboard space and comfort, inflight entertainment, service, food, luggage delivery, delays, and a variety of other factors all have an influence on airline customer sentiment. The purpose of this research is to determine how these important factors influence passenger sentiment, as well as to determine if there is a relationship between these factors and general consumer attitude toward different airlines operating out of Irish airports.

In a connected world, electronic word-of-mouth (eWOM) plays an increasing role in influencing consumer perceptions of airline service, quality, and brand awareness. To enable customers to voice their opinions and score features of the services they have purchased, several online review platforms have evolved. Data can be obtained from these online platforms and businesses can analyse customer perceptions of their product or service offerings using the wealth of information provided by reviewer comments and ratings. A crucial tool that could provide firms an edge over rivals is the collecting and analysis of customer data into pertinent information and analytics.

Understanding consumer expectations and perceptions of satisfaction is crucial for all organisations. Businesses require customer feedback to understand whether customers are satisfied with their goods and services. There are several methods for obtaining customer feedback. Surveys of customer satisfaction that are conducted prior to, during, or after a trip is a typical traditional approach. A small sample of customers may make up the survey's respondent pool, which may not be reflective of the larger population. Survey completion resistance is a frequent occurrence and as such this approach may potentially have a narrower application. Using machine learning techniques, it is relatively simple to analyse and make predictions about qualitative data, such as that obtain via surveys. An NLP strategy is needed for quantitative data, such as written text reviews from Twitter.

Customers are becoming more involved in the creation of travel-related advice and reviews rather than just being passive consumers of it. Customers can now share their positive or negative feelings or experiences with an audience online. To keep businesses informed about customer satisfaction with their goods and services, this customer feedback can be acquired and analysed. Airlines can leverage this source of data to help focus on offering goods and services that elicit favourable emotional responses from clients to attract new and repeat business and stand out from rivals in a positive way.

3 Research Question & Objectives

Sentiment analysis can identify how customers feel about the features and benefits of airline products and services. This can help uncover areas for improvement that airlines may otherwise not have been aware of.

3.1 Research Question

Can accurate insights on customer satisfaction towards airline services be obtained by performing sentiment analysis on data collected from the social media platform Twitter?

3.2 Research Objectives

Table 1. Research Objectives

Research Objective One:	Critically review previous literature on sentiment analysis using both NLP and Machine Learning techniques.
Research Objective Two:	Extract and analyse appropriate publicly available textual data from Twitter, conduct pre-processing, feature extraction, cleaning, stemming, and visualisation of the data.
Research Objective Three:	Build three word polarity sentiment analyses, one semantic score sentiment analysis using existing lexicographical dictionaries, and two machine learning sentiment analysis models.
Research Objective Four:	Compare and contrast four lexicographical sentiment analyses with two machine learning sentiment analysis techniques. Critically analyse the results of the textual sentiment classification models and describe their commercial application and value.

4 Research Methodology

The research will use a modified version of the Cross Industry Standard Process for Data Mining (CRISP-DM), which has seven steps instead of the usual six (see Figure 1 below). Text mining via connection to the Twitter API and subsequent pre-processing of the data will be conducted. Data cleaning techniques, including anonymisation, error removal, removing special characters, punctuation, whitespace, stopwords, and word stemming will follow. Visualisations will be made to convey essential characteristics of the data. Feature engineering will involve appropriate variable manipulation, vectorisation, the creation of textual corpus, and term document matrices (TDM) for appropriate preparation of the data. The selection of appropriate sentiment lexicon libraries and creation of two machine learning classification models will be carried out. The last stage will involve publishing results, assessing the findings, comparing them to those of other studies, and examining any ramifications of the findings. The findings will also be examined for any potential bias or ethical concerns.

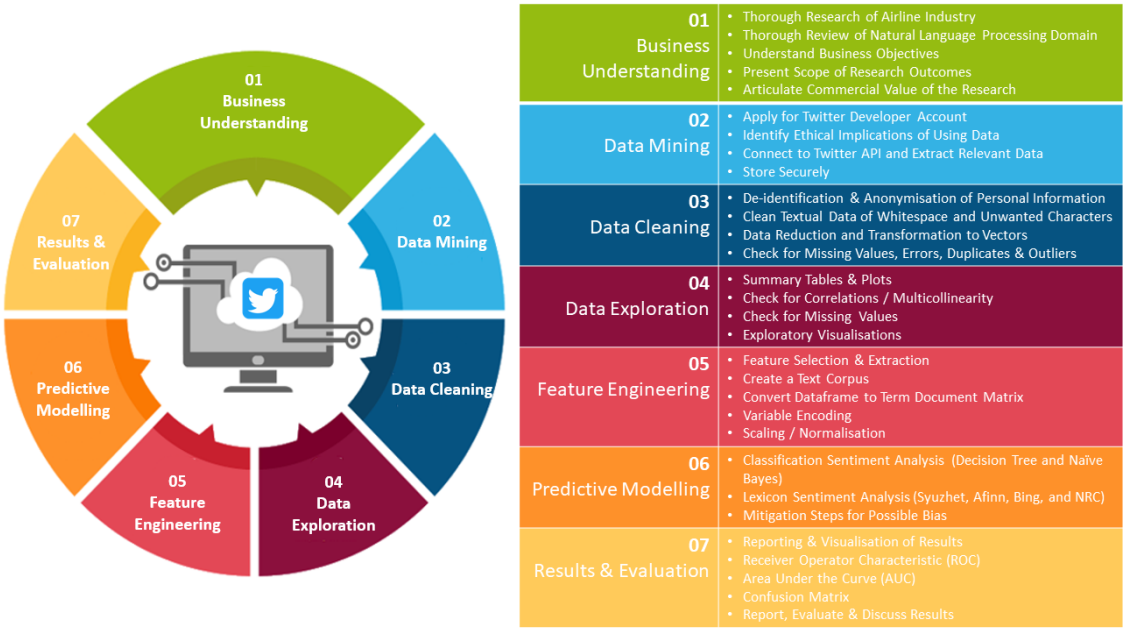

Figure 1. Modified CRISP-DM Research Methodology Infographic.

5 Related Work

5.1 Introduction to Previous Literature

Positively differentiating a product or service from competitors can provide competitive edge leading to increased revenue and market share. In competitive industries companies must create competitive advantage by acquiring, nurturing, and sustaining customers (van Doom et al, 2010). To encourage positive perceptions of a company's brand they must at a minimum satisfy and preferably exceed customer expectations to elicit a positive emotional response. Customer sentiment and satisfaction are influenced by how customers emotionally perceive their experiences. The following review of different approaches to determining customer sentiment and satisfaction provides context for the research objectives of this paper.

5.2 Lexicon-Based Approaches

A lexicon is a dictionary of words which have each been labelled with emotions or sentiments. The labels express a particular emotion (such as joy, trust, anger, etc.) or may be labelled as positive, negative, or neutral. To assess if a statement is positive or negative, this lexicon has been annotated with the polarity and strength of each word or group of words (Taboada et al, 2011). The technique aims to determine how likely each phrase is to be good or bad (Chaovalit and Thou, 2005). Lexicon and semantic orientation are two popular methods which are widely used in NLP.

During text mining research of online customer reviews of airlines using NLP a Latent Dirichlet Allocation (LDA) model to find textual terms and phrases associated with airline customer satisfaction was published in the Journal of Air Transport Management Research (Lucini et al, 2020). The LDA model produced an accuracy of 80% when data from 419 airlines throughout the world was analysed and modelled.

Another paper outlining sentiment analysis, text classification, and predictive modelling of reviews from the online travel agency, Booking.com, to study the aspects of perceived service quality and their contribution to the overall rating of service experience (Kolomoyets, 2017). To classify the textual data into positive, neutral, or negative sentiments, the researcher employed the *word2vec* library, Support Vector Machine (SVM) and an unsupervised LDA algorithm for sentiment classification. Following the completion of data pre-processing, the author built several predictive models using Naive Bayes, SVM, and Random Forest. The models were then assessed for accuracy using five-fold cross validation, Cohen's Kappa, and F1 score. The findings showed that perceived quality in relation to purchases was most strongly influenced by hotel staff responsiveness, staff empathy, and perceived transactional simplicity of online booking. Although the author proposes that hotel managers may use this knowledge to enhance customer relations, it is challenging to envision how managers could foster greater empathy in their personnel. The other two most powerful predictors, the simplicity of the booking process and the responsiveness of the staff, are two aspects of the travel experience with scope to be improved.

A paper describing a machine learning-based methodology to assess the many factors that influence consumer satisfaction and using data from 100,000 customer reviews acquired from

Airbnb (Chiny et al. 2021) the authors applied an NLP approach to labelling words and phrases. Subsequently, using multiple linear regression, support vector regression (SVR), and artificial neural networks (ANN), independent variables such as description accuracy, cleanliness, check-in, communication, location, and value were used to predict overall customer satisfaction as well as satisfaction within different customer segments. The results show that customer interest in satisfaction measures is not uniform. Depending on the client's segment, variations were found. According to the authors, ANN had reasonable predictive capabilities for each of the three customer classes, with the ability to forecast anticipated customer satisfaction among individual guests to a maximum of 67%, couples to a maximum of 68%, and families to a maximum of 71%. It is reasonable to question whether results that only demonstrate moderate predictive accuracy should serve as the basis for a recommendation to alter the platform's current customer satisfaction and review metrics. The authors of the study suggest that Airbnb's rating system be improved to make it more appropriate for each group in which the guests fall considering their findings.

5.3 Machine Learning Approaches

No single indicator can accurately predict positive customer sentiment leading to customer loyalty. If companies want to customers to have positive sentiments towards their products, they must take a balanced and holistic approach to multiple aspects of the customer experience (Keiningham et al. 2007). This includes pricing, product or service range, customer service, redress of customer complaints, marketing and advertising, value for money, and various other industry specific customer metrics. The researchers discovered that a multivariate strategy proved to be reliable and accurate predictive model using both single and multiple logistic ordinal regression algorithms. Repurchase and customer recommendations were strongly correlated to perceived metrics of value, service expectations, and overall customer satisfaction, according to the authors, who evaluated the model using Sensitivity, Specificity, Receiver Operating Characteristic, and Area Under the Curve (AUC). According to their results, incorporating more customer-related variables resulted in more accurate model outputs.

Research was conducted to investigate whether words and phrases used to describe hotel guest experiences could be used to predict whether a guest would rate a hotel as positively or negatively (Sánchez-Franco, Navarro-Garca, and Rondán-Catalua 2019). The authors created a Naïve Bayes model to predict customer satisfaction using Yelp data on 33 hotels in Las Vegas, USA. Their predictive model produced a maximum accuracy of 86%, and the findings demonstrate how hotels can enhance their offerings by considering elements like employee experience, professionalism, tangible and experiential qualities, and gambling-related attractions.

Using crowdsourced data from blogs, social networks, forums, and customer reviews, (Catal and Nangir 2017) implemented a novel Multiple Classifier System (MCS) to predict people's sentiments regarding a range of goods and services offered in the Turkish market. According to the authors, their unique MCS model was a slightly better predictor than Naive Bayes and SVM in terms of predictive accuracy, with a maximum accuracy of 86% when compared to 85% and 83% respectively.

During research on predicting the likelihood of customers making repeat trips on airlines several algorithms were used on data gathered from a 2017 survey (Hwang et al. 2020). Each model had moderate to strong predictive accuracy of customer loyalty. The research differentiated between full-service airlines and low-cost carriers (LCCs), as respondents from these two market segments appeared to place greater emphasis on various service-related factors, with value for money being most important to LCC customers and good overall service being more important to full-service airline customers. A maximum prediction accuracy of 80% was found using the XGBoost classifier (a gradient boosted decision tree ensemble method). The authors found that airlines that focused on providing high-value services had higher statistically significant predictions about whether customers would make additional purchases from them.

3.5 million guest reviews of 13,000 hotels in 80 countries were obtained via web scraping from the TripAdvisor website and used in a multivariate analysis of customer satisfaction in the Hotel Industry (Radojevic et al, 2017)). The number of global online hotel reviews posted to the platform by users between the years of 2002 and 2016 showed exponential growth during the initial exploratory analysis of the data, demonstrating the platform's quick rise in popularity and the growing amounts of data available for such analysis. Overall Satisfaction (dependent variable), Reason for Travel, Location, Cleanliness, Rooms, Service, Sleep Quality, and Value (independent variables) were the criteria to create a multilevel regression. It was discovered that the goal of the trip has a significant impact on how satisfied customers are with hotel services, with business travellers being less likely than leisure tourists to be dissatisfied with any aspect of hotel service.

Another paper describes the use of several Machine Learning categorisation algorithms to client feedback in the form of Twitter Tweets (Gautam and Yadav, 2014). Author postings, or tweets, were classified as positive, negative, or somewhere in between by extracting adjectives from the dataset and building feature vectors using the lexical database *WordNet*. Using three machine learning-based classification algorithms, Naive Bayes, Maximum Entropy, and SVM, these vectors were then used to create classification predictions on the likelihood of consumers' semantic opinions. Each classifier was assessed using recall, precision, and accuracy. The researchers found that their model had a maximum accuracy of 90% in predicting whether customers would recommend a good or bad product or service.

Text mining of online airline reviews from the Skytrax website were used in a study to classify customer sentiment (Jain et al, 2019). After conducting a sentiment analysis to determine whether reviews could be labelled as positive or negative, three different classification algorithms, k-Nearest Neighbours (kNN), SVM, and Decision Tree, were used to build predictive models. The researchers found that SVM had a maximum accuracy of 83%, compared to Decision Tree's 75% and kNN's 65%. The very small sample size of data used in their research was one of its limitations, and it is proposed that a larger sample drawn from a wider range of customer reviews may have an impact on the accuracy of results.

Using data from 350 airlines obtained from the online airline review website airlinequality.com research was conducted to assess service aspects of airlines in an effort to make accurate predictions about how customers rank airline service quality (Alkhatib and Migdadi, 2014). The authors employed a stepwise multiple linear regression approach to forecast a passenger's numerical rating of an airline. They separated the market into the short,

medium, and long-haul routes, as well as international, domestic, each with a range of projected accuracy based on distinct predictor criteria. Despite their somewhat arbitrary and unclear choice of independent variables, including the number of seats per aircraft type, the authors showed that this technique produced moderate to strong predictive accuracies.

An innovative method for analysing airline service quality, perceived value, satisfaction, and behavioural intentions for flyers using a structural equation model (SEM) is presented in a research study (Chen 2005). SEM models, which combine factor analysis and regression, are suitable for those whose independent variables exhibit significant multicollinearity. The results show that rather than being only dependent on service quality, customer contentment is affected by a combination of service quality and other variables such as price and perceived brand strength.

In a comparative study on the effectiveness of employing an ANN, Naive Bayes, and SVM for a variety of products and services, including textual movie reviews (Moraes et al, 2013) the authors showed that ANN and SVM were strongly capable at classifying reviewer sentiment. A reported accuracy of 87% was reported for ANN, 84% for SVM, and 80% for Naïve Bayes.

To better understand geographic disparities in service perceptions and the weight of customer expectation placed on these services when deciding whether to give their airline experience a positive or negative rating, a study using data obtained via a web scrape to gather passenger evaluations from the Skytrax website was conducted (Punel et al, 2019). The authors found that each service category affected total customer satisfaction predictability using sentiment analysis and scoring each service criterion. Results showed that information obtained from online review websites, such as Skytrax, may be used to generate accurate forecasts on airline passenger sentiment.

Another study using data from Skytrax (Lacic et al, 2013) proposed two research questions. Firstly, which airline service criteria had the biggest impact on total customer satisfaction. Secondly, which factors in the Skytrax airline review data are the best predictors of passenger satisfaction with airlines. Along with the numerical information in the Skytrax reviews, the authors also employed the textual information by labelling the n-grams within using a technique called Suffix Tree Clustering (STC). The authors classified the material as positive, negative, or neutral before using Naïve Bayes, decision trees, and random forests to predict total passenger satisfaction. The authors found that inflight cabin personnel, onboard seat comfort, and perceived value for money were the factors of highest relevance when predicting overall happiness with an airline, with the best performing model returning a maximum prediction accuracy of 96%.

Building on earlier research, research using an ensemble of machine learning algorithms to classify airline consumer sentiment to service criteria was undertaken (Wan and Gao, 2015). Using ten-fold cross validation across six algorithms, Nave Bayes, SVM, Bayesian Network, Decision Tree, and Random Forest, were created and evaluated. The researchers found that using an ensemble of classifiers rather than one to generate predictions enhanced overall classification accuracy, with a maximum resulting predictive accuracy of 84%.

A study on using a variety of machine learning classifiers on Twitter data relating to airline customer sentiment describes three machine learning algorithms, SVM, ANN, and Convolutional Neural Network (CNN) (Kumar and Zymbler, 2019). By mapping categories

of Tweets to passenger sentiment, the authors were able to classify phrases included in Tweets posted by passengers about their experience traveling with a variety of US-based airlines. With an accuracy of 92% after 2700 iterations, CNN was the most accurate of the three methods employed.

5.4 Summary of Previous Research

Table 2: Summary of Previous Best Performing Research Methods

NLP Technique	Choice of Algorithm	Maximum Reported Accuracy	Authors
Sentiment Analysis	Random Forest	84%	Wan and Gao 2015
Sentiment Analysis	Random Forest	96%	Lacic et al 2013
Sentiment Analysis	Convolutional Neural Network (CNN)	92%	Kumar & Zymbler 2019
Sentiment Analysis	Multiple Classifier System (MCS)	86%	Catal & Nagal 2017
Sentiment Analysis	Support Vector Machine (SVM)	83%	Jain et al 2019

5.5 Gaps in Research

While several papers have been published in the domain of airline customer sentiment using Twitter Data, the majority of these have been based on data obtained from a now defunct website, CrowdFlower. The CrowdFlower dataset relating to US based airlines is popular on Kaggle, an online platform for the data science and machine learning community. There is no published record of the method for labelling the sentiment for this dataset and as such the papers must be viewed with some scepticism as data may be outdated or indeed the sentiment labelling methodology may be flawed.

The research outlined in this paper describes a reproducible way to connect to the Twitter API extract real-time data relating to user specified airlines, create a variety of sentiment analysis models using a variety of techniques, and outlines the strengths and weaknesses of their predictive outcomes. This method can provide most up to date and relevant data and provide airline executives with a valuable and low-cost tool to monitor customer sentiment to their services.

6 Research Methodology

There are several methods which can be employed to perform a sentiment analysis on textual data. Figure 2 below (Medhat, Hassan and Korashy, 2014) shows an overview of the different techniques. In a lexicon-based method, sentiment polarity is determined by the text’s words or phrases semantic orientation. A text's subjectivity and viewpoint are measured by its "semantic orientation." This rule-based technique scans a document for opinion words before categorising it according to the proportion of positive and negative terms. It categorises using dictionary polarity, negation words, booster words, idioms,

emoticons, and words with conflicting meanings. Each review is represented by statistical models as a combination of latent features and ratings. To cluster words into aspects and sentiments into ratings, on the premise that aspects and their ratings may be represented by multinomial distributions. By training on a given dataset, a machine learning technique employs a variety of supervised learning algorithms to determine the emotion. This paper describes lexicon-based methods and machine learning-based methods.

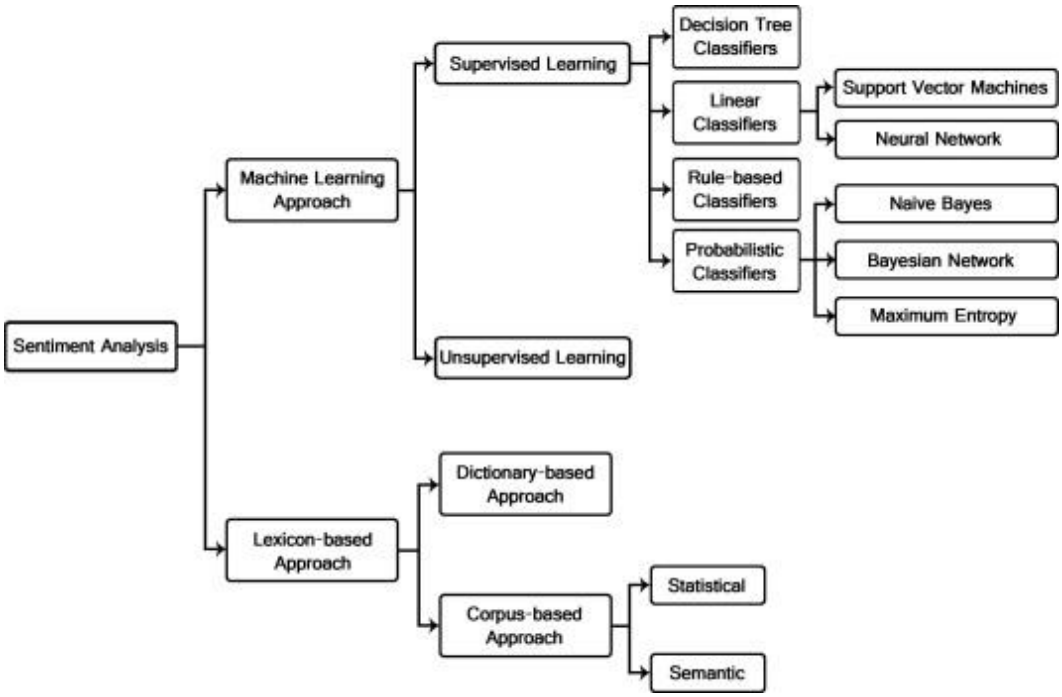

Figure 2. Varieties of Sentiment Analysis Methods.

6.1 Lexicon-based Sentiment Analysis

A lexicon is simply the vocabulary of a language and can be thought of as a semantic dictionary. The simplest method for conducting sentiment analysis is to use lexicon-based techniques. This method takes use of a dictionary that contains words that have already been pre-tagged with appropriate sentiment applied to different words within that dictionary. There are many different lexicons which can be used to perform sentiment analysis and the R programming language has several libraries containing different lexicons.

This paper outlines four approaches to sentiment analysis of Twitter Tweets using available packages in R: *Syuzhet*, *Rsentiment*, *SentimentR*, and *SentimentAnalysis*. These packages disregard syntax and grammar, instead relying on a bag-of-words method where the sentiment is inferred from the specific words that appear in the text. The text's words are compared to one or more lexicons that list positive and negative terms and associate them with a level of either positive or negative intensity. To arrive at an overall numeric score expressing the overall sentiment at the level of interest, the number of matches between the words in the text and the lexicons as well as the degrees of intensity of the sentiment connected to those words are calculated. The packages, accompanying lexicons, and functions used by each to calculate sentiment are described in the following sections.

6.1.1 Syuzhet Lexicon

Syuzhet has the option to select from one of four sentiment lexicons or create our own unique lexicon. Syuzhet, Afinn, Bing, and NRC are the four common lexicons. The most widely used lexicon is Syuzhet, which was developed in the Nebraska Literary Lab (Jockers, 2015) It comprises 10748 words with emotive values between [1, 1]. Negative keywords account for 7161 of the total 10748 words, leaving 3587 positive words.

6.1.2 AFINN Lexicon

The AFINN lexicon (Nielsen, 2011), which contains profane words and Internet slang phrases. Starting with a collection of obscene words it has been steadily expanded to include acronyms and abbreviations by looking at tweets and sets of words pulled from the Urban Dictionary and Wiktionary. 2477 words make up the vocabulary that results from this. There are more negative words (1598) than positive words (878) and neutral words (1), combined. The score range is substantially broader than that used in the Syuzhet lexicon at [5, 5].

6.1.3 Bing Lexicon

The Bing lexicon [Bing and Liu, 2004], often known as the Opinion Lexicon, has 6789 words, 2006 of which are positive, and 4783 of which are negative. The assigned score can be either -1 or 1.

6.1.4 NRC Lexicon

Compared to the previous lexicons, NRC (Mohammad and Turney, 2013) is different as it assigns sentiment by eight emotional categories: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. NRC consists of 13889 words. The calculation of emotions finds each word in each element of the input vector and calculates the algebraic sum of the values of its emotions. NRC gives each sentence a score for each semantic category instead of determining which category a word is in and obtaining algebraic scores for positive and negative words.

Table 3: Summary of Lexicons

Lexicon	Word Total	Positive Words	Negative Words	Resolution	Calculation Method to Obtain Score	Classification Method
AFINN	2477	878	1598	11	Score individual words & sum	Manual
Bing	6789	2006	4783	2	Number of positive words – number of negative words / total words	Manual
NRC	5555	2312	3324	2	Number of positive words – number of negative words / total words	Amazon Mechanical Turk
Syuzhet	10748	3587	7161	16	Score individual words & sum	Manual

7 Lexicon Based Approaches

The process of converting textual data into valuable visualisation using a lexicon-based approach is summarised in Figure 3 below.

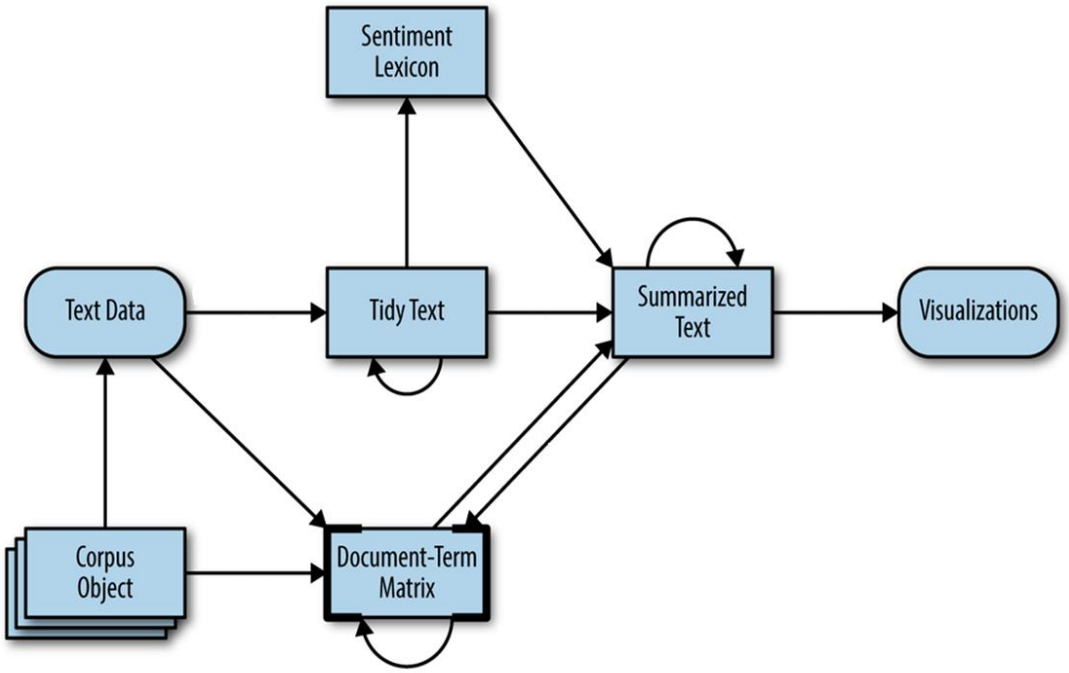

Figure 3. Lexicon Process Flow Diagram

8 Machine Learning Process

There are two types of machine learning methods for sentiment classification: supervised and unsupervised. For the purposes of this research two supervised machine learning methods have been chosen, Decision Tree and Naïve Bayes.

8.1 Sentiment Classification via Machine Learning Process Flow Diagram

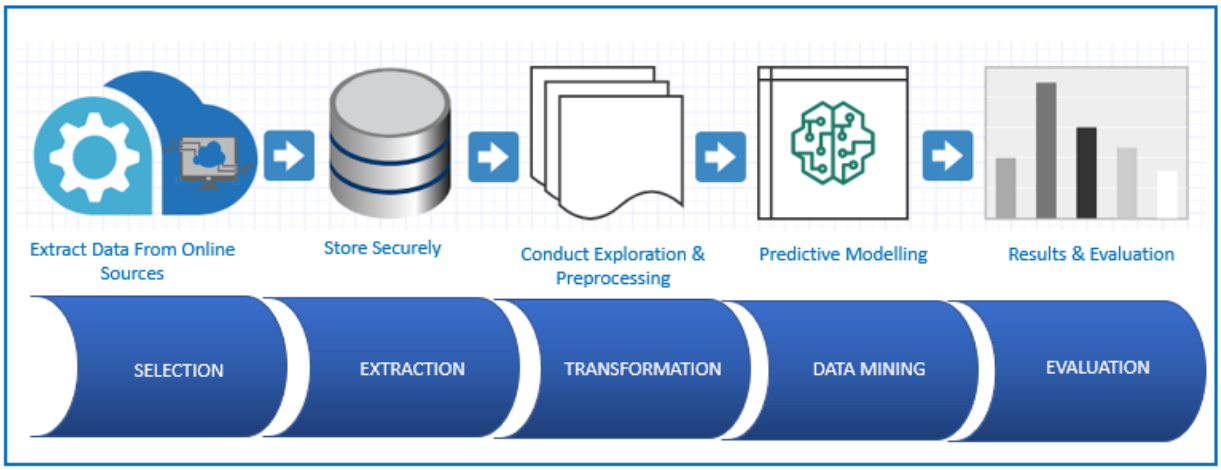

Figure 4. Machine Learning Process Flow Diagram

8.2 Decision Tree Classifier

A Classification and Regression Trees (CART) algorithm, often referred to as a Decision Tree, is the first machine learning technique to be used. The algorithm is used to discover whether the independent variables, in this case the words or phrases within text, can accurately predict the three levels of the dependent variable, namely "positive," "negative," or "neutral" sentiment. Decision trees iteratively divide and segment data into binary nodes and are a non-parametric method to determine outcomes based on the data rather than predetermined assumptions. Entropy and Gini impurity are the two primary factors used to construct decision trees. Entropy values range from 0 to 1, whereas the Gini impurity index ranges from 0 to 0.5. Entropy is a metric for how uncertain it is to assign a variable to a certain category. Gini impurity measures the likelihood that a variable will be misclassified during random splits at each Decision Tree node. These factors work together to balance the Decision Tree method, making it a flexible but occasionally prone to overfitting.

8.3 Naïve Bayes Classifier

One of the approaches most frequently for text data classification is the Naïve Bayes algorithm. The Naïve Bayesian technique assumes that dataset variables are independent of one another (hence the term "naïve"). When a certain class, or classes, must be predicted, the Naïve Bayes Classification method is useful as the algorithm determines whether each attribute has a greater or lesser likelihood of falling into a certain category, in this example, "positive," "negative," or "neutral" sentiment, based on its individual qualities.

The Naïve Bayesian classifier interprets the text of each Twitter Tweet as a bag-of-words, much like the Lexicon-based classifier does. A single tweet document is passed to the Naïve Bayesian classifier, which then estimates the probability of each feature occurring a tweet for one of three sentiments: "positive," "negative," or "neutral". A tweet's sentiment is categorised into one of the three sentiments classes that has the highest likelihood of occurring.

Based on the distribution of the words in the document, the Naïve Bayes classification model determines the posterior probability of a class. The model employs a bag-of-words feature extraction technique that ignores the word's placement within the text. To determine the likelihood that a given feature set is associated with a specific label, it applies the Bayes Theorem.

9 Design Specification

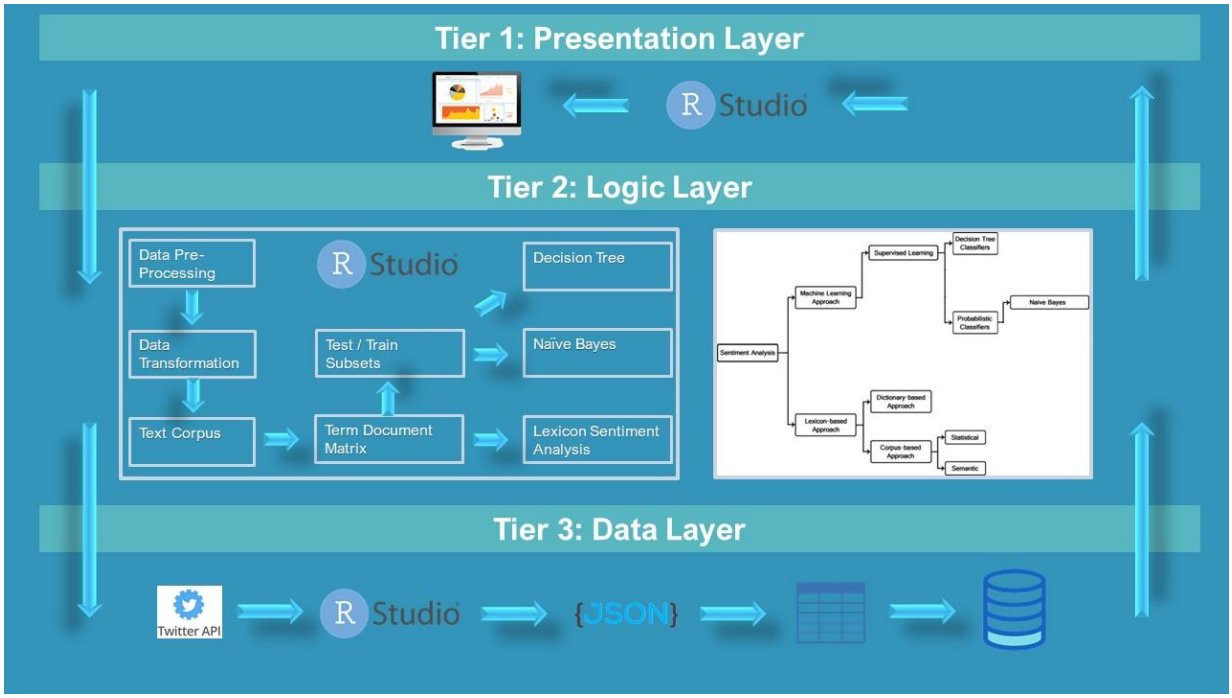

Figure 4. Sentiment Analysis Classification Design Diagram

9.1 Tier 3: Data Layer

The first tier of the design architecture is the Data Layer which contains the connection to the Twitter API, extraction of suitable Tweet data into a dataframe, and storage in a secure relational database.

9.2 Tier 2: Logic Layer

The second tier of the design architecture is concerned with pre-processing procedures are implemented such as discarding irrelevant columns, dropping null vales, creating a text corpus and Term Document Matrix (TDM), removing special characters, excluding stop words, etc., to prepare the data for analysis and modelling. After pre-processing and transformation of data, a variety of sentiment analysis techniques will be employed.

9.3 Tier 1: Presentation Layer

The final layer is to display the results. This will be mainly done via visualisations for the lexicon-based sentiment analyses, such as word frequency and word cloud plots. For the machine learning sentiment analyses models (Naïve Bayes and Decision Tree) confusion matrix outputs, such as accuracy, specificity, and sensitivity will be used, along with plots such as the Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) plots.

9.4 Programming Language & Tools

The R programming language was chosen for the technical aspects of the research and the RStudio Integrated Development Environment (IDE) was used to develop the solution. R has several open source libraires allowing connection to the Twitter API, a full range of sentiment and sematic lexicon libraries, the ability to create text corpora and Term Document Matrices (TDM). R also can implement a wide range of machine learning classification algorithms and R's superior data visualisation outputs also made it suitable for the technical tasks.

9.5 Twitter Data

Data obtained via Twitter APIs returns Tweets encoded using JavaScript Object Notation (JSON). JSON is a free open-source file format and data exchange format that employs language that can be read by humans to store and send data objects made up of arrays and attribute-value pairs. It is a widely used data format for electronic data exchange, notably between servers and online applications. Key-value pairs with named attributes and associated values. Objects are described using these properties and their states.

9.6 Data Extraction via Twitter API

The data acquired for this research was obtained via an authorised Twitter developer account. By connecting to the Twitter API and conducting a search of tweets relating to five airlines who operate from/to Irish airports were chosen to be the source of the data: Ryanair, Aer Lingus, Delta, Emirates, and Lufthansa. The most recently created 1000 tweets per airline were collected (5000 in total) and stored as an object in R. This real-time collection of Tweet data provides relevant data for analysis and modelling.

9.7 Twitter Developer Account & Access Tokens

Twitter allows users to apply for a free Developer Account under their terms of service. Academic users are permitted access on the basis that data is fully anonymised. A Twitter App is created on the platforms website and within this are contained unique access credentials, including *consumer_key*, *consumer_secret*, *access_token*, and *access_secret*.

9.8 twitteR Package

The R package *twitteR* allows for connection access to the Twitter API. Authorisation to the API's functionality is available using this package if the correct Twitter Developer Account access credentials have been set during, as specified in the previous paragraph.

9.9 searchTwitter Function

Hashtags (#) are used in Twitter to identify relevant Tweets and can be searched and extracted using several hashtags at a time. The number (n) of Tweets can also be set, in this case it was 1000 tweets for each of the seven airlines selected, Ryanair, Aer Lingus, Delta, United Airlines, British Airways, Emirates, and Lufthansa. A total of 5000 Tweets were used during analysis at any one time. The Tweets are extracted in real-time, so the data is dynamic as the selection process selects the most recent Tweets.

10 Data Pre-Processing

10.1 Dataset Description

The extracted Twitter JSON data was converted to a suitable format for storage in a dataframe consisting of 5000 rows and 50 columns. Most columns of the data were not required for sentiment analysis as it was only the main text of the Tweet which was required for analysis and modelling. Of the 50 columns 46 were removed to anonymise the tweets and to remove unwanted data. The reduced dataset is summarised in Table 4 below.

Table 4. Summary of data used during research.

Attribute	Description
user_id	Unique Tweet Identifier
created_at	Date Tweet Posted
screen_name	Airline Name
Text	Textual Content of Tweet

An important first step when working with text data is pre-processing to ensure the data is ready for analysis. Text data collected from Twitter is relatively unstructured and noisy an requires several pre-processing steps. Spelling mistakes, incorrect grammar, use of slang, presence of unwanted symbols and stop words are commonly contained within Tweets.

10.2 Text Stemming

The Snowball framework is one of the most widely used stemming tools for the English language in R. This helps to streamline text documents during the pre-processing stage by reducing word count. Plural forms and word derivatives, for instance, can be combined into a single phrase. An example is the words “accounts", "account", "accounting", and misspellings such as "acounting", "acount", "acounts", "accounnt". Each of these words can be stemmed into a single word “account”.

10.3 Removing Stop Words

Before processing natural language textual data, stop words are removed from a stop list. There is no single universal list of stop words used by all NLP tools, nor any agreed upon

rules for identifying stop words. It is possible to select a group of words that will serve as stop words for a particular purpose. Generally, words which do not impact or convey an emotional state are removed. Words such as "at", "is", "the", "in", "for", and "to" are examples. Some custom stopwords were also removed, such as spelling mistakes and abbreviations which are not present in the lexicon dictionaries and as such will not be labelled.

10.4 Tokenizing Words

Tokenization is a process of dividing raw text sentences into words and assigning these words as tokens. The tokens help in understanding the context or developing NLP models. The tokenization helps in interpreting the meaning of the text by analysing the sequence of the words. Once text data has been structured via tokenisation it can be interpreted and used for mathematical analysis.

10.5 Word Frequency Distribution

By counting how frequently particular words appear, term (or word) frequency analyses the significance of terms in a text or collection of texts. This covers percentages and raw and relative frequency counts.

10.6 Visualising the Data

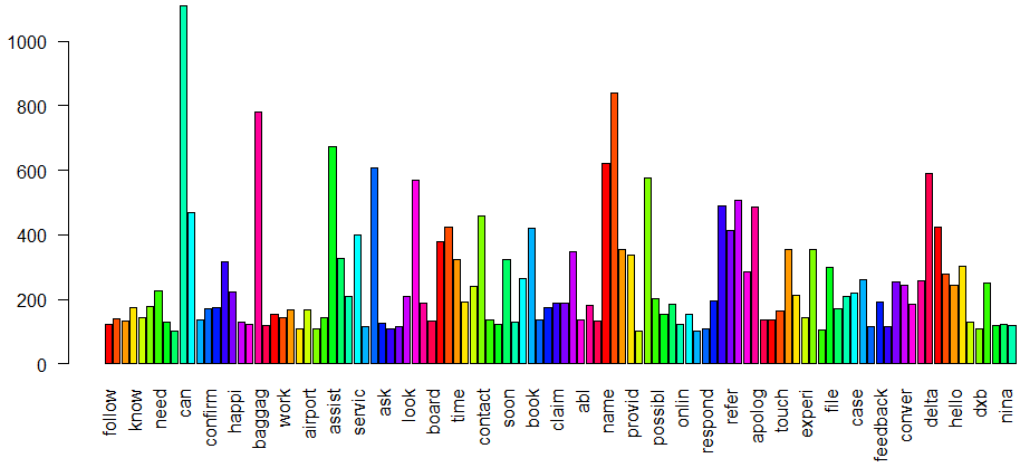

Figure 5. Barplot of Word Frequency.

Figure 6. Wordcloud of Word Frequency.

10.7 Document-Term Matrix

The document-term matrix (DTM) is one of the most typical structures used by text mining software. Each row represents one tweet, each column represents one phrase, and each value reflects the number of times that term appears in that document. DTMs are commonly implemented as sparse matrices and these objects can be used in a similar way as matrices are, such as accessing specific rows and columns, but they are stored in a more efficient format.

10.8 Test & Training Data

In addition to the pre-processing required for lexicon-based sentiment analysis, test and train subsets of the data were created for use with the Decision Tree and Naïve Bayes classifiers. These subsets were used to evaluate predictive capability and goodness of fit of each model. The data was split into 30% testing data and 70% training data subsets.

11 Evaluation

11.1 Evaluating Lexicon-based Models

Lexicon models do not create many statistical outputs upon which to evaluate their accuracy. By their nature, using pre-labelled dictionaries the words contained in Tweets are associated to the existing labels contained within each dictionary. However, using lexicon-based approaches valuable visual outputs can be obtained. Figures 7, 8, and 9 show the different outputs obtained using each of the four lexicons.

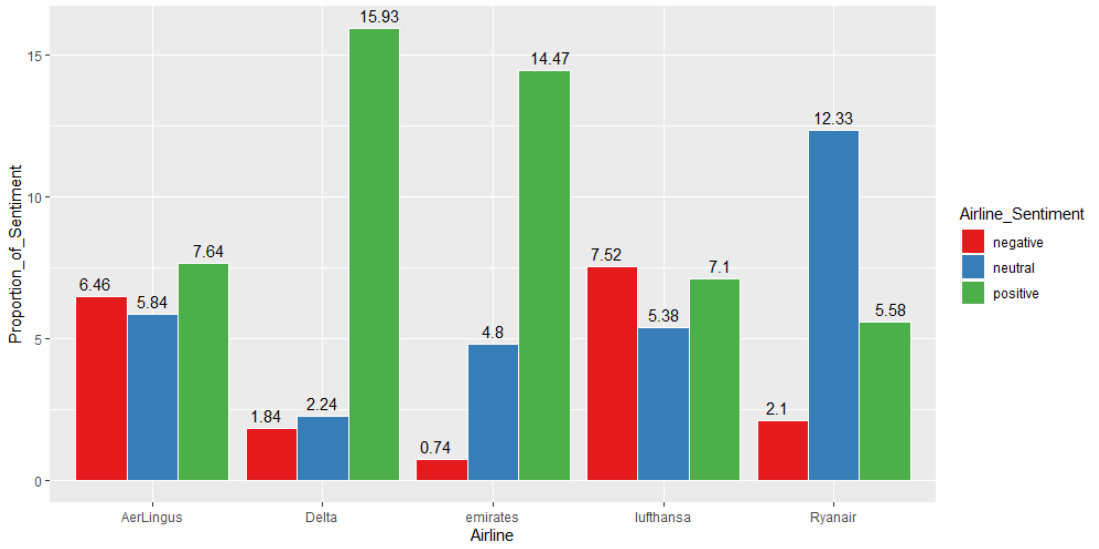

Figure 7. Barplot of Tweet Sentiments using AFINN Lexicon

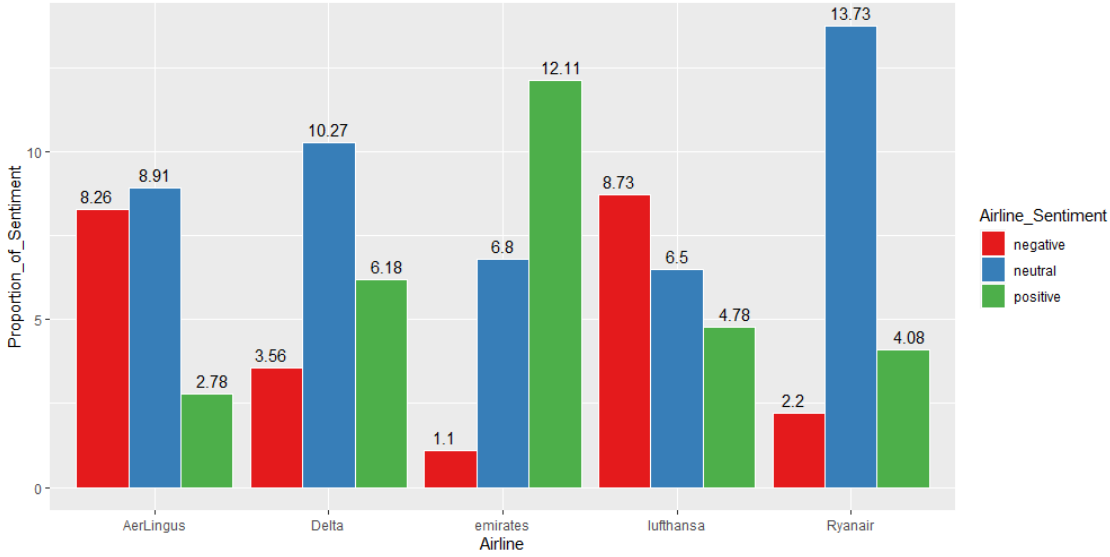

Figure 8. Barplot of Tweet Sentiments using Bing Lexicon

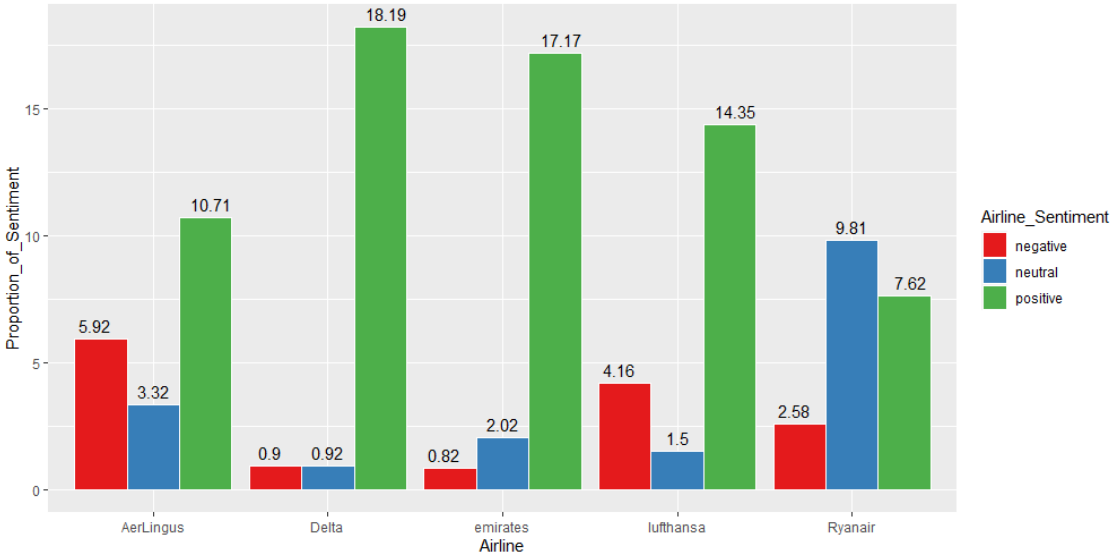

Figure 9. Barplot of Tweet Sentiments using Syuzhet Lexicon

11.2 Differences in Results Between Lexicons

Appending the sentiment scores from all four lexicons it was possible to compare the results to see what percentage of words were classified equally as positive, negative, or neutral using the first three lexicon dictionaries. The fourth lexicon uses emotional labelling and as such was not included in the comparison. The results show that just under 52% were labelled the same, as per Table 5 below.

Table 5. Percentage of Sentiments Matching Across AFINN, Bing & Syuzhet Lexicons

False	True
51.68%	48.32%

11.3 NRC Lexicon

The NRC Lexicon outputs are slightly different to the previous three described above as this method applies the emotions continued in Tweets rather than classifying them as positive, negative, or neutral. Figures 10 below show the most common emotions contained in the Tweets. From the outputs, the overwhelming positive emotional sentiment contained in most Tweets was “trust”. The most common negative emotional sentiment was “sadness”.

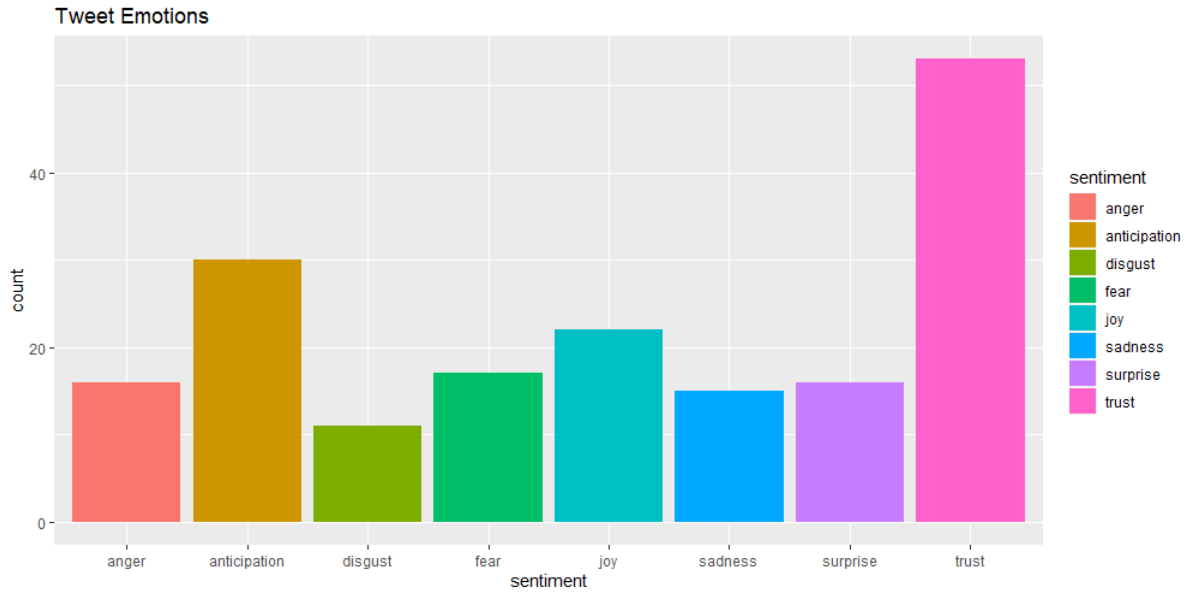

Figure 10. Barplot of Tweet Emotions using NRC Lexicon.

12 Evaluation Metrics

12.1 Accuracy

Accuracy is the measure of identifying how well the model performs based on the training data used. It is the metric used to indicate which model is best at finding relationships and patterns between variables in a dataset. It can be expressed as:

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

Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

12.2 Precision

Precision is simply the amount of positively classified sentiments out of the total amount of sentiments. It is calculated by the number of true positives divided by the number of true positives plus the number of false positives and can be expressed as:

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

12.3 Recall

Recall is the number of true positives to all other true positives plus all the false negatives and can be expressed as:

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

12.4 F1-Score

The F1-Score is the harmonic mean of precision and recall. This makes it possible to evaluate a model and describe a model's performance while accounting for both precision and recall in a single score. It can be express as:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

13 Evaluating The Machine Learning Models

Table 6. Evaluation Metrics Summary

	Accuracy	Precision	Recall	F-Measure
Decision Tree	73.08%	70.63%	71.79%	70.30%
Naïve Bayes	69.29%	79.05%	74.21%	75.55%

13.1 Decision Tree

The Decision Tree model was built to classify three levels of response variable, positive, negative, and neutral. To evaluate the model a confusion matrix was used. The model’s overall accuracy was 70.12%. A Sensitivity (True Positive rate) of 59.05% and the Specificity (True Negative rate) of 74.21% was archived. Overall, this model can be said to

have a moderate predictive capability. A visual representation of how well the model classified sentiment can be seen in Table 5 below and the evaluation metrics contained in Table 6 show

Table 5. Number of Tweets Classified by Decision Tree.

	Negative	Neutral	Positive
Negative	421	241	271
Neutral	63	1223	243
Positive	69	606	1860

13.2 Naïve Bayes

A Naïve Bayes model was built and employed tenfold cross validation and the resultant model accuracy of 69.29% was archived, showing moderate sentiment classification accuracy. Accuracy was calculated on the proportion of correct number of classifications out of the total number of predictions correctly classified, i.e., were the Tweets positively or negatively classified correctly.

Tenfold cross validation was used to better estimate the out-of-sample prediction error and minimise the loss function of the prediction. Using this technique, the data is divided into 10 pieces at random. Nine of those parts for training and one for testing. Repeat the process ten times, saving a different tenth for testing each time.

A Sensitivity (True Positive rate) of 59.05% and the Specificity (True Negative rate) of 74.21% was archived. Overall, this model can be said to have a moderate predictive capability. A visual representation of how well the model classified sentiment can be seen in Figure 12 below.

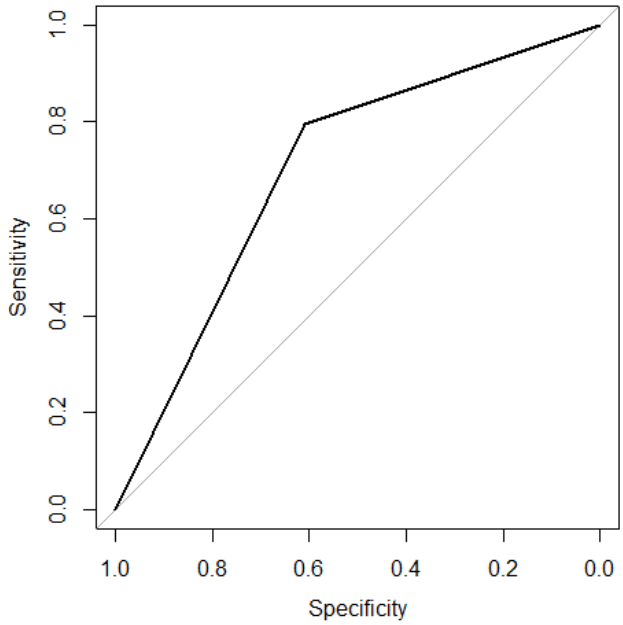

Figure 12. Receiver Operating Characteristic Curve Plot

Cohen's Kappa was reported as 32.22%. Kappa measures observed accuracy against expected accuracy if the predictions were left to random chance. The relatively week Kappa shows a moderate agreement that the model is better at predicting Tweet sentiment when

compared to leaving the prediction to random chance. A limitation of Naïve Bayes is that the model creates predictions based on a lack of relationship (independence) between the predictor variables, so it lacks the ability to account for variables which may have some relationship. However, as can be seen in Table 6 above, the overall results of both classification models show moderate capabilities across all evaluation metrics.

14 Conclusions

The purpose of this research was to implement several sentiment analyses to determine the underlying emotional responses related to consumer perceptions of airlines contained in Tweets from the social media platform Twitter. The study proposes classification techniques which classify Twitter textual data using supervised machine learning methods and conventional lexical-based methods.

The results show, as a proof of concept, that valuable information can be obtained from freely available data which can enable airlines to monitor customer views and determine service criteria which produce positive and negative sentiments. The research has demonstrated that airlines have a low-cost method of monitoring customer sentiment towards their own services and that of their competitors and that moderately strong classifications of sentiments can be obtained. Obtaining data from social media platforms, such as Twitter, and applying a range of sentiment analyses customer attitudes to their experiences may be evaluated. Airlines can be informed of shifts in consumer perceptions of their services and those of their rivals. Negative tweets may harm a company's brand, therefore monitoring and responding to them is useful from a business perspective.

15 Discussion & Future Work

While value can be extracted from extracting data from Twitter, the limited number of Tweets available to extract via the Twitter API limits the size of historical data which can be analysed. Furthermore, while being able to determine emotional responses towards airline services is valuable, further research into the causes of those emotions is recommended. This would enable airlines to fully address the root causes of both positive and negative sentiments amongst customers.

References

Grimes, Seth (1 August 2008). "Unstructured Data and the 80 Percent Rule". Breakthrough Analysis - Bridgepoints. Clarabridge.

Van Doorn, J. et al. (2010) ‘Customer Engagement Behavior: Theoretical Foundations and Research Directions’, Journal of Service Research, 13(3), pp. 253–266. DOI: 10.1177/1094670510375599.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics, [online] 37(2), pp.267–307. DOI:10.1162/coli_a_00049.

Chaovalit, P. and Zhou, L., n.d. Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches. Proceedings of the 38th Annual Hawaii International Conference on System Sciences. DOI: 10.1109/hicss.2005.445.

Lucini, F., Tonetto, L., Fogliatto, F., Anzanello, M. (2020) "Text mining approach to explore dimensions of airline customer satisfaction using online customer reviews", Journal of Air Transport Management, 83, 101760. DOI: 10.1016/j.jairtraman.2019.101760

Yuliya Kolomoyets, "A Text Mining Approach to Measuring and Predicting Perceived Service Quality from Online Chatter", Proceedings of the ENTER2020 Ph.D. Workshop, The International Federation for Information Technology and Travel & Tourism; University of Surrey, School of Hospitality and Tourism Management, vol 2020, pages 97-100.

Mohamed Chiny, Omar Bencharef, Moulay Youssef Hadi, Younes Chihab, "A Client-Centric Evaluation System to Evaluate Guest's Satisfaction on Airbnb Using Machine Learning and NLP", Applied Computational Intelligence and Soft Computing, vol. 2021, Article ID 6675790, 14 pages, 2021. DOI: 10.1155/2021/6675790

Keiningham, T.L., Cooil, B., Aksoy, L., Andreassen, T.W. and Weiner, J. (2007), "The value of different customer satisfaction and loyalty metrics in predicting customer retention, recommendation, and share-of-wallet", Managing Service Quality: An International Journal, Vol. 17 No. 4, pp. 361-384. DOI: 10.1108/09604520710760526.

Sánchez-Franco, M.J., Navarro-García, A. and Rondán-Cataluña, F.J. (2019). A naive Bayes strategy for classifying customer satisfaction: A study based on online reviews of hospitality services. Journal of Business Research, 101, pp.499–506. DOI: 10.1016/j.jbusres.2018.12.051.

Cagatay Catal, Mehmet Nangir, "A sentiment classification model based on multiple classifiers", Applied Soft Computing, Volume 50, 2017, Pages 135-141, ISSN 1568-4946, DOI: 10.1016/j.asoc.2016.11.022.

Syjung Hwang, Jina Kim, Eunil Park, Sang Jib Kwon, "Who will be your next customer: A machine learning approach to customer return visits in airline services", Journal of Business Research, Volume 121, 2020, Pages 121-126,ISSN 0148-2963, DOI: 10.1016/j.jbusres.2020.08.025.

Radojevic, T., Stanisic, N. and Stanic, N. (2017) "Inside the Rating Scores: A Multilevel Analysis of the Factors Influencing Customer Satisfaction in the Hotel Industry", Critique of Anthropology, 58(2), pp. 222–242. DOI: 10.1177/0308275X19842920.

Gautam, G.; Yadav, D., 2014. Sentiment analysis of Twitter data using machine learning approaches and semantic analysis. 2014 Seventh International Conference on Contemporary Computing (IC3). DOI: 10.1109/IC3.2014.6897213

Jain, P.K. et al., 2019. Airline recommendation prediction using customer generated feedback data. 2019 4th International Conference on Information Systems and Computer Networks (ISCON). DOI: 10.1109/ISCON47742.2019.9036251

Saleh F. S. Alkhatib & Yazan K. A. Migdadi (2018) “Operational determinants of airline service quality: Worldwide cross-regional analysis”, Quality Management Journal, 25:4, 186-200, DOI: 10.1080/10686967.2018.1515525

Ching-Fu Chen & Ya-Ling Kao (2010) Relationships between process quality, outcome quality, satisfaction, and behavioural intentions for online travel agencies – evidence from Taiwan, The Service Industries Journal, 30:12, 2081-2092, DOI: 10.1080/02642060903191108

Moraes, R., Valiati, J.F. and Gavião Neto, W.P. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Systems with Applications, 40(2), pages 621–633. DOI: 10.1016/j.eswa.2012.07.059

Aymeric Punel, Lama Al Hajj Hassan, Alireza Ermagun, "Variations in airline passenger expectation of service quality across the globe", Tourism Management, Volume 75, 2019, Pages 491-508, ISSN 0261-5177, DOI: 10.1016/j.tourman.2019.06.004.

Lacic, E., Kowald, D. and Lex, E. (2016) “High enough?: Explaining and predicting traveler satisfaction using airline reviews,” in Proceedings of the 27th ACM Conference on Hypertext and Social Media. New York, NY, USA: ACM. DOI: 10.1145/2914586.2914629

Yun Wan and Qigang Gao (2015). An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis. 2015 IEEE International Conference on Data Mining Workshop (ICDMW). DOI: 10.1109/ICDMW.2015.7

Kumar, S., Zymbler, M. (2019) “A machine learning approach to analyze customer satisfaction from airline tweets”, Journal of Big Data, 6(1). DOI: 10.1186/s40537-019-0224-1

Medhat, W., Hassan, A. and Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), pp.1093-1113. DOI: 10.1016/j.asej.2014.04.011

Jockers, M., 2022. Introduction to the Syuzhet Package. [online] Cran.r-project.org. Available at: <https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html> [Accessed 12 August 2022].

Finn Årup Nielsen A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages 718 in CEUR Workshop Proceedings 93-98. 2011 May. http://arxiv.org/abs/1103.2903.

Minqing Hu and Bing Liu, ``Mining and summarizing customer reviews.", Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004), Seattle, Washington, USA, Aug 22-25, 2004.

Saif Mohammad and Peter D Turney, “NRC emotion lexicon”, National Research Council Canada (2013), pp 234. DOI: 10.4224/21270984