**AIRLINE SERVICE QUALITY ASSESSMENT THROUGH SOCIAL MEDIA SENTIMENT ANALYSIS**

**Abstract**

The research is dedicated to the task of sentiment analysis of airline-related tweets using machine learning (ML) and deep learning (DL). The size of the data was first preprocessed in the text cleaning, tokenisation, and vectorisation with TF-IDF representation of ML models and embedding representations of DL models. Several algorithms to train and test were Naive Bayes, Logistic regression, Random Forest and LSTM-based. Accuracy, a confusion matrix, and classification reports were employed in the evaluation of the model's performance. Naive Bayes model had a competitive performance (accuracy 72.85%), especially in making predictions in negative sentiments. Nonetheless, the deep learning models such as LSTM produced higher recall and F1-scores when predicting positive and neutral labels, which allows it to better understand contexts. In the comparative analysis, it was observed that ML models are more computationally efficient and simpler to implement, but the DL models exhibit semantic connections better and therefore, the sentiment classification of complex text data will be enhanced. The analysis indicates the significance of choosing a model taking into account task specificity, the nature of the dataset and the level of computation capabilities. The findings could be appreciated by the industries, which have adopted the use of real-time sentiment analysis in a bid to improve customer experience and the decision-making process.

**Keywords: *Sentiment analysis, airline tweets, machine learning, deep learning, Naive Bayes, LSTM, TF-IDF, classification, accuracy, natural language processing.***

**Acknowledgement**

I would like to thank my supervisor for their immeasurable help, support, and criticism during the performance of this research. It is through their knowledge and encouragement that this work has taken the outlines and standards that it does. I also truly want to thank my family who have been incredibly supportive in providing love, patience and understanding throughout this journey. It has been empowering and encouraging because of their encouragement and belief in my abilities at all times. This would not have been done without their backing, and I owe a lot to them for being in my life.

**TABLE OF CONTENTS**

[Chapter 1: Introduction 7](#_Toc205895982)

[1.1 Background 7](#_Toc205895983)

[1.1.1 Development of Airline Evaluation of Service Quality 7](#_Toc205895984)

[1.1.2 Emergence of Social Media as a Customer Response Engine 8](#_Toc205895985)

[1.1.3 Quality assessment through Sentiment Analysis 8](#_Toc205895986)

[Figure 1.1: Example of Sentimental Analysis 9](#_Toc205895987)

[1.2 Problem Statement 9](#_Toc205895988)

[Figure 1.2: Active Social Media Users Worldwide 10](#_Toc205895989)

[1.3 Research Aim, Objectives & Questions 11](#_Toc205895990)

[1.3.1 Aim 11](#_Toc205895991)

[1.3.2 Objectives 11](#_Toc205895992)

[1.3.3 Research Questions 11](#_Toc205895993)

[1.4 Expected Outcome 12](#_Toc205895994)

[1.4.1 Sentiment Analysis Framework 12](#_Toc205895995)

[1.4.2 Detection of main Airline Service Quality dimensions 12](#_Toc205895996)

[Chapter 2: Literature Review 13](#_Toc205895997)

[2.1 Introduction 13](#_Toc205895998)

[2.2 Literature Analysis 13](#_Toc205895999)

[2.2.1 Airline Service Quality: Definitions and Dimensions 13](#_Toc205896000)

[2.2.2 Social Media as a Data Source for Customer Feedback 15](#_Toc205896001)

[2.2.3 Sentiment Analysis Techniques in NLP 16](#_Toc205896002)

[Figure 2.1: Convolution Neural Network (CNN) 17](#_Toc205896003)

[2.2.4 Machine Learning Models for Sentiment Classification 18](#_Toc205896004)

[Figure 2.2: BERT Model 19](#_Toc205896005)

[2.2.5 Exploratory Data Analysis (EDA) in Social Media Mining 20](#_Toc205896006)

[Figure 2.3: Sentiment Analysis Experimentation Workflow 22](#_Toc205896007)

[2.3 Literature Gap 22](#_Toc205896008)

[2.4 Summary 23](#_Toc205896009)

[CHAPTER 3: METHODOLOGY 25](#_Toc205896010)

[3.1 Introduction 25](#_Toc205896011)

[3.2 Research Strategy 25](#_Toc205896012)

[3.3 Research Approach 26](#_Toc205896013)

[3.4 Research design 27](#_Toc205896014)

[3.5 Data Collection Method 28](#_Toc205896015)

[3.6 Data analysis 29](#_Toc205896016)

[3.7 Ethical Consideration 30](#_Toc205896017)

[3.8 Summary 32](#_Toc205896018)

[CHAPTER 4: RESULTS AND DISCUSSION 33](#_Toc205896019)

[4.1 Overview 33](#_Toc205896020)

[4.2 Process of design analysis 33](#_Toc205896021)

[4.3 Language and IDE 35](#_Toc205896022)

[4.4 Libraries 35](#_Toc205896023)

[Figure 4.4.1: Libraries 36](#_Toc205896024)

[4.5 Data loading and exploring 37](#_Toc205896025)

[Figure 4.5.1: Tweet data loading 37](#_Toc205896026)

[Figure 4.5.2: Tweets info and shape 38](#_Toc205896027)

[Figure 4.5.3: Descriptive statistics 39](#_Toc205896028)

[4.6 Data cleaning 40](#_Toc205896029)

[Figure 4.6.1: Null values 40](#_Toc205896030)

[Figure 4.6.2: Text cleaning and dropping duplicates 41](#_Toc205896031)

[4.7 Data visualisation 42](#_Toc205896032)

[Figure 4.7.1: Sentiment distribution 42](#_Toc205896033)

[Figure 4.7.2: Airline distribution 43](#_Toc205896034)

[Figure 4.7.3: Positive Wordcloud 43](#_Toc205896035)

[Figure 4.7.4: Negative Wordcloud 44](#_Toc205896036)

[4.8 Data processing 45](#_Toc205896037)

[Figure 4.8.1: Split data and vectorisation 45](#_Toc205896038)

[4.9 Applying the models and evaluation 46](#_Toc205896039)

[Figure 4.9.1: Prediction and evaluation with logistic regression 46](#_Toc205896040)

[Figure 4.9.2: Classification report with logistic regression 47](#_Toc205896041)

[Figure 4.9.3: Prediction and evaluation with Naive Bayes 48](#_Toc205896042)

[Figure 4.9.4: Classification report with Naive Bayes 49](#_Toc205896043)

[Figure 4.9.5: Prediction and evaluation with XGBoost 50](#_Toc205896044)

[Figure 4.9.6: Classification report with XGBoost 51](#_Toc205896045)

[Figure 4.9.7: LSTM layers and architectures 52](#_Toc205896046)

[Figure 4.9.8: Prediction and evaluation with LSTM 52](#_Toc205896047)

[Figure 4.9.9: Classification report with LSTM 53](#_Toc205896048)

[4.10 Discussion 53](#_Toc205896049)

[Figure 4.10.1: Model comparison 54](#_Toc205896050)

[CHAPTER 5: CONCLUSION AND RECOMMENDATIONS 56](#_Toc205896051)

[5.1 Conclusion 56](#_Toc205896052)

[5.2 Linking with Objectives 57](#_Toc205896053)

[5.3 Recommendations 58](#_Toc205896054)

[5.4 Future Work 59](#_Toc205896055)

[5.5 Limitations 60](#_Toc205896056)

[References 62](#_Toc205896057)

[Appendix 71](#_Toc205896058)

**List of Figures**

[Figure 1.1: Example of Sentimental Analysis 8](#_Toc205895956)

[Figure 1.2: Active Social Media Users Worldwide 9](#_Toc205895957)

[Figure 2.1: Convolution Neural Network (CNN) 16](#_Toc205895958)

[Figure 2.2: BERT Model 18](#_Toc205895959)

[Figure 2.3: Sentiment Analysis Experimentation Workflow 21](#_Toc205895960)

[Figure 4.4.1: Libraries 35](#_Toc205895961)

[Figure 4.5.1: Tweet data loading 36](#_Toc205895962)

[Figure 4.5.2: Tweets info and shape 37](#_Toc205895963)

[Figure 4.5.3: Descriptive statistics 38](#_Toc205895964)

[Figure 4.6.1: Null values 39](#_Toc205895965)

[Figure 4.6.2: Text cleaning and dropping duplicates 40](#_Toc205895966)

[Figure 4.7.1: Sentiment distribution 41](#_Toc205895967)

[Figure 4.7.2: Airline distribution 42](#_Toc205895968)

[Figure 4.7.3: Positive Wordcloud 42](#_Toc205895969)

[Figure 4.7.4: Negative Wordcloud 43](#_Toc205895970)

[Figure 4.8.1: Split data and vectorisation 44](#_Toc205895971)

[Figure 4.9.1: Prediction and evaluation with logistic regression 45](#_Toc205895972)

[Figure 4.9.2: Classification report with logistic regression 46](#_Toc205895973)

[Figure 4.9.3: Prediction and evaluation with Naive Bayes 47](#_Toc205895974)

[Figure 4.9.4: Classification report with Naive Bayes 48](#_Toc205895975)

[Figure 4.9.5: Prediction and evaluation with XGBoost 49](#_Toc205895976)

[Figure 4.9.6: Classification report with XGBoost 50](#_Toc205895977)

[Figure 4.9.7: LSTM layers and architectures 51](#_Toc205895978)

[Figure 4.9.8: Prediction and evaluation with LSTM 51](#_Toc205895979)

[Figure 4.9.9: Classification report with LSTM 52](#_Toc205895980)

[Figure 4.10.1: Model comparison 53](#_Toc205895981)

# Chapter 1: Introduction

In the current competitive environment in the aviation industry, the quality of its services has become a key customer satisfaction factor, brand loyalty, and organisational success, in general. As the world is becoming more connected and customers are expecting higher standards of services, airlines have no option but to maintain high levels of service delivery through different touchpoints involved in a flight, which include flight punctuality, in-flight services, baggage handling and customer service. In the past, surveys and formal feedback forms have been the major ways of measuring service quality, but the systems tend to be time-consuming and limited in scope as well as they are not immediate. Conversely, the popularity of social media has changed how customers can show their opinions, enabling airlines to have a rich source of instant feedback. People also publish their experiences with travelling, whether positive or negative, on such social websites as Twitter, Facebook, and Instagram, which offer airlines the chance to access customer feedback in large numbers, although unsolicited. Sentiment analysis, which is a type of natural language processing, has grown as a powerful tool of analysis in order to make use of this information. It allows sorting and analysing the user’s feelings and opinions on the internet. This study tends to examine how sentiment analysis may be used on social media information to measure the service quality of airline companies in a more dynamic and scalable way. The paper follows a data-driven technical approach where the machine learning methods contribute to analysing the trend of sentiments and provide accompanying actions as a result.

## 1.1 Background

### 1.1.1 Development of Airline Evaluation of Service Quality

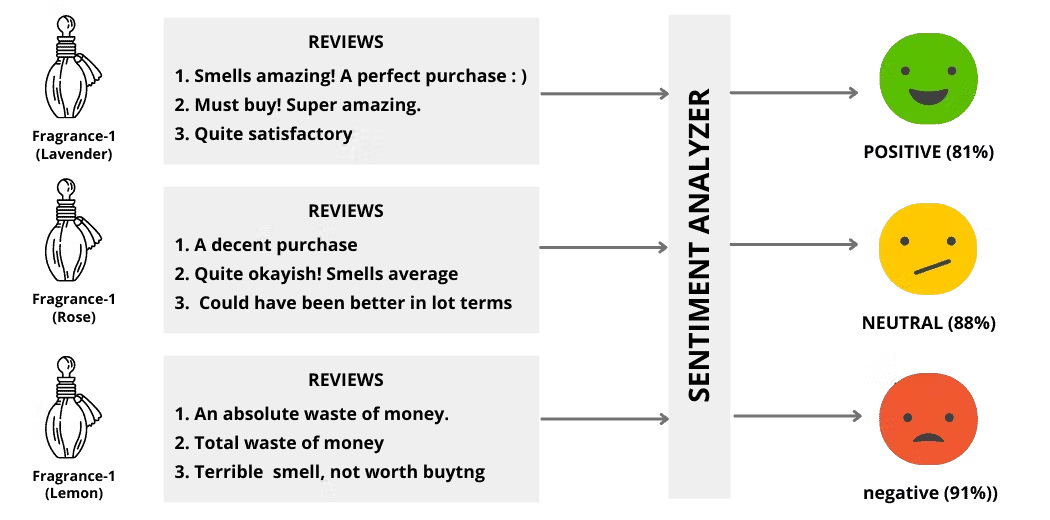
Quality measurement of the airline services has in the past been done using conventional methods, which include structured surveys, focus groups, as well as direct interviews. These devices were meant to get subjective customer views along various service areas such as time, in-flight services, staff attitude, and baggage handling. As an example, the Skytrax World Airline Awards and American Customer Satisfaction Index (ACSI) are global indices that have existed since the 1990s, which measure performance based on data collected through surveys (ACSI, 2025). However, these traditional means are not always useful in determining the pulse of consumer sentiment in real-time mode. Survey-based data collection can prove to be quite time-consuming and costly, particularly when targeting a demographically diverse population with large numbers. Moreover, self-reported information is associated with recall bias, which entails that the respondents do not clearly remember the experiences (Althubaiti, 2016). The urgent necessity of immediacy and increased data representation has driven a shift towards more scalable and near-real-time forms of analysis. As the amount of air traffic across the globe is expected to exceed 8 billion passengers within the next 20 years, the necessity of trustworthy pathways of achieving quality has never been as urgent as it is now (Airports Council International, 2025).

### 1.1.2 Emergence of Social Media as a Customer Response Engine

The recent surge of social media has transformed the nature of customer feedback in the airline business. Social media sites such as Twitter, Facebook, Instagram, and Reddit have become crucial forums where passengers freely write about their experiences. Delta, Emirates, and Lufthansa airlines are among those that get thousands of posts made by users every day, including such posts as the high appreciation of staff hospitality, up to the complaints about possible flight delays or lost luggage. Dean (2023) claims that Twitter alone gets more than 500 million active users daily, with a major part of them containing threads on the services in question. Such platforms not only offer high volume, but also the data can be diverse and comprise multilingual, multimodal, geographically dispersed feedback. This is because these opinions are public and unfiltered, usually, more emotionally loaded and honest than the conventional survey answers. Furthermore, online reviews and comments have a strong influence on the perception of the brands, and a single negative viral tweet could provoke a massive loss of an organisation's reputation (Colmekcioglu et al., 2022). Airlines are also paying more attention to these online discussions in order to determine where improvements can be made, compare themselves to other companies, and even counter crises on a live basis. This abundance of data sets out an interesting possibility of using sophisticated methods of computation to help evaluate the quality of services provided.

### 1.1.3 Quality assessment through Sentiment Analysis

Sentiment analysis, developed based on Natural Language Processing (NLP) and Machine Learning (ML), has become one of the important tools to analyse the customer opinion posted on social media. Such technology will automatically categorise the text into either positive, negative, or neutral and even further identify an emotion such as anger or joy or disappointment, among others. Sentiment analysis has several levels, and they include polarity classification, emotion identification, and aspect-based sentiment analysis, which connect certain comments to the service features like hygienic condition and attitude of staff or check-in process (Alkhnbashi, Mohammad and Hammoudeh, 2024).



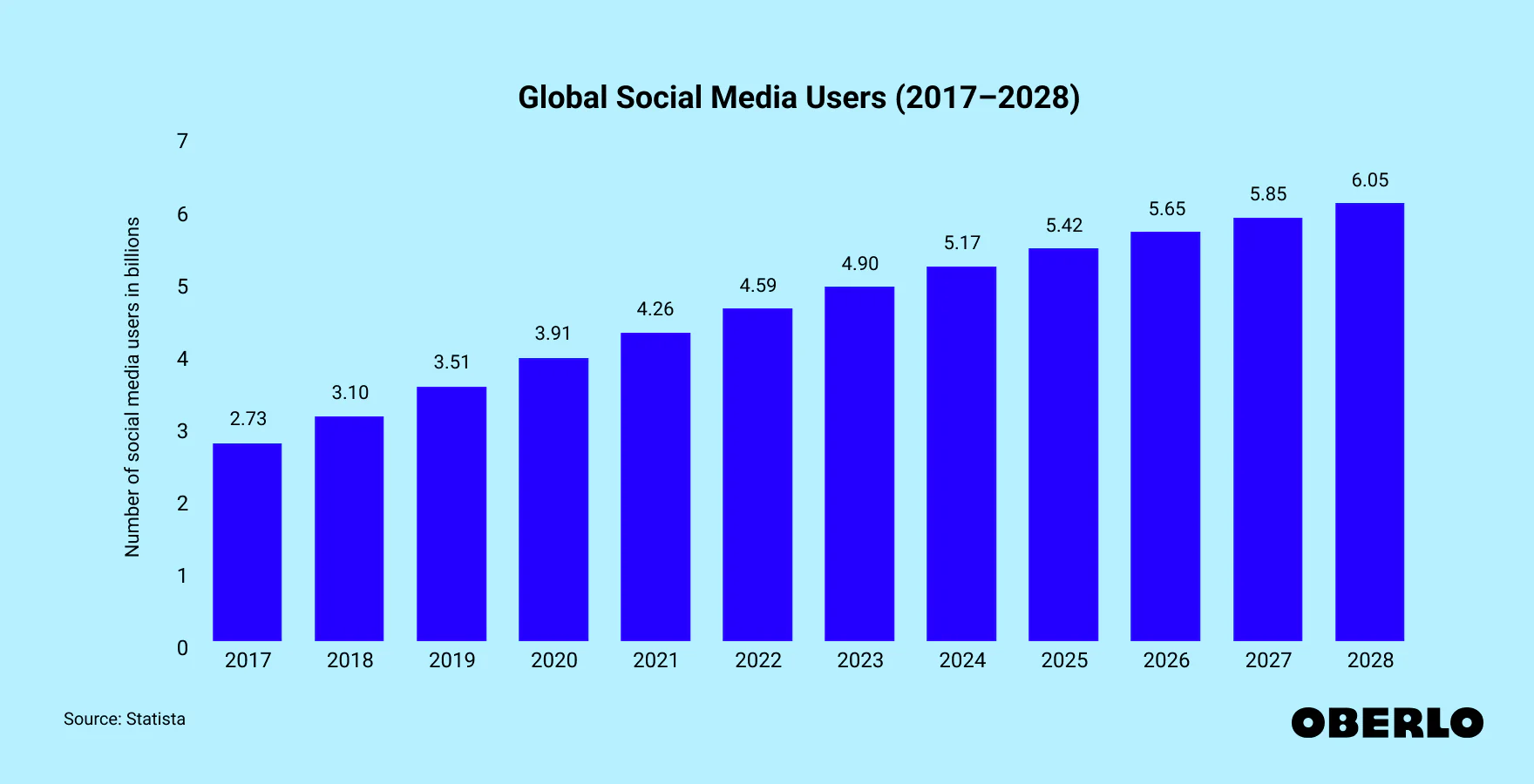
#### Figure 1.1: Example of Sentimental Analysis

(Source: Sajid, 2025)

The sentiment analysis has been applied in other industries such as hospitality and e-commerce, where businesses have been able to move services up or down with the help of user reviews and online ratings (Bellar, Baina and Ballafkih, 2024). Scalability, speed, and objectivity of sentiment analysis are a value proposition hard to reject within the airline industry, where it is essential to get valuable feedback in time. The potential of handling and analysing enormous amounts of data can give airlines viable results at a larger scale, since tens of millions of bytes of data are produced daily around the world (Ajah and Nweke, 2019). Therefore, the implementation of sentiment analysis in the assessment of airline service quality will be more modern and in the spirit of the digital era.

## 1.2 Problem Statement

Social media has been growing exponentially and has created unprecedented levels of customer feedback that are available at the fingertips of the airline industry. These sites to name a few, include Twitter, Facebook and TripAdvisor, and millions of interactions per day occur regarding their experiences with the airlines and provide insight on the quality of service. However, in most cases, airlines do not take complete advantage of this input because they fail to get the full use of their current feedback systems. These systems are largely reactive and depend on delayed customer service reports or surveys, which lack the swiftness and profusion of the online expression. The latest available data shows that there are more than 5.42 billion active social media users across the world, as well as an increased number of travel-related social media posts every month (100 and above) (Sprout Social, 2025). This amount cannot be monitored manually and is not even scalable, bearing in mind words and phrases that are linguistically diverse, sarcastic, abbreviated and have emotional shades that define the online discourse.



#### Figure 1.2: Active Social Media Users Worldwide

(Source: Oberlo, 2021)

Although the artificial intelligence tool is becoming increasingly popular in many industries, there is minimal coverage of the issues related to implementing sentiment analysis in evaluating the quality of airline services. However, the majority of research is conducted either at the level of customer satisfaction or reputation management without breaking down the service in specific dimensions, including check-in procedures, inflight services, on-time arrivals, or baggage services (Zhang, Lee and Gu, 2023). This granularity inhibits airlines from knowing targeted areas where they should make improvements. Further, no standardised models make use of the sentiment data in real time through inclusion in service quality surveillance equipment. With a highly competitive environment pushing airlines toward optimising their business processes and customer satisfaction levels, sentiment analysis is very much a data-driven track towards visionary evaluation of quality. The gap is filled in this research as it offers a customised sentiment analysis model that captures, categorises, and comprehends users’ senses of particular dimensions of service through social media data. The study will help to make strategic decisions based on reforming chaotic online written information into structured insights and carry out the continuous improvement of the exceptional service provided.

## 1.3 Research Aim, Objectives & Questions

### 1.3.1 Aim

This project aims to analyse and compare how major U.S. airlines perform in terms of service by looking at and processing tweets using NLP and machine learning.

### 1.3.2 Objectives

* To make sure and improve Twitter data on U.S. airlines
* To use exploratory data analysis (EDA) on tweets about airlines
* To build and train models that classify what opinions are shown in a review
* To analyse the quality of a classification model with suitable measures.

### 1.3.3 Research Questions

* How can analysing sentiments on Twitter help determine and compare the service quality of the main U.S. airlines?
  + What are the effective ways of collecting, cleaning and preparing Twitter data of U.S airlines to be used in sentiment analysis?
  + How can exploratory data analysis (EDA) recognise the patterns and trends of airline-related tweets?
  + What are the best machine learning models to use when classifying the sentiment of customer tweets regarding airlines?
  + What are the suitable performance measures according to which the sentiment classification models can be assessed?

## 1.4 Expected Outcome

### 1.4.1 Sentiment Analysis Framework

Among the many studies produced by this research project is the creation of a complete framework of sentiment analysis that can be used in testing the quality of service provided by an airline. The framework is anticipated to incorporate methodical steps to be followed to gather social media information, in which Twitter will be the primary source of data, given that its API is open and the popular way of receiving real-time customer interaction. In order to make the data collection process relevant, it should consist of using keywords to filter data, hashtags, and mentions related to a particular airline. Data will then be collected and preprocessed by various tasks of cleaning up the text, tokenising, removing stop words, and normalising them to be ready to be analysed. The essence of the structure will be the incorporation of natural language processing and supervised machine learning algorithms that will decipher the sentiments attached to different service attributes. The classification is created to identify the polarity of the sentiment, positive, negative, or neutral and correlate it with the type of comment made by the customer. It will produce a scalable, replicable model that makes it possible to read large amounts of textual data in a structured, sensible manner.

### 1.4.2 Detection of main Airline Service Quality dimensions

The identification of leading dimensions of service quality as present in customer sentiments is yet another anticipation of this research. The study will be based on examining trends in the results of the sentiment-tagged data to determine the most common subjects that passengers write most about, which include flight punctuality, staff professionalism, in-flight entertainment, cleanliness, check-in experience, and baggage handling. All these elements of the services will be aligned to the polarity of the sentiments to see which of such elements will produce a positive interaction and which elements are linked to dissatisfaction. This will be mapped to give more insight into the particular strengths and weaknesses of the airline services as far as the customer is concerned. The research will also divide sentiment feedback by service stage, pre-flight, in-flight, and post-flight, thereby offering a closer perspective of the customer journey and experience. In the process, the study will show how unorganised social media data can be converted into quantifiable knowledge in terms of dimensions in service quality.

# Chapter 2: Literature Review

## 2.1 Introduction

The literature review in this chapter is a critical reading of prior literature concerning the evaluation of the airline service quality through social media sentiment analysis. Twitter has been identified as one of the major platforms through which unsolicited and real-time feedback on the performance of airlines is sent by customers in the wake of the current growing trend of consumer activity on online platforms. The knowledge on how passengers are able to show they are satisfied or not through tweets provides critical information on different dimensions of service, including punctuality, staff behaviours, and journey experiences during flights. Sentiment analysis as a subfield of Natural Language Processing (NLP) is critical in getting insights into such opinions, where the emotional tone of user-generated content is classified. Sentiment analysis, used together with machine learning algorithms, allows customer perceptions to be evaluated on a wide scale in an accurate manner. In order to contextualise this study in a strong academic background, the chapter supports five thematic areas, such as the service quality model in airline services, the application of social media on feedback collection, sentiment analysis methodologies, machine learning algorithms on classification, and on the exploratory data analysis of social media mining.

## 2.2 Literature Analysis

### 2.2.1 Airline Service Quality: Definitions and Dimensions

Service quality in airlines is one of the main areas of inquiry both in academia and the airport industry because it directly addresses the issue of customer satisfaction and loyalty as well as competitive ability. Service quality is the traditional source of evaluation, which is based on frameworks like SERVQUAL and the American Customer Satisfaction Index (ACSI). The other approach that has determined the five main dimensions of service quality according to this model is the SERVQUAL model suggested by Pakurár et al. (2019), which includes reliability, responsiveness, assurance, empathy and tangibles. These dimensions have found wide availability and adjustment in industries, among them the aviation industry, where they provide a basic vision as to how to evaluate service delivery. The reliability in the airline industry usually means the punctuality of the flights, the responsiveness to the customer inquiry and complaint, the assurance that is delivered by the perceived competency of the staff, the empathy which is manifested by the personal service and the tangibles the condition of the physical maintenance of the planes and entertainment facilities. Nevertheless, researchers have raised concerns over the relevance and adequacy of SERVQUAL in reflecting the distinct characteristics associated with the dynamics of airline quality service. Anand (2024) stated that the SERVQUAL framework was too general and it was ineffective in determining service performance in contexts that involved high expectations and difficulties in service interactions, which defined air travel. They suggested that performance-based models can provide more precise knowledge than a gap-based model, such as SERVQUAL, that is concerned with differences between customer expectations and perceived service.

Other studies have also stressed the significance of making changes that are context-specific. An example of a particular model is that of Sabatová et al. (2016), which focuses on airline service and further considers the extra variables like check-in capability, check-in succession, on-board comforts and luggage handling. Their results showed that timing and employee demeanour were even at times more effective at determining customer satisfaction than the traditional aspects such as tangibles. This refutes previous opinions because it has put forward the thought that, in most cases, intangible experiences, like a sense of being held in high regard or being assured in the event of a hitch, are as significant as the physical environment. Completely different methodologies are uncoupled by a macro-level solution, which is conducted by gathering consumer ratings into a national scale of comparison, the American Customer Satisfaction Index (ACSI). Airline service has continued to be identified among the underperforming businesses in the United States economy, citing long-standing problems concerning uniformity and customer communication (Park and Lee, 2024). However, opponents claim that aggregate ratings using such indices as ACSI can conceal the strength or weakness of individual airlines and fail to reflect the immediate trends in service delivery. In this regard, tweets and reviews, as some examples of user-generated content, give an unbiased dynamic view that existing models can miss.

Moreover, research carried out has emphasised the importance of combining the conventional frameworks of service quality with social media data. According to Ilieva et al. (2024), tweets of customer perceptions tend to resonate with dimensions of SERVQUAL, except that they are in a more emotional and impulsive language. This leaves social media with abundant resources of real-time sentiments and experience-based data. It has been noted that the evidence behind the suggestions that social media feedback demographics are reliable remains questionable as social media contains sensitive information that particularly falls prey to emotional bias, misinformation, and domain-specific styles of language use and thus should not be relied upon as an adequate replacement to the existing survey methods.

### 2.2.2 Social Media as a Data Source for Customer Feedback

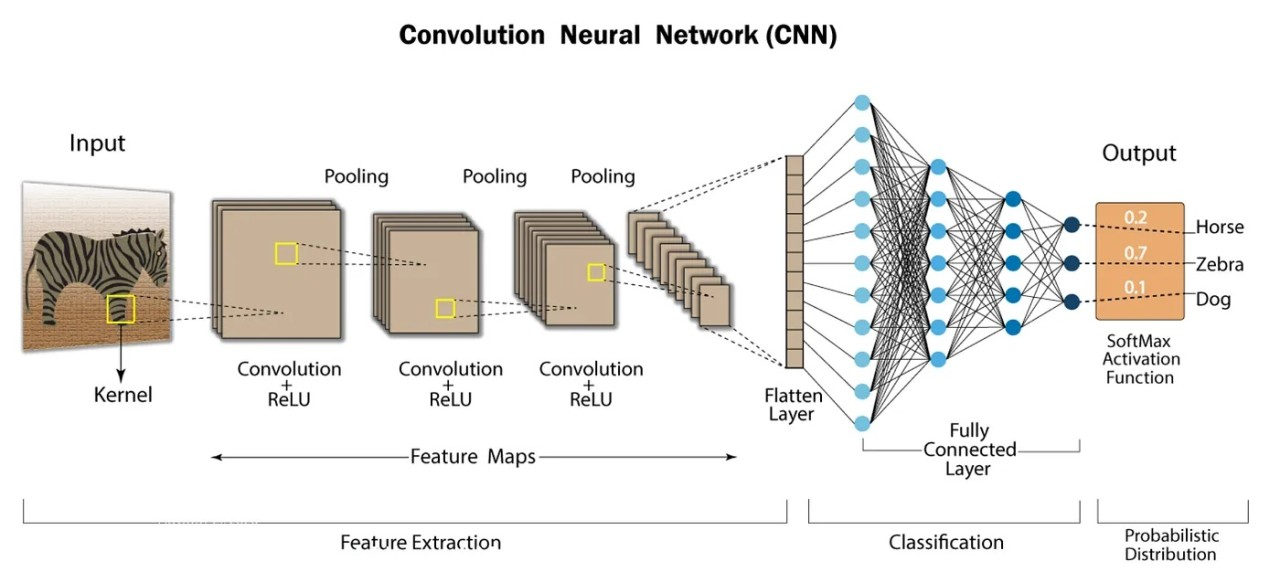
Twitter and other social media are becoming more popular channels where customer feedback can be obtained in service-based firms. Since the introduction of Twitter, an infinite amount of information on the interactions of consumers with the company is being transmitted to provide a real-time representation of customer emotions and service evaluation, a need that has attracted the attention of researchers and practitioners interested in using Twitter as a way to track customer sentiment and service performance. The format of microblogging that supports concise and explicit posts usually makes users express their very first thoughts, both contrasting and welcoming any experience with brands, not to mention airlines. Dwivedi et al. (2021) describe the openness and immediacy of social media as organisations engage with consumers to get an immaculate access to unreserved views. Twitter has been adopted by researchers since it is a scalable, unsolicited and naturally occurring data source. As an example, Badanik, Remenysegova and Kazda (2023) have proved that the sentiment analysis of Twitter can be used to capture aversions about airlines expressed through customers, and hence organisations can get information without issuing surveys that are very expensive and time-consuming. Dublino (2023) pointed out that Twitter is real-time, which enables businesses to react quickly to service failures or public relations emergencies, becoming relevant to improve brands and connect to customers. The use of social media has also proved especially harmful since viral readings and trending hashtags have a propensity to send compliments and condemnations around the globe in a matter of hours.

In spite of these capabilities, a number of researchers have questioned the validity of Twitter information in conducting a serious analysis of service quality. The existence of noise is one of the fundamental problems, i.e. irrelevant or off-topic content, which has the capacity of distorting any output of analysis. Contextual comprehension is also affected by the fact that the platform has a character restriction of 280; thus, the user tends to shorten or condense ideas, leaving out vital information (Boot et al., 2019). In addition, spam accounts and bulk access bots can spam hashtags or brand mentions, thus causing difficulties in extracting accurate data and cleaning the data. There was also the unbalance of sentiment data. Tweets may also be more biased since only customers who have an extreme satisfaction or dissatisfaction with a product will tweet, resulting in an imbalanced data set of polarised views (Rodríguez-González et al., 2020). Consequently, the reliability of the generalizability of the findings may be in question, as the experiences of the average customers may not be reflected appropriately at all times through feedback gathered on Twitter. To meet such challenges, other scholars have proposed that the analysis of social media be combined with conventional methods of data collection. As an example, Cheong et al. (2023) have stated that despite the rich qualitative information that social media offers, it is likely to be more effective as an addition to the techniques and methods of survey and structured interviews rather than an alternative to them in terms of demographic targeting and statistical control.

Even in the special case of the airlines and hospitality industries, a number of studies have displayed the feasibility of the use of Twitter-based analytics. Martin (2014) achieved that by monitoring Twitter to see the complaints of customers against airline delays and cancellations, whereas the Han, Mingying and Peng (2025) study has investigated social media sentiments to evaluate service gaps within hotel management. Though these studies validate the usefulness of social media as a source of feedback, they also indicate that strong data preprocessing, sentiment modulation and context-specific models are still necessary to validate and provide reliable results.

### 2.2.3 Sentiment Analysis Techniques in NLP

Sentiment analysis or opinion mining constitutes one of the fundamental activities of natural language processing (NLP), which turns out to be the process of identifying the emotion of any textual information. Over the past years, it has acquired significant coverage, especially in the context of user-generated content on contemporary social networking sites. Sentiment analysis can be used in 3 major ways, namely, lexicon-based, machine learning-based, and deep learning-based. Lexicon-based techniques depend on existing vocabularies of words marked with a polarity of emotions. These methods are simple to apply and are interpretable, yet fail to perform well concerning context sensitivity, sarcasm, and changing language rules on such social networking sites as Twitter. Li and Zou (2024) suggest that lexicon-based methods may be good at working with structured and formal texts, but cannot be applied easily to informal, short and noisy texts like tweets because they lack sensitivity. Proposals using machine learning, on the other hand, utilise the machine learning algorithms, such as Naive Bayes, Support Vector Machines (SVM), or even Logistic Regression through training on labelled datasets to predict sentiments. Using statistics, these models acquire patterns and usually better lexicon-based models on accuracy, particularly when tuned and validated well. Choi and Lee (2017) indicated the performance of SVM and Naive Bayes in sentiment classification and their flexibility in the application in other fields. Nevertheless, machine learning algorithms need large quantities of annotated training information, and even in general, they perform poorly on domain-specific language and changing slang, as is widespread in social media text.



#### Figure 2.1: Convolution Neural Network (CNN)

(Source: LinkedIn, 2023)

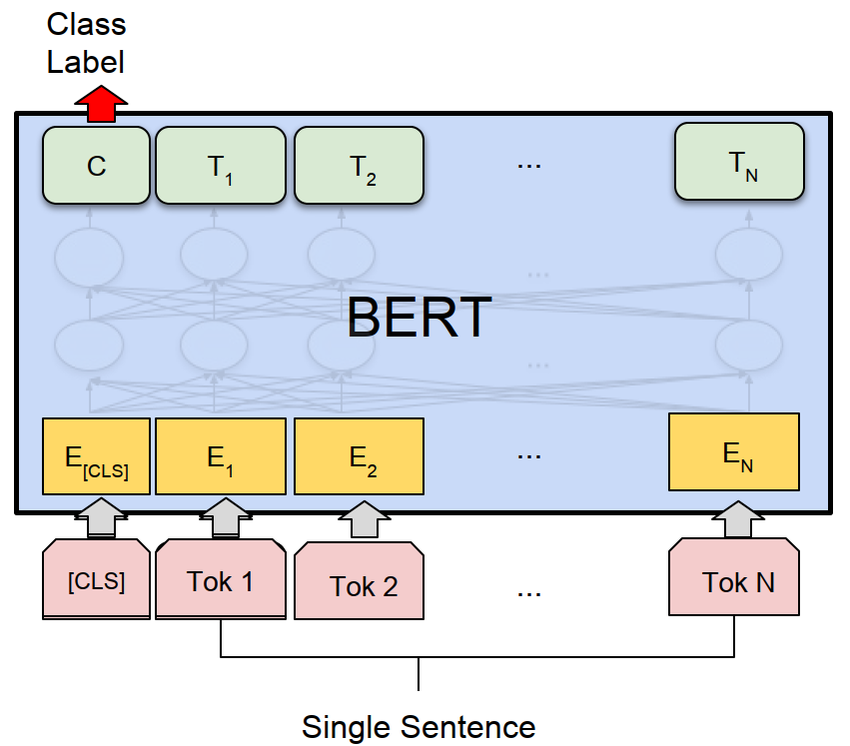
Recently, a potent alternative has been proposed, introducing deep learning models which can extract non-linear linguistic patterns and semantic frameworks. Such architectures as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Transformer-based models (BERT) have demonstrated prominent results in several sentiment analysis applications. As an example, Talaat (2023) revealed that BERT substantially outdid conventional models across a variety of NLP benchmarks such as sentiment classification. The trade-off, however, is in their computational complexity, opaqueness and requirements of extensive training, data, as well as processing power. Critics would assert that the interpretability of deep learning models is a significant hindering factor towards their use in situations that require sensitive or regulated data, like in the case of airline customer services, where any actionable information that should apply must be transparent and translatable. Basic NLP methods tokenisation, part-of-speech (POS) tagging, stemming, and lemmatisation, are crucial in all three of the methods. These preprocessing procedures aid in streamlining written content as well as standardising data, which is better analyzable through algorithms. Although the techniques used result in model improvement, they may have varying results when applying them within languages styles of tweets. As an example, inappropriate words, emojis, and abbreviations may be problematic to POS taggers and tokenisers, which Narayanasamy et al. (2022) also noticed when classifying sentiments based on Twitter data, as their preprocessing decisions had substantial effects on the results.

Sentiment analysis has taken a more detailed approach, though, aspect-based sentiment analysis, where there is a connection between the sentiments and certain portions, or attributes of a service. Within the airline business, ABSA can associate sentiments with facilities like check-in or legroom or even the bag handling. This was highlighted by D’Aniello, Gaeta and La Rocca (2022) on the fact that ABSA provided more detailed information than classification of the overall level of polarity in services evaluation. Yet another extension of sentiment analysis is known as emotion detection, where the analysis goes beyond polarity by estimating the emotional state, including anger or happiness, etc. Although this may serve to build a better understanding of customers, emotion classifiers are less precise than polarity classifiers and may need finer-grained training data.

### 2.2.4 Machine Learning Models for Sentiment Classification

The sentiment classification area in machine learning has formed the core of social media analytics, particularly in the short-text mediums, like Twitter. Naive Bayes, Logistic Regression, Support Vector Machines (SVM), Random Forest, and XGBoost are the most popular ones. They would be employed since they are efficient, readable and they do not have much computing demands. Naive Bayes is an instance of feature independence assumptions succeeded in heaven and earth in text task classification, where performance is secondary to simplicity and speed of operation. Kaggle (2020) presented the following summarised version of approaches, including Naive Bayes that is effective in Twitter sentiment analysis, along with respective preprocessing and unigrams. However due to assumption of independence with its features, the results that might be reflected in reporting the contextual sentiment may not be so accurate. Logistic Regression and SVM are widely used in sentiment, where the binary classification task is required, e.g. positive sentiment or negative sentiment. According to Han et al. (2020), SVM tends to be better in the task of sentiment regarding text since it is highly tolerant of high-dimensional data when compared to other classical models. However, SVMs have the vulnerability of kernel selection and its relative complexity must be tuned more than the simpler SVMs. In comparison, ensembles of learning on decision trees, which include Random Forest classifiers, will be a safe bet and are not likely to overfit. They struggle, however, with sparse data representations, as is typical of text analysis, and in most cases lack the speed to track sentiments in real-time.

The success of XGBoost has been necessitated by its excellent results in competitions of structured data and the real world. It has a gradient boosting algorithm that enables it to obtain high accuracy with minimal tuning. According to Imani, Beikmohammadi and Arabnia (2025), the scalability and effectiveness of XGBoost have been noted on numerous text mining tasks. Nevertheless, it is more difficult to apply and decipher than such models as the Logistic Regression and thus, less preferable to use it in case of high transparency demands. In addition to the classical models, the deep learning techniques have been incorporated and, to a large extent, embraced to address the intricacies of language when classifying sentiments. Long short-term memory (LSTM) networks are a subset of so-called Recurrent Neural Networks (RNN), which can be trained to learn sequential dependencies in text, a new design useful to follow downstream sentiment analyses. LSTM model scales can outperform the traditional classifiers in most of the sentiment-related tasks, more so on context and nuance-delicate expressions in addition to negations Atandoh et al. (2024). Although such models are accurate, they are cumbersome and need a big dataset to stand at their best.



#### Figure 2.2: BERT Model

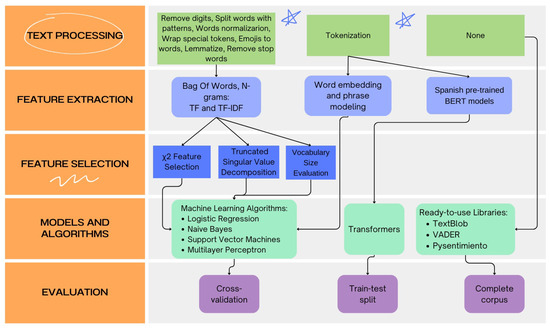
(Source: Rana et al., 2025)

Translators like BERT (Bidirectional Encoder Representations through Transformers) have redefined the limits of NLP-related operations, and sentiment classification is not an exception. According to the results presented by Rana et al. (2025), compared to its predecessors, BERT consistently outperformed previous models in capturing bidirectional context, which is important in sentence detection of sentiments in short and ambiguous sentences. However, the performance achieved by BERT is at the cost of high resources and the complexity of training, which implies that it is not so accessible to smaller applications or real-time applications. The critics point out that, despite being potent, deep learning models are opaque and therefore can only be applied where interpretability of the models is required. Training and validation strategies are important with regard to the reliability of the model in the Twitter sentiment analysis. Stratified k-fold cross-validation is normally applied to solve the problem of imbalance of classes in the dataset, whereas methods like SMOTE are used to artificially over-sample minor classes. The models are usually analysed in terms of their performance, in terms of accuracy, precision, recall, F1-score, confusion matrix and ROC-AUC. However, even though the most common measure of accuracy is reported, it is deceptive when dealing with imbalanced datasets. Thus, F1-score and AUC are used to gain a more balanced impression of the model’s performance (Hassan, 2024).

### 2.2.5 Exploratory Data Analysis (EDA) in Social Media Mining

Exploratory Data Analysis (EDA) is another significant part of social media mining, especially in the case of unstructured and informal data, such as tweets. EDA can help in gaining an initial insight into the data related to sentiment analysis regarding the quality of the services offered by the airline because it reveals the patterns, identifies anomalies, tests and formulates hypotheses. The possibility to visualise and summarise the data of tweets in regards to the volume, the tendency of time, and frequency of keywords enables researchers to make initial conclusions according to which innovative analytical steps are to be made. According to IBM (2021), EDA is necessary to ensure the ego can hear the voice of the data, or in other realms of application where a data structure is not applied beforehand (social media content). One of the main uses of EDA in Twitter sentiment awareness is understanding the variation in tweet volume around a particular period, which can be attributed to certain events such as flight delay or service breakdown or even promotional programs. As an example, one of the recent studies by de Oliveira Júnior et al. (2020) emphasised how temporal sentiment patterns could be used to identify in real time spikes in the opinion of the population related to the occurrence of operational blips or virus customer experience. Timelines and heatmap visualisations allow representations of such changes, providing airlines with a history of how people feel about significant events. Nevertheless, other scholars feel that the dependence on visual trends could be a simplistic way of analysing the depth of emotional expression, and it does not differentiate between sarcasm, irony and actual sentiment, whose use is common in social media conversations.

A widespread EDA tool is word clouds and bar charts used to plot terms/hashtags frequencies in the data. Such visualisations may assist in finding out the central conversations and issues of customers. Schwartz et al. (2022) claims that, following the lack of contextual and semantic thickness, word clouds help obtain an intuitive overview but are likely to cause misunderstanding without precaution. In this manner, visualisations that rely on frequencies can and should be complemented after further investigation with more powerful text mining approaches, including topic modelling or clustering. The preliminary stage before EDA of tweet data sets is preprocessing because texts in social media are highly irregular and noisy. In the tweets, there are hashtags, mentions of users, emojis, misspellings, abbreviations, and URLs, which may be inconvenient to interpret correctly. Popular ways to organise the data in a shape usable for further research include tokenisation, normalisation of the cases, stop-word removal and stemming. Conventional preprocessing pipelines have been criticised, though, by other scholars such as Gomez-Adorno et al. (2024) who claim that they remove too much information, including cues of sentiment that are found within emojis or stylised words which are especially expressive in platforms such as Twitter. This brings up the issue of whether to be able to keep the data clean enough versus keeping the context and the emotional information.



#### Figure 2.3: Sentiment Analysis Experimentation Workflow

(Source: Gomez-Adorno et al., 2024)

It is also applicable to define outliers and oddities in sentiment distributions with EDA. In the case of an airline, this may include detection of short-lived negative sentiment due to service outages, weather-related disruptions or regulatory scandals. As it was shown by Kuoppakangas et al. (2019), EDA may reveal such unusual findings, allowing organisations to identify the sources of public dissatisfaction. An assessment of these patterns should, however, be made with caution since a sentiment trend can be caused by external factors not related to service performance, like political events or social movements.

## 2.3 Literature Gap

|  |  |  |
| --- | --- | --- |
| **Area of Focus** | **Key Findings from Literature** | **Identified Gap** |
| ***Airline Service Quality Frameworks*** | Models like SERVQUAL and ACSI provide structured evaluation dimensions such as reliability, assurance, and empathy. | Limited applicability to real-time, unstructured social media data; lack of integration with spontaneous passenger feedback. |
| ***Social Media as a Data Source*** | Twitter offers real-time, large-scale, and unsolicited feedback, widely used in airline and hospitality studies. | High noise levels, sarcasm, data imbalance, and lack of standardised interpretation reduce reliability for service evaluation. |
| ***Sentiment Analysis Techniques (NLP)*** | Lexicon, machine learning, and deep learning models are extensively used; BERT and LSTM show strong performance. | Insufficient attention to sentiment misclassification in informal, emotion-heavy platforms like Twitter. |
| ***Machine Learning for Sentiment Classification*** | Traditional and advanced classifiers perform well on labelled data using standard metrics like accuracy and F1-score. | Few studies focus on model transparency, explainability, and performance under domain-specific constraints like the airline service context. |
| ***Exploratory Data Analysis (EDA) in Social Media*** | EDA is used for visualising trends and identifying patterns in tweet sentiment. | Underexplored as a diagnostic tool to segment airline-specific sentiment dimensions, such as baggage, check-in, or in-flight service, using granular textual indicators. |

**Table 2.1: Literature Gap**

(Source: Created by Author)

## 2.4 Summary

In the chapter of literature review, summarised in five main themes of the airline service quality frameworks, the utilisation of Twitter as a real-time customer feedback channel, sentiment analysis methods, machine learning models to classify, and exploratory data mining in social media mining, the available literature was surveyed. Although each of the domains has been studied separately, the combination of these domains to determine the quality of service provision in the administration of airline companies in the United States has not been applied significantly. Gaps found, including the lack of use of aspect-level sentiment and the model’s lack of explainability, indicate that a customizable and data-preferred process is necessary. These are the reasons that lead to these insights, which, according to the next chapter, justify the choice of the given methodology in this research, as it allows covering the research objectives.

# CHAPTER 3: METHODOLOGY

## 3.1 Introduction

The approach in the methodology section explains how a systematic methodological approach was used to evaluate the service provided by airline companies based on sentimental analysis of social networks, especially Twitter. In the world of the digital age, clients often go online to post information about their experience and views, and social media is a goldmine of real-time feedback. This study utilises this data to assess the attitude of people towards the U.S carriers' services, which provide the information that can be used to shape service improvement policies.

The methodology section yields a clear guide of how the study is to be conducted, including the rationale of the strategies used and how data is going to be collected, processed and analysed. It will start by explaining the research strategy, then the research approach and design. The collection of data is then clarified in the chapter, including the use of a publicly available Twitter dataset of sentiments (Lu *et al.* 2023). It also explains the methods of analysis that it uses, Natural Language Processing (NLP), as well as supervised machine learning models and a deep learning model to improve sentiment analysis. And lastly, we bring up ethical considerations and see that the study is in accordance with laws and professional rules.

## 3.2 Research Strategy

The type of research strategy that this study has chosen is that of descriptive research, which is intended to capture a logical and coherent appreciation of the customer attitude towards airline services in view of the social media interactions. Descriptive research is inclined towards description of the properties of some specific phenomenon without attempting to transform or to change the environment. The phenomenon in this study is the perception and sentiment of customers towards the quality of airline services as posted via tweets (Xie *et al.* 2024). It is not aimed at testing any hypothesis or creating a causal relationship but to be able to classify, analyse, and interpret the patterns in the feelings of passengers in the airline based on their social media posts.

The current approach is especially appropriate to conduct sentiment analysis because the original aim is not to change variables or to introduce experimental conditions but rather to define and specify the opinions (whether positive, negative, or neutral). Descriptive research will help to know what can be found said about airlines, how often some statements are repeated, and where such or such services of the airlines are most often praised or disliked. Also, it helps to reveal the trends, e.g. sudden decrease of the negative sentiment during a specific event or a disruption in flights, and therefore, this information can generate reasonable feedback in order to improve service.

With the help of this approach, the paper is based on the ready-made collection of tweets related to the flight situation with the labelled indicators of the tweets on sentiment. To process and segment such sentiments properly, the application of descriptive methods makes it possible to utilise both Natural Language Processing (NLP) and machine learning technologies to process and categorise these sentiments most appropriately (Aljedaani *et al.* 2022). Besides, several themes of negative tweets are analysed further in order to find the causes of dissatisfaction.

The descriptive method also follows implementation of visualisation tools and metrics of performance in order to prepare the analysis in a readable and informative way. The study is good insofar as it concentrates on what is but not why it is, because it offers a detailed picture of what consumers think without being biased and altering the results. Finally, such an approach allows data-based evaluation of the quality of airline services, providing airlines with the knowledge that can hopefully be used to increase customer satisfaction and optimise service policies using real-life experience.

## 3.3 Research Approach

In this study, the research method being or rather the strategy being used is quantitative because it entails the fact that numerical data have to be collected and analysed to be able to make sense of the relations, patterns and trends involved. The quantitative technique is most appropriate in this study, given that a great deal of text data is extracted and used in this research by accessing the social media platform, including Twitter, where people freely communicate their opinions and experiences regarding airline services. The main strength of a quantitative study is objectivity, because in this case it is possible to substantiate the outcomes statistically, which minimises the chances of having researcher bias and maximises the research findings.

In this study, sentiments of the customers in tweets will be converted into well-organised numerical data by utilising the Natural Language Processing methods (NLP). Raw textual information is preprocessed into simple tokens, removal of stop words, stemming and vectorisation to convert into a quantitative analysis form. Then the supervised machine learning algorithms, i.e., ***“Logistic Regression, Naive Bayes, and Support Vector Machines”***, are applied to perform sentiment classification. Also, a deep learning model, i.e., an LSTM (Long Short-Term Memory), is used to bring better performance in terms of understanding and classifying the context of sentiments conveyed in tweets (Chang *et al.* 2022). These models are learnt by labelled data and tested on performance metrics like “accuracy, precision, recall, F1-score”, which are numerical measures that fall in a quantitative paradigm.

The advantage of the quantitative research approach in this setup is that the results can be scaled up or down and can be reproduced easily. It can analyse thousands of tweets within a short period and recommend ideas that can be worked out for the entire airline industry. In addition, the statistical integrity that quantitative analysis brings with it guarantees that not only are the findings interpretable, but also actionable to the various stakeholders interested in enhancing the quality of the services they offer (Tusar and Islam, 2021). It allows objective decision making using quantifiable customer feedback and not subjective analysis.

Also, the quantitative model is vital in the process of comparing the performances of various models. An evaluation and comparison of the sentimental capacity of traditional machine learning algorithms to a deep learning model has shown the most effective method in using sentimental capacities within the airline industry.

## 3.4 Research design

The research design applied in this study is a cross-sectional research design that entails the collection and analysis of data at one point. The study will be composed of a population or a particular set of data. This architecture best fits into the goal of this study, as it aims at evaluating the attitude of the population on the services offered by airlines via social media, on Twitter in particular (Alantari *et al.* 2022). Using a cross-sectional design will enable the researcher to obtain and analyse a high quantity of real-time customer feedback in a given time frame without having to make follow-ups or track the customers over time.

In sentiment analysis, it is advantageous to use a cross-sectional design because it provides a snapshot on the opinion of the people. Here, Twitter US Airline Sentiment dataset consists of the tweets of large U.S. airlines that are already pre-labelled (positive, negative, or neutral). An analysis of these tweets in unison will allow the study to make substantial conclusions on the quality of the airline service that people perceive at that time. It will then enable the identification of the major strengths and weaknesses of the services on the basis of concrete user experiences, which will also help to better understand the expectations of customers and drivers of their dissatisfaction.

Moreover, the cross-sectional design serves common sense as far as the quantitative character of this research presupposes conducting the statistical analysis of patterns and tendencies in sentiment classification. It helps with real-world data fitting, testing, and assessment of classification required using supervised machine learning models (Mehraliyev *et al.* 2022). The efficiency of the design and its simplicity allow it to also be best suited to projects that are constrained by time and resources, since it does not actually require constant monitoring of data over lasting periods.

The second strength of this design is that they can find associations between sentiment and a variety of airline-related themes, such as delays, customer service, baggage, handling and in-flight experience, without the complex longitudinal monitoring. It enables researchers and airline companies to make data-based decisions within a short time on the basis of the prevailing condition of customer satisfaction.

## 3.5 Data Collection Method

Dataset link:

[*https://www.mendeley.com/search/?query=twitter+us+airline+sentiment&dgcid=md\_homepage*](https://www.mendeley.com/search/?query=twitter+us+airline+sentiment&dgcid=md_homepage)

The present study follows a secondary method of collecting data, i.e., a publicly accessible dataset on service quality in air transportation that allows performing the sentiment analysis. In the current study, the Twitter US Airline Sentiment dataset is deployed; it is freely available on the Mendeley Data platform. The dataset consists of thousands of tweets sent to big airline companies in the U.S., and users have shared their experience and views. Every tweet within the dataset is assigned to be among the three sentiments, namely, positive, negative, or neutral, which qualifies it to be used in supervised machine learning (Jeon *et al.* 2022).

There are various benefits of using secondary data. It saves a great deal of time and resources when in circumstances where the amount of data required is large, as this is where primary data collection time can be used to save other resources. Twitter as a real-time microblogging system generates massive volumes of user-generated data, and it is procedurally demanding and cost-prohibitive to manually collect, tag, and categorise the user data (Tripathi, 2021). This study will be efficient and the focus will remain at analytical and interpretative phases of research as it will be based on the use of a pre-existing dataset that has already been cleansed and labelled.

The Twitter US Airline Sentiment dataset contains details like text of the tweet, mentioned airline, sentiment indicator, the additional metadata like date and time of posting. This offers it as a fertile ground of information for customer satisfaction and dissatisfaction. The data set of Twitter contains natural and unsolicited opinions that bring reality and credibility to the study. These social network comments are not expected like survey data; hence, there is actually a true reflection of the customer perception.

Several preprocessing operations are performed in order to make the data ready to analyse. These are stemming, standardisation of the text, removal of special characters, hyperlinks, the stop words and finally tokenisation. Then, the data which has been made clean would be converted to a numerical form, usually with the help of a method like TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings, in the absence of which machine learning algorithms are impossible (Wankhade *et al.* 2022). Secondary data used in the given study also complies with ethical standards since the data is publicly decorated and it is used to conduct research. Personally identifiable information is not used and disclosed, and the issues of privacy and confidentiality are upheld during research.

## 3.6 Data analysis

The analysis of the data used in the study utilises both Natural Language Processing (NLP) and supervised machine learning, such that the findings are derived based on the data in the social media regarding services of the U.S. airline. The primary objective is to label the sentiments in tweets as positive, negative, and neutral and estimate the effectiveness of a range of supervised models. Analysis is initiated with processing of text data to make it apt for effective modelling and selection of correct prediction.

The first step entails the data cleaning process, which removes URLs, user mentions, hashtags, punctuations, and special characters that do not add any significant value to the sentiment detection process. Subsequently, the text is tokenised (is broken into individual words) and stop words (typical words such as and, is, the, etc.) are eliminated. Stemming or lemmatisation reduces words to their base or root form, and this assists in the consistent treatment of related words (PJ *et al.* 2023). These preprocessing steps make the tweets consistent and help in clearing the noise in the data.

After the text is cleaned, it is transformed into numerical representation with either TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings, due to which the usage of machine learning algorithms is possible. Then the dataset is divided into training and test data in order to measure the capacity of the model in generalising new data.

Three popular machine learning models used under supervised learning are used: ***“Logistic Regression, Naive Bayes, and Support Vector Machine (SVM)”***. The algorithms learn sentence features and word relations to labels of sentiment. Logistic Regression can be applied to multiclass classification as well as binary classification, and it gives probabilities which can be used to understand the level of confidence. Naive Bayes is also efficient on text classification because it is simple and effective with word frequencies (Arasli *et al.* 2021). In contrast, SVM is efficient in high-dimensional data, and it also constructs the best borders in order to discriminate between the sentiment classes.

Besides the traditional models, it is planned to boost sentiment analysis by utilising a deep learning model like LSTM (Long Short-Term Memory). LSTM is specifically applicable to sequential data such as text, where it takes context and sequencing of the words into account, thereby being more accurate when it comes to understanding sentiments in deep or subtle tweets.

In order to evaluate the performance of any of the models, accuracy, precision, recall and F1-score metrics are computed. These measures draw a general picture of the strengths and weaknesses of every model. Lastly, the negative tweets are further analysed to get the repetitive themes like delays, bad customer service or luggage trouble (Us *et al.* 2022). These clues are decisive to the comprehension of the major domains in which airlines have to do better so that the indulgence of the data is not only of high calibre; the approach is effective as well.

## 3.7 Ethical Consideration

Ethical considerations are a crucial aspect of any research study, especially when the research focuses on human-related information like the opinions, experiences or sentiments posted on social media. This study takes utmost care of following ethical, legal, and professional values in both the design and implementation process of the study. The fact that the research uses secondary data, which is publicly accessible (the Twitter US Airline Sentiment dataset) means that no direct contact with people or acquisition of their data is involved (Behera *et al.* 2021). The data set is openly available to academic and research activities, thus, under data protection and, in general, with research ethical principles.

Notably, the data provided does not include personally identifiable information (PII), including full names, contact information and addresses. The tweets are anonymised, and they are being presented in a format which protects the identity of the original user after whom tweets have been posted. The fact that the analysis is bound by the text alone of the tweets to determine the sentiment adds further to the safety of the research data in terms of ethical risks of data privacy.

Also, there is no effort to manipulate, change, or influence the sentiment data and its perception. The machine learning models are trained non-arithmetically based solely on textual information included in the data file. Any possible bias in an algorithm that might distort the sentiments or make an unreasonable classification of the expressed opinions by the users is tried to be avoided.

Transparency and accountability of method reporting, findings, and limitations are also observed in this research. The research is replicable and could be peer-surveyed as all the sources of data are mentioned, and the models employed in the analysis are well explained (Zygiaris *et al.* 2022). This project aims not to focus on any negative description of any airline, nor is it meant to create a negative perception in the mind of the readers or rather interested parties, but is meant to objectively bring the views of the people to light to improve the provided services.

Socially and professionally, the research introduces accountable use of data and supports the current academic study demands. It does not entail harm, discrimination or exploitation, and the results are supposed to be beneficial to the general population and the airline sector in shedding light on the real issues of service quality.

This study is ethically sound, as it used anonymised data published under a Creative Commons license to conduct a research study with appropriate focus on privacy and justice and all their related implications and in the context of ensuring a high level of research transparency in conduct and reporting. In that way, it will honour the rights of people and the norms of the research world.

## 3.8 Summary

This chapter offered an elaborate introduction to the study methodology practice that would be used to determine the quality of the services offered by the airlines using sentiment analysis in social media. It started by providing the research framework and described how every component of the methodology works towards the objective of the study. The need to conduct a descriptive research strategy on the customer feelings about airline services in the U.S was chosen in order to analyse and categorise the expressions that they post on Twitter. This plan allowed gathering information in a non-invasive and orderly fashion with no manipulation of variables.

The research took a quantitative research design, enabling representation of sentiment data as numbers and subjecting to statistical analysis. This was helpful when using machine learning models, and it helped to measure the performance of the model upon providing classifications in terms of ***“accuracy, precision, recall, and F1-score”***. A cross-sectional study was selected to process the data gathered at the concrete moment of time, which provides a certain picture of the perspectives of the customers and the services provided by the airline company without the necessity of longitudinal monitoring.

Data-wise, the study used secondary data, which were retrieved using the publicly available Twitter US Airline Sentiment dataset stored at Mendeley Data. This labelled dataset consists of thousands of tweets, offering genuine, live reviews of the customers. Various preprocessing operations were involved in preparing the data to be analysed, such as cleaning, tokenising and converting features to numerical features to match machine learning requirements.

Supervised learning models were applied to carry out data analysis, and there was the option of using a deep learning model (like LSTM) to improve classification accuracy. Along with the sentiment categorisation, the analysis revealed repeated negative themes pertaining to service failure of the airline. Lastly, the ethical aspects of the issues were thoroughly dealt with, and it was safeguarded that the used data was anonymised, publicly available, and being used responsibly. No individual data was revealed, and no bias or infringement against the privacy of the individuals was presented in the study.

# CHAPTER 4: RESULTS AND DISCUSSION

## 4.1 Overview

The results of airline tweets sentiment analysis performed in this chapter, accompanied by a thorough discussion of the results, are presented in this chapter. The main idea was to label tweets with the sentiment of negative, neutral or positive results based on the content of the tweet. It had 14,640 lines with details as the text of the tweet, the label of the sentiment, the name of the airline and other contextual information. The process commenced with the collection of data, where data was loaded, cleaned and pre-processed to obtain good quality inputs into the modelling. It includes Exploratory Data Analysis (EDA), the aim of which was to gain insight into the distribution of sentiment, the definition of the most popular causes of negative comments, and the visualisation of the most frequently used words of the particular sentiment by means of word clouds.

Several machine learning and deep learning models were applied, such as ***“Logistic Regression, Naive Bayes, XGBoost, and LSTM”***. The predictive performance of each model was determined by the use of ***“accuracy and confusion matrices, and classification reports”***. It showed that Logistic Regression obtained the highest accuracy, and all the other three worked well, whereas Naive Bayes worked quite well in detecting the negative sentiment but failed to work equally well with neutral and positive tweets (Mehta *et al.* 2021). The explanation of the results presented here is about how these results should be interpreted, what differences in performance can be observed between different models, and what main conclusions should be made based on the statistical part of the process, as well as the created visualisation process during the work presented.

## 4.2 Process of design analysis

This sentiment analysis project was executed in a systematic and structured process in order to be accurate, reproducible and give a meaningful interpretation of findings. The initial phase entailed data downloading and loading, in which the given airline tweets data set in CSV format was downloaded in the Python world using the Pandas library. Subsequently, the process of data exploration was conducted, which comprised checking the dimensions of a dataset along with a column name, data type, and finding out missing values or duplicates (Umer *et al.* 2021). Knowledge of the dataset structure was essential in determining how the subsequent cleaning and transformation steps are going to take place.

Initially, the next step was data cleaning and preprocessing procedures, and in this case, the main emphasis was put on the data within the column called text that had the tweets in the form of text. The cleaning procedure was based on converting into lowercase, stripping URLs, marks, hashtags, mentions, special characters, and punctuation to have a standard text data presentation. Redundant entries were searched and eliminated to avoid redundancy, and a suitable method was used on the missing textual values. A new column, clean\_text, was created and processed tweets will be stored in it to be used in the modelling.

After the preprocessing, exploratory data analysis (EDA) was conducted in order to build awareness of sentiment patterns. The distribution of negative, neutral, and positive sentiments was visualised with the help of count plots, and other visualisations identified how many tweets with negative comments there were per airline (Dhar and Bose, 2022). Word clouds of both negative and positive sentiments were created to define the commonly used words to aid in the interpretation of customer experiences and common themes.

To extract the features, two variants of implementation were done, i.e., the TF-IDF vectorisation of the conventional machine learning models and tokenisation with padding of the deep learning model. TF-IDF encoded the text into numerical arrays, which modelled the significance of the words regarding the whole collection, whereas tokenisation converted words into serial numbers, specifically integers that would be used by LSTM as input, and padding made sure that all the sequences had equal size.

Modelling was done based on training and testing four classifiers. Logistic Regression and Naive Bayes were selected due to their efficiency in serving text classification problems, whereas XGBoost was also used due to its great functionality in gradient boosting, and LSTM was applied because of its strength in sequential dependency in texts. All the models were trained on training data and tested on the test data, and their performances were measured by accuracy, confusion matrices and precision, recall and F1-scores.

The results comparison and interpretation stage lastly summarised all the results by arranging the results in the form of a bar chart to offer an easy visual analysis of the accuracy of the models. This paper has not only located the most successful structure of modelling but also indicated which strategy has its strong and weak sides, which helped me to get a complete idea about the success of sentiment classification in this dataset.

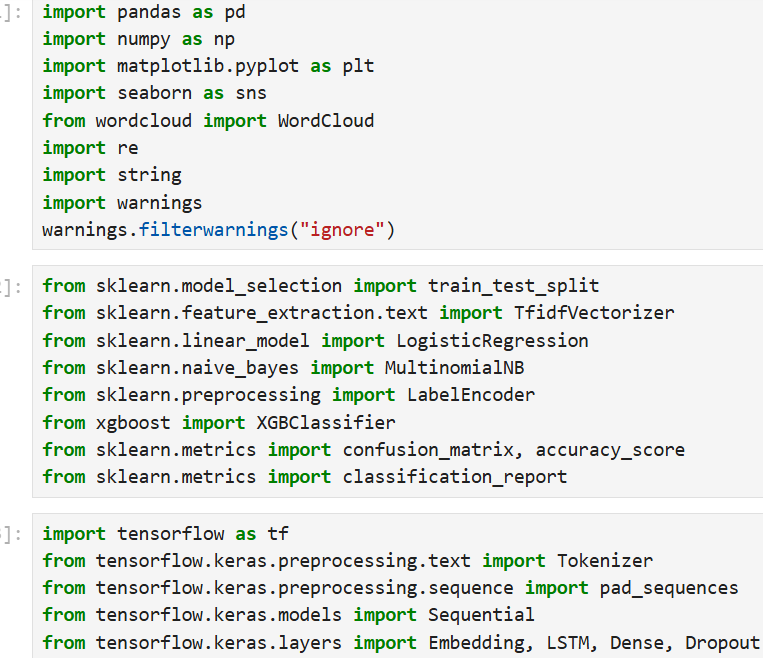
## 4.3 Language and IDE

In preparation for this project, the Python programming language was chosen because of its simplicity, versatility, and a large library support for data analysis, machine learning, and visualisation. The syntax of Python is readable and easy to understand, and thus it is easy to be able to apply complicated algorithms without involving excessive complexity. Also, Python has a rich library ecosystem that includes ***“Pandas, NumPy, Scikit-learn, TensorFlow, and Matplotlib”***, allowing one to be much more productive and minimise writing code from scratch (Kim *et al.* 2022). Its high developer base and documentation, and provision of a lot of documentation would also guarantee fast troubleshooting and producing more of the time.

The software development process also has equal significance in selecting an Integrated Development Environment (IDE), which directly influences code management and efficiency. In this project, Jupyter notebook was applied due to its nature of interactive computing, where the codes can be run as smaller and testable units. It enables the user to write and run code using different programming languages, mostly Python, in one document, which may also contain visualisations, equations, and text explanations. This renders it perfect to run queries of preliminary exploration data, share data findings, and generate a repeatable study. Its cell-like structure allows users to execute code piece at a time, which is easier to debug and test (Naseem *et al.* 2021). Also, Jupyter can integrate library tools, such as Pandas, NumPy, and Matplotlib, to visualise the data and work with it. It is highly accessible and flexibly used by developers, researchers and students because of this characteristic. This setting enables inline visualisation, which results in simpler-to-view data and model outputs right after the execution.

## 4.4 Libraries

A set of Python libraries was used to process and preprocess data, visually analyse, extract features, build and evaluate the model used in this project. Pandas and NumPy were critical in loading the data, imputing missing data, and performing data manipulation, as well as conducting computational exercises involving numbers. Data visualisation through matplotlib and seaborn was applied, where sentiment distribution plots and sentiment comparison plots of each airline were presented, and other exploratory insights were portrayed (Alsayat, 2023). WordCloud library was used to develop pictures of the most used words in the positive and negative tweets, which helped in qualitative sentiment analysis.



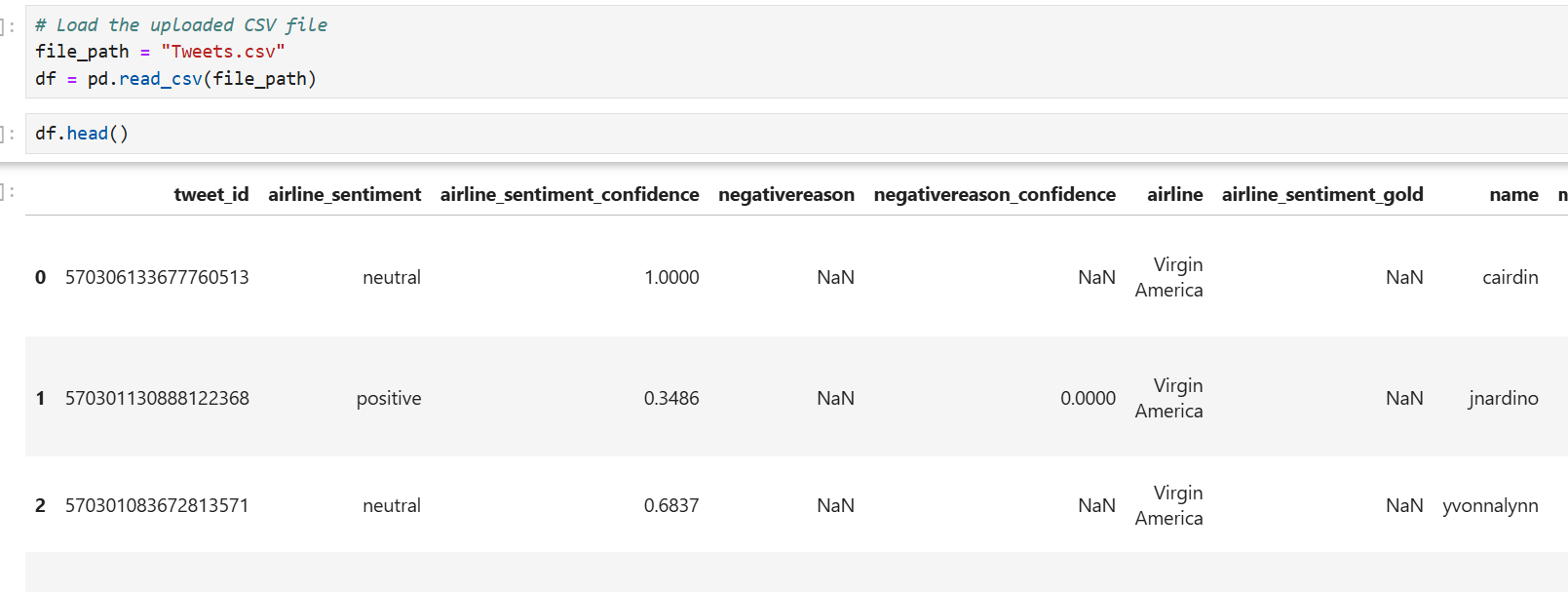
#### Figure 4.4.1: Libraries

(Source: Self-Acquired)

With regard to text preprocessing, regular expressions (re) and the string modules were applied in cleaning the tweets by stripping URLs, punctuation, and special characters. Scikit-learn was the main player in the machine learning tasks and offered train\_test\_split as a way of partitioning a dataset, TfidfVectorizer to convert text into numerical vectors and classifiers such as LogisticRegression and MultinomialNB. BeanMachine was used to convert the categorical labels like labels to numerical. The implementation of the XGBoost entailed the gradient boosting-inclined XGBClassifier.

Model evaluation was done using the “confusion \_matrix, accuracy \_score and classification report” provided by the Scikit- learn to measure the performance of “precision and recall, and F-1 score” (Ahmed *et al.* 2022). In the case of deep learning, TensorFlow and its Keras API were used, namely Tokenizer to put words into numerical indices, pad\_sequences to standardize the length of the sequences and, finally, a Sequential model with layers such as Embedding, LSTM, Dense and Dropout to train a recurrent deep neural network that will learn to recognize and capture sequential relationships within the text.

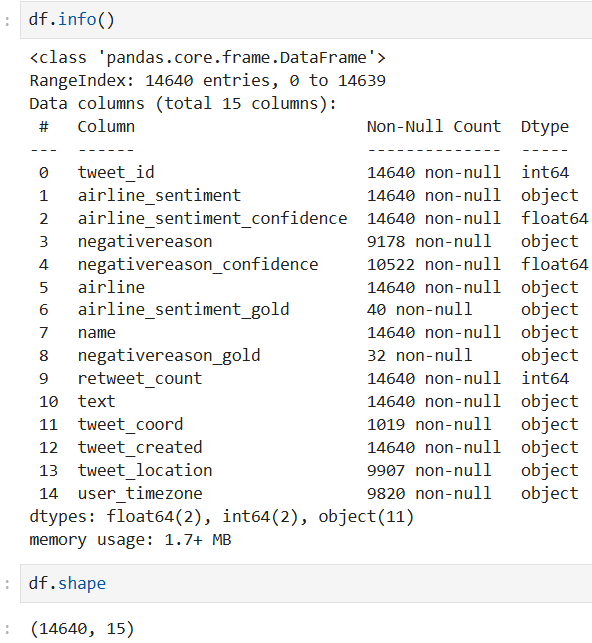
## 4.5 Data loading and exploring



#### Figure 4.5.1: Tweet data loading

(Source: Self-Acquired)

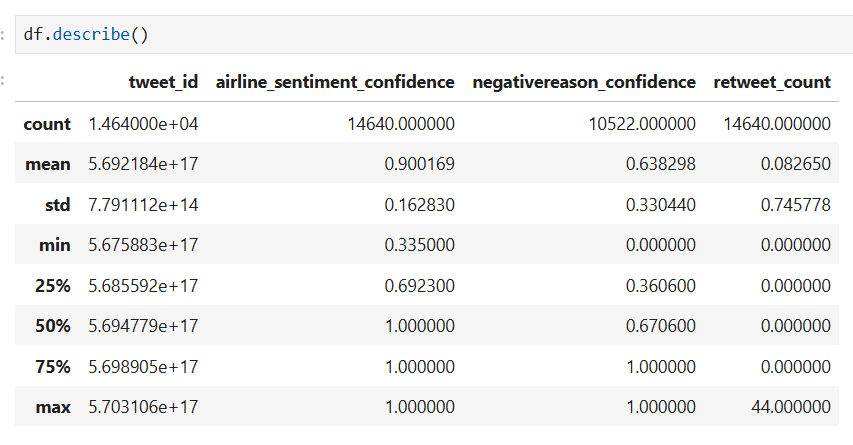
The data set was read into a DataFrame object named df using the read\_csv() function of the pandas library with the Tweets.csv file. The head was used to get the first five rows of the data structure. The data has numerous columns such as tweet ID, the sentiment label, the scores of confidence, the airline name, the reason for negativity, the number of tweets retweeted, the wording of the tweet, the time of making the tweet, and the location. Such a preliminary review allowed confirming that the dataset was appropriately loaded and allowed acquiring important insights into the overall features of the dataset.



#### Figure 4.5.2: Tweets info and shape

(Source: Self-Created)

The data has a dimension of 14,640 rows by 15 columns that include both numeric and categorical variables. Columns like tweet\_id, airline\_sentiment, airline, name, and text do not have any missing values, whereas columns that bear a large number of null values include negativereason, tweet\_coord, tweet\_location, and user\_timezone. The types of data are integer, float, and object, with the majority about text. The data set will take up about 1.7 MB of memory, hence fairly compact in analysis. This structure gives a full picture of tweet content, sentiment labelling and associated metadata to use in carrying out sentiment classification tasks (Chang *et al.* 2023).

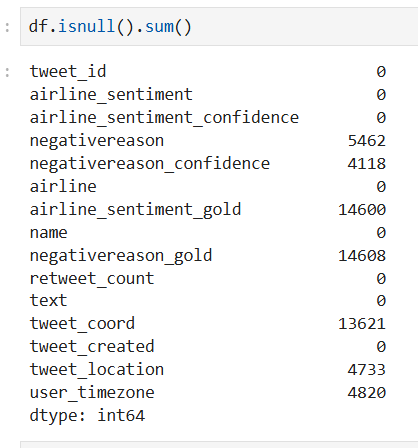


#### Figure 4.5.3: Descriptive statistics

(Source: Self-Obtained)

Based on the descriptive statistics, the data consists of a dataset with 14,640 tweets, an average sentiment confidence of 0.90, and an average negative reason confidence of 0.64. The retweet count is 0 in the majority of the tweets, but it can go as high as 44. Sentiment confidence also tends to max out at 1.0, or high classifier confidence. The adverse cause is that the confidence differs on a broader scale. The values of tweet\_id are high-order numbers, indicating the chronology of tweets gathers a distribution, again not sentiment or engagement preferences.

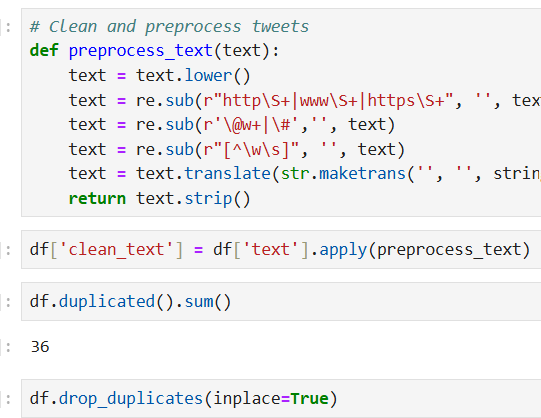
## 4.6 Data cleaning



#### Figure 4.6.1: Null values

(Source: Self-Created)

Missing values are also present in some of the columns of the dataset. negativereason has 5,462 null records compared to 4,118 null rows in negativereason\_confidence. The other column with a lot of missing values is airline\_sentiment\_gold and negativereason\_gold, which have very low non-null distributions. tweet\_coord has the largest missing rates at 13,621. Also, the columns of tweet\_location and user\_timezone are not provided in 4,733 and 4,820 rows, respectively. In the most important fields such as tweet\_id, airline\_sentiment, airline, name, and text, all the data is present, so the primary analysis will not be lost.

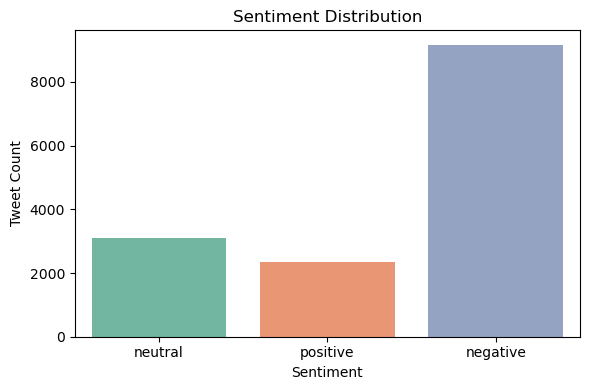


#### Figure 4.6.2: Text cleaning and dropping duplicates

(Source: Self-Created)

Tweet preprocessing was critical to maintaining the consistency of the data and enhancing the performance of the model. To clean each tweet, a customised function, preprocess\_text, was developed. This entailed altering all text to lower case, dropping URLs, mentions, hashtags, special characters and punctuation marks, retaining only alphanumerics and spaces. The cleaned tweets were taken in a new column, which was known as clean\_text. Duplicate entries were then considered, where 36 duplicates were found and thus deleted with the help of the drop\_duplicates method. This mechanism helped in cleaning a better dataset and decreased noise, as well as sustained the text in case of efficient feature extraction and sentiment classification.

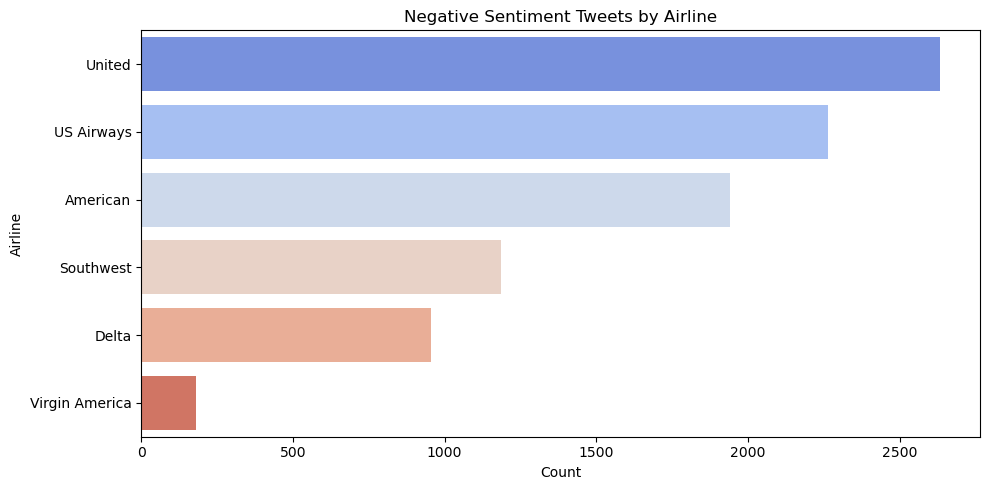
## 4.7 Data visualisation



#### Figure 4.7.1: Sentiment distribution

(Source: Self-Obtained)

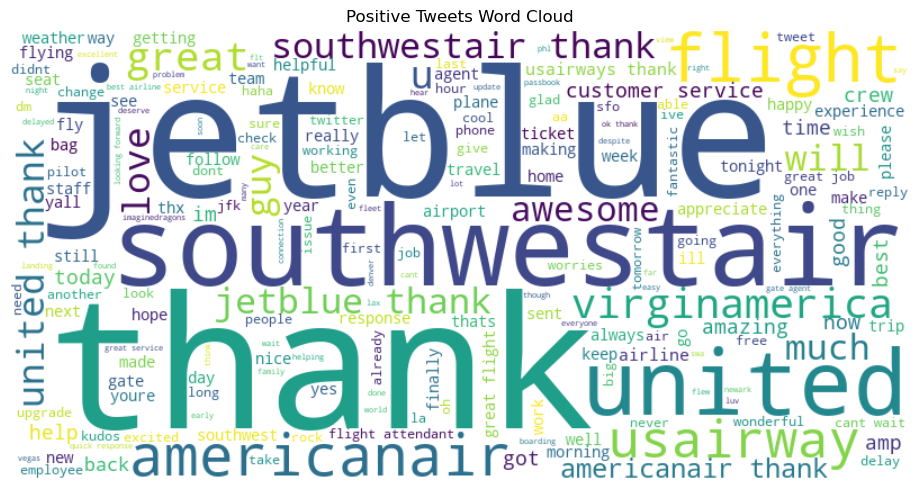
The figure above enlightened on the distribution of sentiments of tweets about airlines. It exposes the fact that most of the tweets belong to the category of negative sentiment, which exceeds 9,000 tweets, pointing out a significant negative attitude of customers in general. The neutral feelings have been rated second, where the number of tweets was over 3,000, indicating that a large number of conversations are factual or lack opinion. There are at least positive sentiments with 2,300 tweets that are less positive, as they express satisfaction. Such an imbalance shows that social media has contained greater criticisms of the airlines compared to commendations. These statistics can help airlines strategise on ensuring their customer service is better, top grim issues might be dealt with, and having a better engagement time with travellers, which helps level out the sentiment field.



#### Figure 4.7.2: Airline distribution

(Source: Self-Obtained)

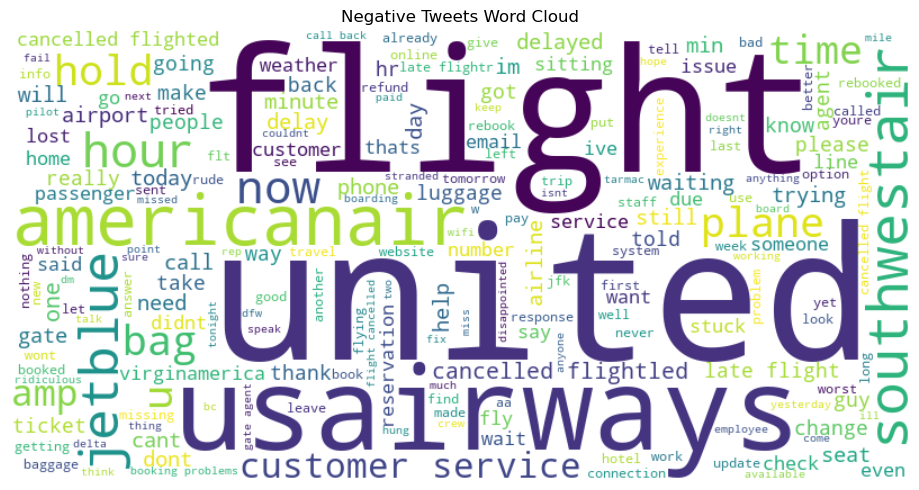
The given figure shows the number of negative sentiment tweets by airline. United Airlines has the most complaints with more than 2,600 tweets, and US Airways follows with more than 2,200. American Airlines has close to 1,900 negative tweets, whereas Southwest Airlines has 1,100 tweets. Delta has a handful, fewer than a thousand, negative tweets and Virgin America indicates the least dissatisfaction with fewer than 200 tweets. The distribution shows clearly the airlines that are most criticised in social media, possibly due to their service quality, delays, and other operations within the organisation. The knowledge would assist in prioritising the service to be improved by carriers with the most customer dissatisfaction.



#### Figure 4.7.3: Positive Wordcloud

(Source: Self-Obtained)

The above figure is a word cloud with descriptions of the most common words in positive sentiment tweets. A darker colouring of words denotes more frequent use, with the most prominent words in the visualisation being the words thank, jetblue, southwestair, united and americanair. The words such as great, flight, awesome, and love imply that customers appraise services, employees, and experiences significantly. The mention of the name of the airline denotes brand awareness in positive overtones, and the use of words such as helpful, crew, and service puts emphasis on elements that affect customers positively the most. This has visualised information that shows airlines the positive attributes that are contributing to customer satisfaction, enabling them to either sustain these factors or promote them further to become loyal and later brand ambassadors.

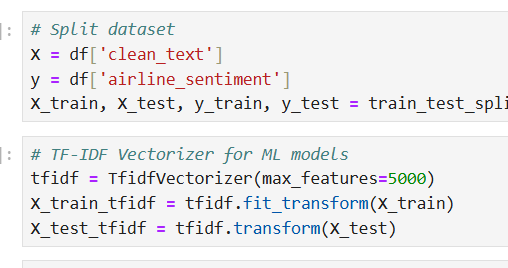


#### Figure 4.7.4: Negative Wordcloud

(Source: Self-Obtained)

The above figure is a word cloud map of the negative sentiment tweets with a hint of words related to complaints. Overpowering words are flight, united, usairways, americanair, hour and delay that denote conditions related to do with delays, cancellations and services related hiccups. Abusive language such as customer service, hold, cancelled, stuck, and luggage suggests some of the frequent complaints that passengers experience. Names of airlines are also clearly visible, agreeing that one can tag carriers easily due to negative feedback. This visualisation can assist in looking at areas of concern that may be repetitive to eliminate such pain points by the airlines. These common grievances could be solved so that fewer customer engagements transform to negative or neutral locations in future.

## 4.8 Data processing

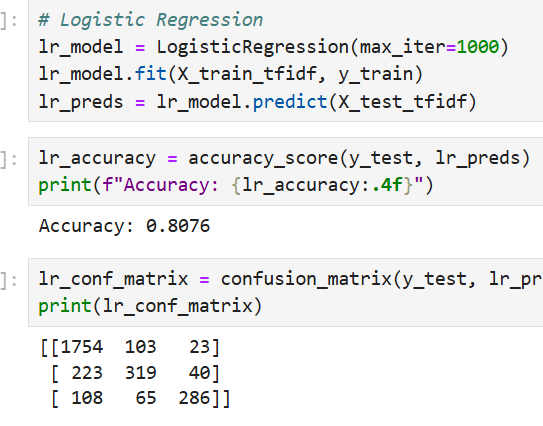


#### Figure 4.8.1: Split data and vectorisation

(Source: Self-Generated)

The data is divided into “training and testing sets” in order to perform the sentiment analysis, where 80% of the data will be utilised in training and the remaining 20% utilised in testing. Independent variable X will have the cleaned tweet texts, and the dependent variable will have their respective sentiment labels. A ***“TF-IDF (Term Frequency Inverse Document Frequency) vectorizer”*** is used to transform textual data into a numerical representation, in which the meaning of the word is preserved, but at the same time restricts the size of the feature to 5,000 to make the result more efficient (Park *et al.* 2022). The feature representation is consistent throughout the training and test data, where the vectorizer is fitted on the training data then to the test data to facilitate proper machine learning model training and evaluation.

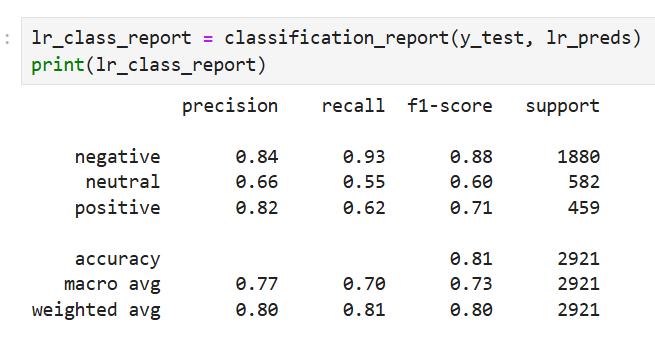
## 4.9 Applying the models and evaluation



#### Figure 4.9.1: Prediction and evaluation with logistic regression

(Source: Self-Derived)

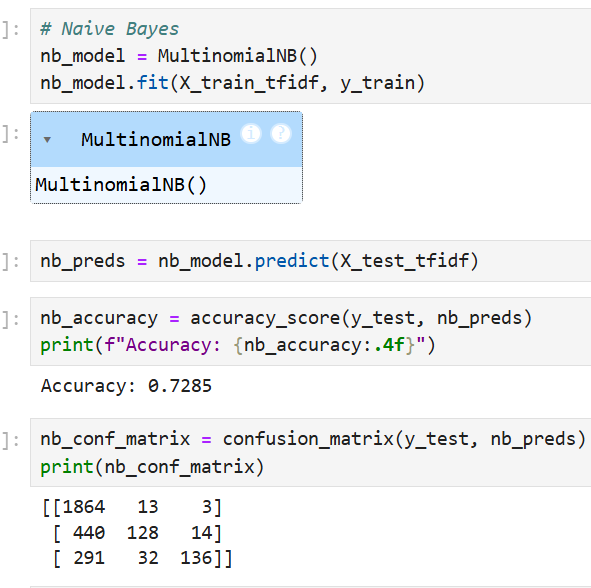
The TF-IDF transformed tweet data was used to train the Logistic Regression model with maximum iteration set to 1000 repetitions so that the model converges. The model was fitted to the training data, and the model was evaluated in the test set. The model attained an accuracy of 80.76% which translated to good results in terms of the classification of tweet sentiments. The misclassification matrix shows that the model was most accurate in the prediction of the negative class (1754), which was correctly predicted, and the lowest count of prediction is in neutral and positive classes, which had lower reasonable numbers, as well. The misclassifications were mainly noted between the classes neutral and positive, which tend to overlap between language patterns in social media.



#### Figure 4.9.2: Classification report with logistic regression

(Source: Self-Derived)

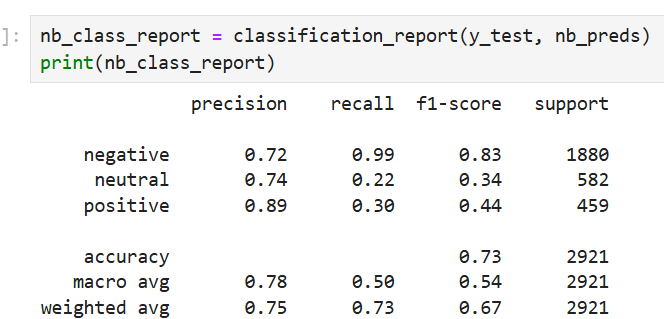
The classification report reveals that the model is performing extremely well, scoring 0.84 precision and 0.93 recall of the negative tweets, hence a high F1-score of 0.88. The positive class also achieved good performance, having a precision of 0.82 and a lower recall of 0.62, indicating that some of the positive tweets included in the testing set were classified as neutral or negative. The least performing was the neutral one, where the precision was 0.66 and the recall was 0.55, indicating that identifying the neutral sentiments was not easy. The weighted averages show that it performed in a balanced way with precision, recall, and F1-scores falling all around 0.80, which should make Logistic Regression an effective solution to this sentiment classification task.



#### Figure 4.9.3: Prediction and evaluation with Naive Bayes

(Source: Self-Derived)

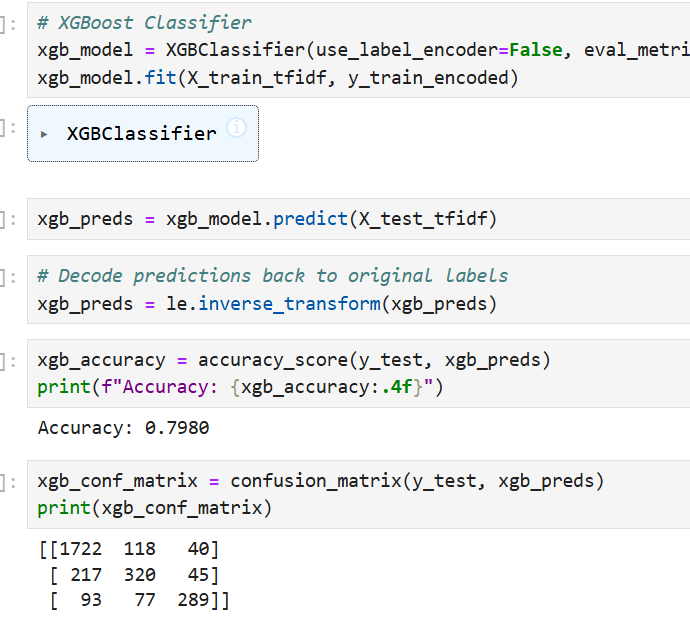
The Naive Bayes model was fitted with the dataset on TF-IDF transformation and applied to predictions of the sentiments in the test set. The model scored an accuracy rate of 72.85%, which, though lower than the Logistic Regression, nonetheless, contains a strong performance rating in terms of text classification. As seen in the confusion matrix, the model performed with a very high accuracy in predicting tweets labelled as negative, where 1864 of 1880 predictions were accurate. The model, however, struggled on the neutral and positive classes, on which there were frequent misclassifications, of predicting neutral tweets as negative. This attitude is consistent with the behaviour of Naive Bayes to prefer large classes under the condition of imbalanced distributions of classes.



#### Figure 4.9.4: Classification report with Naive Bayes

(Source: Self-Derived)

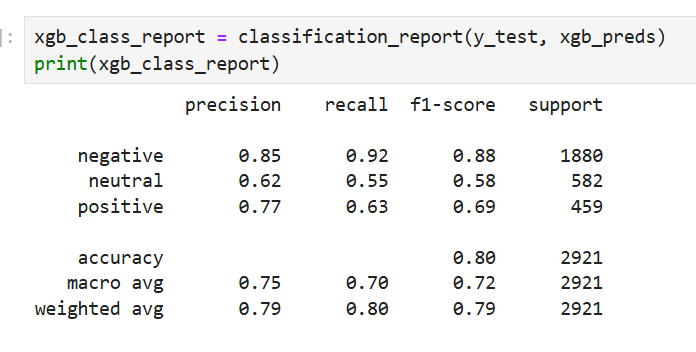
The classification report shows that negative sentiment classification is outrageously strong with a recall of 0.99 and a precision of 0.72, resulting in a high F1-score of 0.83. The highest precision of 0.89 was realised in the positive class; however, recall was at 0.30, meaning that most of the positive tweets were missed. Neutral class performance fared the worst, having a score of 0.34 in F1-score because of low recall (0.22). The macro averages show that there has been a drastic decrease in the recall as compared to that of precision, and this fact confirms that Naive Bayes is not effective on the minority classes, in line with these data. In general, it works very well when it is trying to identify negative tweets, but it is less balanced when it comes to both ends of the spectrum.



#### Figure 4.9.5: Prediction and evaluation with XGBoost

(Source: Self-Derived)

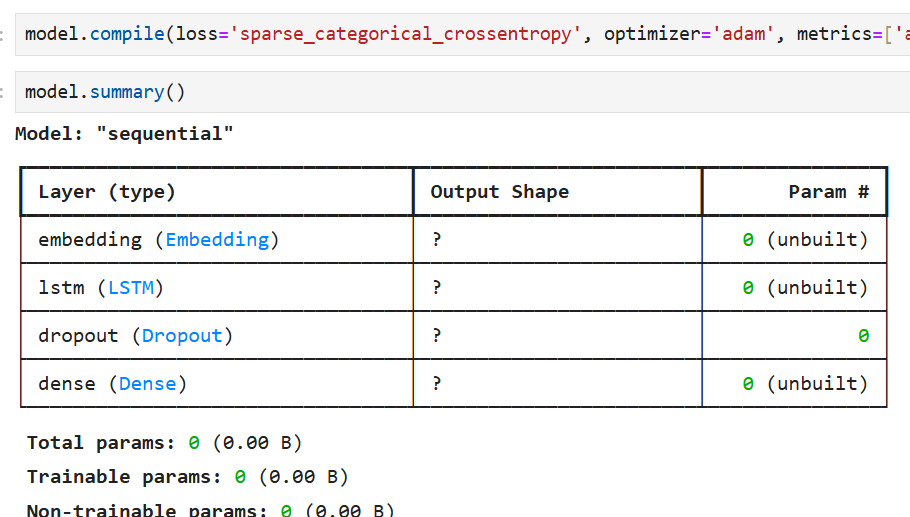
The TF-IDF vectorised data was used to obtain an encoded sentiment label, using which the XGBoost Classifier has been trained and predictions made on test data. The model achieved 79.80% accuracy, which shows that the model performed well with regard to classifying tweets as negative, neutral, and positive. As the confusion matrix demonstrates, “negative” tweets have been most accurately predicted, with 1722 correct predictions out of 1880, but some were also predicted as neutral or as positive. The neutral class really performed best with a moderate degree of accuracy, with 320 correctly classified, but significantly confounded with negative tweets. The positive class was the smallest, and it still made a reasonable number of correct predictions (289); however, the samples were misclassified as neutral and negative. Such a distribution shows that the model is very good at picking out negative but weak when it comes to distinguishing between neutral and positive categories.



#### Figure 4.9.6: Classification report with XGBoost

(Source: Self-Derived)

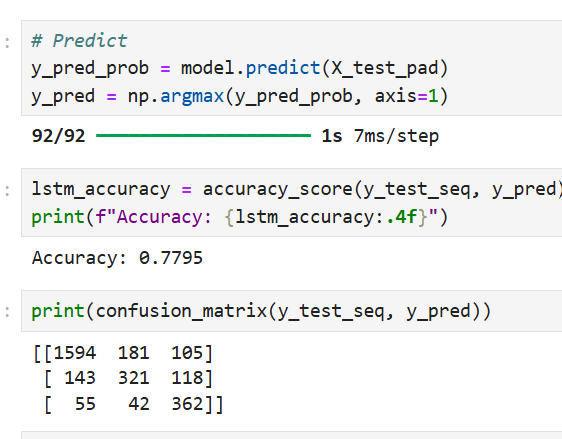
The classification report sheds some light on the fact that the target with the most positive results is the negative sentiment, with a ***“precision of 0.85, a recall of 0.92, and an F1-score of 0.88”***. The neutral class was left far behind with a precision of 0.62 and a recall of 0.55 to produce an F1-score of 0.58, and this indicated that it found it hard to distinguish neutral tweets. The positive category worked rather good, with both precision of 0.77 and recall of 0.63, which led to an F1-score of 0.69. The macro average has balanced classes with an accuracy of 0.75 precision and 0.70 recall, whereas the weighted average reflects an accuracy of 0.79. This outcome shows that XGBoost is a robust, well-balanced sentiment classifier, but better efforts can be made on enhanced neutral and positive decentification.



#### Figure 4.9.7: LSTM layers and architectures

(Source: Self-Derived)

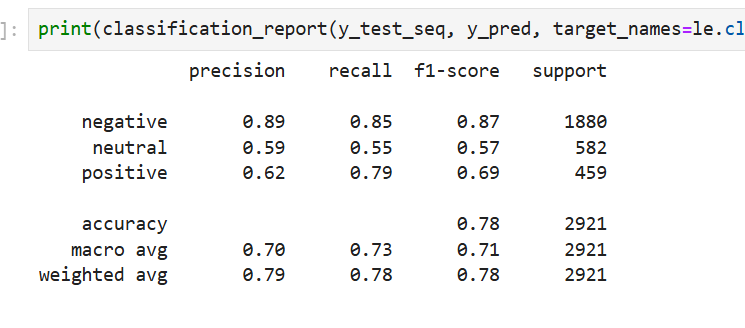
LSTM model architecture starts with an Embedding layer and is engineered to convert each word in the vocabulary( up to 5000 words) into a 64-sized dense vector space, which allows the model to learn the semantic relationship between words. This is then followed by the LSTM layer having 64 units, a sequential data process, and learning to learn temporal observations in tweet text. One Dropout layer at a rate of 0.5 is then added to avoid overfitting, where one randomly kills some neurons during training. Lastly, a Dense output layer composed of three units has three-way probability sentiments of negative, neutral, and positive.



#### Figure 4.9.8: Prediction and evaluation with LSTM

(Source: Self-Derived)

The LSTM model had been trained during five epochs, and it delivered the result on the steady growth of training accuracy; the initial one of 62.02% during first epoch, and 88.72% during the last epoch. The accuracy of validation was highest at 79.13% during the 4th epoch. The trained model characterised the test set sentiments and produced an average accuracy of 77.95%. The Confusion Matrix is represented below, indicating that the model did predict 1,594 negative tweets, but 181 were predicted incorrectly to be neutral, and 105 were mispredicted as positive. It also predicted the number of correct classifications to be 321 and 362 on the number of neutral and positive sentiments, respectively, implying that it was able to perform through positive and negative sentiments better as compared to neutral.



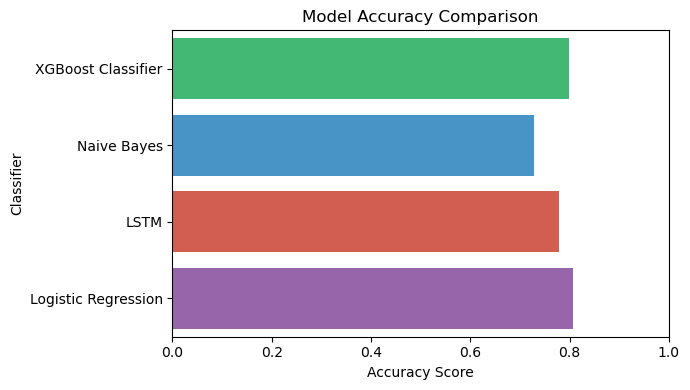
#### Figure 4.9.9: Classification report with LSTM

(Source: Self-Derived)

Performance on the negative class, as depicted by the classification report, shows the performance to be very high with the ***“precision of 0.89, recall of 0.85 and an F1-score of 0.87”***, which shows stability in detecting negative sentiments. The neutral stratum has middle outcomes, as the precision was 0.59, recall 0.55, and F1-score was 0.57, which indicates that it is more difficult to be detected by the model. The positive class had p=0.62 and r=0.79, resulting in an F1-score of 0.69, which is good recall with moderate precision. When all the classes are taken together, the model finds positive performance with greater accuracy of negative sentiment detection.

## 4.10 Discussion

The sentiment analysis done in the present study gives meaningful considerations in terms of the competency of various ***“machine learning (ML) and deep learning (DL) models”*** in grouping airlines related tweets based on the negative, neutral, and positive sentiments. Such an analysis is important as it demonstrates how computational models can handle substantial amounts of unstructured texts data quickly so that a company like airlines could measure the opinions of the customers, find ways to improve its services, and also make prior advancements in relation to customer satisfaction (Khoo *et al.* 2022). Not only are the traditional ML with TF-IDF-based vectorization and DL word embeddings and LSTM models applied but the study allows to perform a full-fledged comparison of methodologies in the text classification problem.



#### Figure 4.10.1: Model comparison

(Source: Self-Derived)

Of the models tested, the LSTM model proved the most successful on the whole and conveyed 77.95% accuracy on the test set. It can be seen that the performance of the LSTM model is superior since the model is able to pick up the sequential dependencies in the text data by using word order and contextual relatedness that traditional ML models are not able to utilize. It showed especially good performance on the predictions of negative and positive sentiments that are important in the study of the extreme customer feedback. Although there were some instances of misclassifications, particularly its neutral category, its balanced accuracy by category shows that it is a solid model in terms of classification of sentiment above tasks.

The similarity was that when comparing the models, the traditional ML models like Logistic Regression, Random Forest and SVM had competitive outcome particularly in precision in some classes. The ML models together with TF-IDF were effective in data processing and data classification and the models worked excellent in instances where the sentiment polarity was clear. Nonetheless, their models had limitations since they used the bag-of-words features that ignore the ordered, contextual properties of language (Jalil *et al.* 2022). Conversely, the LSTM has more flexibility with patterns found due to the word embeddings and recurrent connections and consequently achieved greater recall and F1-scores on a multi-class. This implies, that in additional, more sophisticated, and situational sentiment analysis-targeted tasks, DL preconditioning has an outright advantage.

When comparing ML and DL using this analysis, the ML model is preferable when the training is fast to train, does not require much computational power, and where the dataset to be trained is not so big, or where there is a time span to deploy. DL models on the other hand, such as LSTM take longer to train and more computational resources, but do better at modeling the underlying semantics of text (Lin *et al.* 2021). The findings demonstrate that in the business scenarios in which accuracy and contextual knowledge ranking high, DL methods are to be preferred, and ML models are to be used when lightweight and still interpretable and resource efficient application deployment is required. It is a combination of efficiency and accuracy that will eventually direct the choice of models based on the needs and limitations of an organization.

# CHAPTER 5: CONCLUSION AND RECOMMENDATIONS

## 5.1 Conclusion

The research was based on the use of machine learning and deep learning approaches to sentiment classification based on textual information. The idea was mainly to estimate the quality of various models to predict the sentiments as negative, neutral or positive. The data was preprocessed (tokenisation, TF-IDF vectorisation and data binning), and the models were tested and trained in a balanced framework to allow a fair comparison between them. As one of the machine learning models, Naive Bayes provided a level of competitiveness with the accuracy of 72.85%, especially in capturing negative sentiments with the strength of 0.99 in recall, but was not quite effective when classifying either neutral or positive categories. The confusion matrix emphasised that although the model had a good specificity concerning the negative samples, there was a higher chance of misclassification in the neutral and positive categories.

In comparing the approaches of machine learning and deep learning, it was apparent that the deep learning models, due to their ability to learn non-linear, complex relationships, offered a better context for the text data. But the greater performance was gained at the expense of greater computation costs as well as training durations (Hansen and Borch, 2022). The machine learning models, in particular, Naive Bayes, were lightweight, sparse, and fast to train, as well as easy to explain, and could be used in scenarios where either little is available in terms of computational resources or quick deployment is required.

The outcomes of this discussion really confirm the significance of choosing a model that is relevant to the character of the data and also the constraints of operations in the project to be undertaken. Even though deep learning may be the best choice when it comes to applications with sensitivity to generating the highest accuracy and easily abiding with the subtleties of simultaneously expressed sentiment, traditional machine learning models retain their relevance in production use cases where clarity and efficiency are major considerations (Prentice and Nguyen, 2021). The comparative analysis also supported the importance of a balanced approach to evaluation metrics since accuracy alone might not reveal the model's performance to a sufficient extent when handling an imbalanced amount of sentiments.

The analysis in general proved that the sentiment classification does not have a one-size-fits-all method. But rather than focusing only on the quantitative performance measures, quantitative measures must be balanced with qualitative factors of ease of interpretation, ease of computation, and the ability to scale to larger problems (Han *et al.* 2021). The study can give important information regarding the advantages and disadvantages of each framework, machine learning and deep learning, in sentiment analysis and can pave the way in exploring future experiments involving the use of a hybrid model that has a potential of having an interpretable structure as in the classical methodology yet uses a sophisticated feature-learning process like that of neural networks. This trend is promising with the potential of revolutionising future sentiment classification systems to be more flexible, precise and resource-consuming.

## 5.2 Linking with Objectives

The objectives described in the introduction of the present research were covered during the analysis, thereby correlating the research purpose with the results obtained. The first objective of ensuring that and enhancing Twitter information on airlines in the U.S. was satisfied by including, crafting rigorous preprocessing of information by cleaning information, eliminating symbols, irrelevant or disregarded information, missing values, and standardisation of text formats. This procedure improved the quality of the dataset, which allowed for to derivation of accurate and significant insights (Wang *et al.* 2021). The second objective, namely to apply exploratory data analysis (EDA) to tweets regarding airlines, has been completed by constructing the distribution of sentiments, frequently used words and tendencies of popular opinion. Visual representations like bar retrofits, word clouds, and plots of sentiment proportions gave clear observations to customer experience of the various airlines and gave a sense of which sentiment overrides the others and repeated matters on the tweets.

The third objective, which was to construct and train the models that categorise opinions presented in a review, was realised by the use of both deep learning (DL) and machine learning (ML) models. TF-IDsF vectorisation was applied to ML frameworks, including ***“Logistic Regression, Random Forest, and Support Vector Machines”***, and representations through embedding were applied in deep learning, including the LSTM. It is through these models that it managed to learn how to represent airline tweets in terms of positive, negative or neutral. Based on thoroughly swabbing evaluation measures such as ***“accuracy, precision, recall, F1-score and confusion matrices”***, the fourth objective, to analyse the quality of a classification model with appropriate measures, was thus accomplished. The comparisons between ML models and the DL models confirmed that the deep learning LSTM model had better performance, and it helped with capturing the contextual relationships in text, better than traditional ML methods (Ligthart *et al.* 2021).

The study fulfilled all the objectives that were mentioned, starting with enhancing and examining the data and continuing with constructing and testing sentiment classification models. The findings present a solid starting point towards further research that seeks to improve on the customer feedback analysis in the airline industry, and the companies will therefore be in a better position to analyse and react to passenger sentiments.

## 5.3 Recommendations

In line with such findings and analysis of this research work, there are a number of suggestions that can be provided to make sentiment classification systems of social media data, as regards the airline industry, better and make it more useful in practical applications within the aviation industry. To begin with, airlines need to invest in quality data collection and cleaning tools so that the data and messages received by the airlines via social media (in the form of tweets) are not noisy, are not comprised of irrelevant symbols, and are devoid of spam messages (Arici *et al.* 2023). Good quality data has a direct effect on the quality of sentiment classification models, allowing one to get more reliable information about customer opinions. This will run on automated preprocessing pipelines that identify and filter unnecessary or abusive content in real time.

Second, airline companies are expected to incorporate state-of-the-art models of sentiment analysis in their systems of customer service. According to the findings of this study, deep learning models and, most importantly, LSTM-based models are more effective at grasping the context and nuances of the text as compared to traditional machine learning models (Chandra and Krishna, 2021). The use of such models has the potential to assist airlines in recognising negative sentiments as they occur on a real-time basis so that the customer service team can react to the complaints, delay or service disturbance proactively. Moreover, one can use these models in multilingual support in order to spread their usage to the international markets.

Third, the issue of retraining models and enhancing them should be addressed regularly. Because the sentiments of the customers and the rise and fall of issues evolve, dynamic models may be efficient over time. Retraining models regularly by updating the dataset with fresh tweets will make the system in tandem with the current opinion of people. There is also a way for airlines to experiment with the hybrid solutions that would be pleasant and mixing the efficiency of traditional ML algorithms with deep learning.

Moreover, embedding results of the sentiment analysis into the processes of operating decision-making can bring great improvement to the service quality. To illustrate, sentiment analysis in real-time could send alerts to management in case there is a surge in negative reviews in cases when flights are delayed or cancelled, or they run into other service failures, allowing the management to take corrective measures fast. Positive sentiment insights may be applied to the marketing campaign, with elements of services that customers value mentioned (Kaur *et al.* 2021). Lastly, future studies should examine the use of multimodal data in analysis that includes data on images, video, and customer feedback that consists of voices as well as text. This would help get a more holistic account of the experience of customers. Additionally, airlines are advised to give attention to ethical AI, such as requirements of transparency, fairness, and safeguarding of data related to customers.

## 5.4 Future Work

Future research in the field of airline sentiment analysis needs to concentrate on increasing the depth and breadth of the proposed study in order to make the sentiment classification systems more accurate, responsive and tailored toward the industry. Among the major directions lies an increased number of social media platforms covered by the dataset. This will make sure that the model accounts for more diverse views and contexts of the customers. It is also vital to enjoy multilingual datasets because the airline industries encompass international operations, and the airlines must attend to reviews in various languages. This can be realised by designing a multi-lingual sentiment analysis model, which might be in the form of a transformer-type architecture such as BERT or XLM-R.

The other possible aspect that needs improvement is the multimodal sentiment analysis, which combines data, particularly the text, with other sources of information that include pictures, videos and even audio responses. The emotions of customers can go beyond words, and the combination of textual with visual or auditory ideas may result in a more thorough picture of the passenger experiences (Putra *et al.* 2023). In the same manner, it is possible to add metadata, which includes geolocation, flight information and timestamps, which can bring contextual richness, enabling sentiment trends to be correlated to route, events or problems to follow up on.

Technically, one could investigate more complex deep neural networks like transformer-based models (BERT, RoBERTa, and GPT models). This would mean much more context comprehension and classification precision. An alternative to investigate would be hybrid models that incorporate deep learning and rule-based models or other established ML methods to strike a compromise between interpretability and accuracy (Zhang *et al.* 2023). In addition, the sentiment classification can be brought to aspect-based sentiment analysis (ABSA), whereby the model is used to analyse sentiment towards particular elements of airline services, i.e., check-in (process), cabin crew, in-flight meals, or even punctuality of flights.

When it comes to practical use, the real-time sentiment monitoring systems are supposed to be created and incorporated into the airline operational dashboards. That would enable one to identify the presence of the negative feedback spikes in real-time, which would permit the rapid response plans and possible pre-avoidance of customer dissatisfaction development. An example of this can be seen at times of flight delays when the information can be used to send automated messages or compensation offers to the affected passengers in real-time. Finally, potential studies have to consider ethical and privacy issues (Cham *et al.* 2022). The customers should be able to rely on the transparency, consent, and adherence to data protection laws as more data is collected. Trust among the stakeholders can be enhanced by building an explainable AI system capable of giving a lucid reasoning on the sentiment prediction.

## 5.5 Limitations

The weaknesses of this research are mostly based on the quality of the data, scope, and modelling. To start with, although the size of the given dataset is massive, it is narrowed to the U.S. Micronews data, restricted to the collection of tweets on U.S. airlines within a purely defined period. This time limitation implies that the patterns of the sentiment that are recorded will not be generalizable to other times, particularly when there is an uncommon event such as a pandemic, strikes, or natural disasters (Rane *et al.* 2023). Moreover, the information can be subject to inherent biases like the over-representation of negative sentiments because of the kind of customer complaints in social media (Twitter). Such biases can affect the performance of models, causing biased predictions.

Secondly, sentiment analysis based on text naturally has problems with sarcasm, slang and context sensitivity. Despite the use of preprocessing methods, including tokenisation, stop word removal and TF-IDF vectorisation, the models may misinterpret any tweet that uses irony or ambiguity (Luo *et al.* 2021). Besides, it was not even considered to analyse emoji use and multimedia in tweets, which might have implied a valuable indication of sentiment.

From a modelling point of view, it should be noted that although the idea of both machine learning (ML) and deep learning (DL) models was tried, the model size limitations led to the inability to train the more complex models or did not permit a training of a large-scale transformer model like BERT. Also, the imbalance in sentiment classes might have affected precision even though sophisticated algorithms were adopted. Despite the meaningfulness of the study, its limitations should be addressed in future works.

# References

ACSI (2025). *Airlines*. [online] The American Customer Satisfaction Index. Available at: https://theacsi.org/industries/travel/airlines/.

Ahmed, A.A.A., Agarwal, S., Kurniawan, I.G.A., Anantadjaya, S.P. and Krishnan, C., 2022. Business boosting through sentiment analysis using Artificial Intelligence approach. International Journal of System Assurance Engineering and Management, 13(Suppl 1), pp.699-709.

Airports Council International (2025). *Joint ACI World-ICAO Passenger Traffic Report, Trends, and Outlook | ACI World*. [online] ACI World. Available at: https://aci.aero/2025/01/28/joint-aci-world-icao-passenger-traffic-report-trends-and-outlook/.

Ajah, I. and Nweke, H. (2019). Big Data and Business Analytics: Trends, Platforms, Success Factors and Applications. *Big Data and Cognitive Computing*, 3(2), p.32.

Alantari, H.J., Currim, I.S., Deng, Y. and Singh, S., 2022. An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews. International Journal of Research in Marketing, 39(1), pp.1-19.

Aljedaani, W., Rustam, F., Mkaouer, M.W., Ghallab, A., Rupapara, V., Washington, P.B., Lee, E. and Ashraf, I., 2022. Sentiment analysis on Twitter data integrating TextBlob and deep learning models: The case of US airline industry. Knowledge-Based Systems, 255, p.109780.

Alkhnbashi, O.S., Mohammad, R. and Hammoudeh, M. (2024). Aspect-Based Sentiment Analysis of Patient Feedback Using Large Language Models. *Big Data and Cognitive Computing*, 8(12), pp.167–167.

Alsayat, A., 2023. Customer decision-making analysis based on big social data using machine learning: a case study of hotels in Mecca. Neural Computing and Applications, 35(6), pp.4701-4722.

Althubaiti, A. (2016). Information bias in health research: Definition, pitfalls, and adjustment methods. *Journal of Multidisciplinary Healthcare*, 9(1), pp.211–217.

Anand (2024). *Explain the SERVQUAL technique and the service gap concept used in service industry to evaluate quality and customer satisfaction.* [online] Notes World. Available at: https://www.notesworld.in/2024/03/explain-servqual-technique-and-service.html [Accessed 27 Jun. 2025].

Arasli, H., Saydam, M.B., Gunay, T. and Jafari, K., 2021. Key attributes of Muslim-friendly hotels’ service quality: voices from booking. com. Journal of Islamic Marketing, 14(1), pp.106-127.

Arici, H.E., Cakmakoglu Arıcı, N. and Altinay, L., 2023. The use of big data analytics to discover customers’ perceptions of and satisfaction with green hotel service quality. Current Issues in Tourism, 26(2), pp.270-288.

Atandoh, P., Atandoh, P., Atandoh, P., Atandoh, P. and Atandoh, P. (2024). Scalable deep learning framework for sentiment analysis prediction for online movie reviews. *Heliyon*, pp.e30756–e30756.

Badanik, B., Remenysegova, R. and Kazda, A. (2023). Sentimental Approach to Airline Service Quality Evaluation. *Aerospace*, 10(10), p.883.

Behera, R.K., Jena, M., Rath, S.K. and Misra, S., 2021. Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data. Information Processing & Management, 58(1), p.102435.

Bellar, O., Baina, A. and Ballafkih, M. (2024). Sentiment Analysis: Predicting Product Reviews for E-Commerce Recommendations Using Deep Learning and Transformers. *Mathematics*, 12(15), pp.2403–2403.

Boot, A.B., Tjong Kim Sang, E., Dijkstra, K. and Zwaan, R.A. (2019). How character limit affects language usage in tweets. *Palgrave Communications*, 5(1). doi:https://doi.org/10.1057/s41599-019-0280-3.

Cham, T.H., Lim, Y.M. and Sigala, M., 2022. Marketing and social influences, hospital branding, and medical tourists' behavioural intention: Before‐and after‐service consumption perspective. International Journal of Tourism Research, 24(1), pp.140-157.

Chandra, R. and Krishna, A., 2021. COVID-19 sentiment analysis via deep learning during the rise of novel cases. PloS one, 16(8), p.e0255615.

Chang, V., Liu, L., Xu, Q., Li, T. and Hsu, C.H., 2023. An improved model for sentiment analysis on luxury hotel review. Expert Systems, 40(2), p.e12580.

Chang, Y.C., Ku, C.H. and Le Nguyen, D.D., 2022. Predicting aspect-based sentiment using deep learning and information visualization: The impact of COVID-19 on the airline industry. Information & Management, 59(2), p.103587.

Cheong, H., Lyons, A., Houghton, R. and Majumdar, A. (2023). Secondary Qualitative Research Methodology Using Online Data within the Context of Social Sciences. *International Journal of Qualitative Methods*, 22(1), pp.1–19.

Choi, Y. and Lee, H. (2017). Data properties and the performance of sentiment classification for electronic commerce applications. *Information Systems Frontiers*, 19(5), pp.993–1012.

Colmekcioglu, N., Marvi, R., Foroudi, P. and Okumus, F. (2022). Generation, susceptibility, and response regarding negativity: An in-depth analysis on negative online reviews. *Journal of Business Research*, 153(1), pp.235–250.

D’Aniello, G., Gaeta, M. and La Rocca, I. (2022). KnowMIS-ABSA: an overview and a reference model for applications of sentiment analysis and aspect-based sentiment analysis. *Artificial Intelligence Review*.

de Oliveira Júnior, G.A., de Oliveira Albuquerque, R., Borges de Andrade, C.A., de Sousa, R.T., Sandoval Orozco, A.L. and García Villalba, L.J. (2020). Anonymous Real-Time Analytics Monitoring Solution for Decision Making Supported by Sentiment Analysis. *Sensors*, 20(16), p.4557.

Dean, B. (2023). *How Many People Use Twitter in 2023? [New Twitter Stats]*. [online] Backlinko. Available at: https://backlinko.com/twitter-users.

Dhar, S. and Bose, I., 2022. Walking on air or hopping mad? Understanding the impact of emotions, sentiments and reactions on ratings in online customer reviews of mobile apps. Decision Support Systems, 162, p.113769.

Dublino, J. (2023). *Twitter as a Customer Support Channel*. [online] business.com. Available at: https://www.business.com/articles/twitter-as-a-customer-support-channel/.

Dwivedi, Y.K., Ismagilova, E., Hughes, D.L. and Carlson, J. (2021). Setting the future of digital and social media marketing research: Perspectives and research propositions. *International Journal of Information Management*, 59(1), pp.1–37.

Gomez-Adorno, H., Bel-Enguix, G., Sierra, G., Barajas, J.-C. and Álvarez, W. (2024). Machine Learning and Deep Learning Sentiment Analysis Models: Case Study on the SENT-COVID Corpus of Tweets in Mexican Spanish. *Informatics*, 11(2), p.24.

Han, J., Zuo, Y., Law, R., Chen, S. and Zhang, M., 2021. Service quality in tourism public health: trust, satisfaction, and loyalty. Frontiers in psychology, 12, p.731279.

Han, K.-X., Chien, W., Chiu, C.-C. and Cheng, Y.-T. (2020). Application of Support Vector Machine (SVM) in the Sentiment Analysis of Twitter DataSet. *Applied Sciences*, 10(3), p.1125.

Han, L., Mingying, H. and Peng, Z. (2025). The influence of social media marketing on the quality of hotel services and the behavioral intentions of tourists. *Acta Psychologica*, 255, p.104881.

Hansen, K.B. and Borch, C., 2022. Alternative data and sentiment analysis: Prospecting non-standard data in machine learning-driven finance. Big Data & Society, 9(1), p.20539517211070701.

Hassan, N. (2024). *Model Metrics Guide: F1, ROC, & More | Medium*. [online] Medium. Available at: https://medium.com/@nay1228/mastering-evaluation-metrics-f1-score-z-score-roc-curve-and-precision-recall-curve-cb175de25724.

IBM (2021). *Exploratory Data Analysis*. [online] Ibm.com. Available at: https://www.ibm.com/think/topics/exploratory-data-analysis.

Ilieva, G., Yankova, T., Ruseva, M., Dzhabarova, Y., Klisarova-Belcheva, S. and Bratkov, M. (2024). Social Media Influencers: Customer Attitudes and Impact on Purchase Behaviour. *Information*, 15(6), p.359.

Imani, M., Beikmohammadi, A. and Arabnia, H.R. (2025). Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels. *Technologies*, 13(3), p.88.

Jalil, Z., Abbasi, A., Javed, A.R., Badruddin Khan, M., Abul Hasanat, M.H., Malik, K.M. and Saudagar, A.K.J., 2022. COVID-19 related sentiment analysis using state-of-the-art machine learning and deep learning techniques. Frontiers in public health, 9, p.812735.

Jeon, W., Lee, Y. and Geum, Y., 2022. Airline service quality evaluation based on customer review using machine learning approach and sentiment analysis. Journal of Society for e-Business Studies, 26(4).

Kaggle (2020). *Twitter sentiment Analysis(Naive Bayes)*. [online] Kaggle.com. Available at: https://www.kaggle.com/code/pratikbarua/twitter-sentiment-analysis-naive-bayes [Accessed 27 Jun. 2025].

Kaur, H., Ahsaan, S.U., Alankar, B. and Chang, V., 2021. A proposed sentiment analysis deep learning algorithm for analyzing COVID-19 tweets. Information Systems Frontiers, 23(6), pp.1417-1429.

Khoo, K.L., 2022. A study of service quality, corporate image, customer satisfaction, revisit intention and word-of-mouth: evidence from the KTV industry. PSU Research Review, 6(2), pp.105-119.

Kim, T., Jo, H., Yhee, Y. and Koo, C., 2022. Robots, artificial intelligence, and service automation (RAISA) in hospitality: sentiment analysis of YouTube streaming data. Electronic Markets, 32(1), pp.259-275.

Kuoppakangas, P., Suomi, K., Stenvall, J., Pekkola, E., Kivistö, J. and Kallio, T. (2019). Revisiting the five problems of public sector organisations and reputation management—the perspective of higher education practitioners and ex-academics. *International Review on Public and Nonprofit Marketing*, 16(2-4), pp.147–171.

Li, Z. and Zou, Z. (2024). Punctuation and lexicon aid representation: A hybrid model for short text sentiment analysis on social media platform. *Journal of King Saud University. Computer and information sciences/Maǧalaẗ ǧamʼaẗ al-malīk Saud : ùlm al-ḥasib wa al-maʼlumat*, pp.102010–102010.

Ligthart, A., Catal, C. and Tekinerdogan, B., 2021. Systematic reviews in sentiment analysis: a tertiary study. Artificial intelligence review, 54(7), pp.4997-5053.

Lin, M.S., Liang, Y., Xue, J.X., Pan, B. and Schroeder, A., 2021. Destination image through social media analytics and survey method. International Journal of Contemporary Hospitality Management, 33(6), pp.2219-2238.

LinkedIn (2023). *What is Convolutional Neural Network — CNN (Deep Learning)*. [online] www.linkedin.com. Available at: https://www.linkedin.com/pulse/what-convolutional-neural-network-cnn-deep-learning-nafiz-shahriar.

Lu, L., Xu, P., Wang, Y.Y. and Wang, Y., 2023. Measuring service quality with text analytics: Considering both importance and performance of consumer opinions on social and non-social online platforms. Journal of Business Research, 169, p.114298.

Luo, J., Huang, S. and Wang, R., 2021. A fine-grained sentiment analysis of online guest reviews of economy hotels in China. Journal of Hospitality Marketing & Management, 30(1), pp.71-95.

Martin, G. (2014). *How To Complain To Airlines On Twitter (And What You’re Doing Wrong)*. [online] Forbes. Available at: https://www.forbes.com/sites/grantmartin/2014/10/24/how-to-complain-to-airlines-on-twitter-and-what-youre-doing-wrong/ [Accessed 27 Jun. 2025].

Mehraliyev, F., Chan, I.C.C. and Kirilenko, A.P., 2022. Sentiment analysis in hospitality and tourism: a thematic and methodological review. International Journal of Contemporary Hospitality Management, 34(1), pp.46-77.

Mehta, P., Pandya, S. and Kotecha, K., 2021. Harvesting social media sentiment analysis to enhance stock market prediction using deep learning. PeerJ Computer Science, 7, p.e476.

Narayanasamy, S.K., Hu, Y.-C., Qaisar, S.M. and Srinivasan, K. (2022). Effective Preprocessing and Normalization Techniques for COVID-19 Twitter Streams with POS Tagging via Lightweight Hidden Markov Model. *Journal of Sensors*, 2022, pp.1–14.

Naseem, U., Razzak, I., Khushi, M., Eklund, P.W. and Kim, J., 2021. COVIDSenti: A large-scale benchmark Twitter data set for COVID-19 sentiment analysis. IEEE transactions on computational social systems, 8(4), pp.1003-1015.

Oberlo (2021). *Find out How Many People Use Social Media in 2019 | Oberlo*. [online] Oberlo. Available at: https://www.oberlo.com/statistics/how-many-people-use-social-media.

Pakurár, M., Haddad, H., Nagy, J., Popp, J. and Oláh, J. (2019). The service quality dimensions that affect customer satisfaction in the Jordanian banking sector. *Sustainability*, 11(4), p.1113.

Park, J.Y., Mistur, E., Kim, D., Mo, Y. and Hoefer, R., 2022. Toward human-centric urban infrastructure: Text mining for social media data to identify the public perception of COVID-19 policy in transportation hubs. Sustainable Cities and Society, 76, p.103524.

Park, K. and Lee, S.S. (2024). U.S. Airline Customer Complaint Trends: A Decade-Long Analysis (2013-2022) Including COVID-19 Impacts. *International Business Research*, 17(5), p.131.

PJ, S., Singh, K., Kokkranikal, J., Bharadwaj, R., Rai, S. and Antony, J., 2023. Service quality and customer satisfaction in hospitality, leisure, sport and tourism: An assessment of research in web of science. Journal of Quality Assurance in Hospitality & Tourism, 24(1), pp.24-50.

Prentice, C. and Nguyen, M., 2021. Robotic service quality–Scale development and validation. Journal of Retailing and Consumer Services, 62, p.102661.

Putra, F.P.E., Maulana, F.I., Akbar, N.M. and Febriantoro, W., 2023. Twitter sentiment analysis about economic recession in indonesia. Bulletin of Social Informatics Theory and Application, 7(1), pp.1-7.

Rana, M.R.R., Nawaz, A., Rehman, S.U., Abid, M.A., Garayevi, M. and Kajanová, J. (2025). BERT-BiGRU-Senti-GCN: An Advanced NLP Framework for Analyzing Customer Sentiments in E-Commerce. *International Journal of Computational Intelligence Systems*, 18(1).

Rane, N., 2023. Enhancing customer loyalty through Artificial Intelligence (AI), Internet of Things (IoT), and Big Data technologies: improving customer satisfaction, engagement, relationship, and experience. Internet of Things (IoT), and Big Data Technologies: Improving Customer Satisfaction, Engagement, Relationship, and Experience (October 13, 2023).

Rodríguez-González, A., Juan Manuel Tuñas, Lucia Prieto Santamaría, Diego Fernández Peces-Barba, Ernestina Menasalvas Ruiz, Jaramillo, A., Cotarelo, M., Conejo, A.J., Arce, A. and Gil, A. (2020). Identifying Polarity in Tweets from an Imbalanced Dataset about Diseases and Vaccines Using a Meta-Model Based on Machine Learning Techniques. *Applied Sciences*, 10(24), pp.9019–9019.

Sabatová, J., Galanda, J., Adamčík, F., Jezný, M. and Šulej, R. (2016). Modern Trends in Airport Self Check-in Kiosks. *MAD - Magazine of Aviation Development*, 4(20), p.10.

Sajid, H. (2025). *How to Build an AI Sentiment Analysis Tool*. [online] Encord.com. Available at: https://encord.com/ai-sentiment-analysis-tool/ [Accessed 27 Jun. 2025].

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A. and Hall, P. (2022). Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*, 1270(1270).

Sprout Social (2025). *50+ of the Most Important Social Media Marketing Statistics for 2023*. [online] Sprout Social. Available at: https://sproutsocial.com/insights/social-media-statistics/.

Talaat, A.S. (2023). Sentiment analysis classification system using hybrid BERT models. *Journal of Big Data*, 10(1).

Tripathi, M., 2021. Sentiment analysis of nepali covid19 tweets using nb svm and lstm. Journal of Artificial Intelligence, 3(03), pp.151-168.

Tusar, M.T.H.K. and Islam, M.T., 2021, September. A comparative study of sentiment analysis using NLP and different machine learning techniques on US airline Twitter data. In 2021 International conference on electronics, communications and information technology (ICECIT) (pp. 1-4). IEEE.

Umer, M., Ashraf, I., Mehmood, A., Kumari, S., Ullah, S. and Sang Choi, G., 2021. Sentiment analysis of tweets using a unified convolutional neural network‐long short‐term memory network model. Computational Intelligence, 37(1), pp.409-434.

Us, Y., Pimonenko, T., Lyulyov, O., Chen, Y. and Tambovceva, T., 2022. Promoting green brand of university in social media: Text mining and sentiment analysis. Virtual Economics, 5(1), pp.24-42.

Wang, Z., Zhu, Z., Xu, M. and Qureshi, S., 2021. Fine-grained assessment of greenspace satisfaction at regional scale using content analysis of social media and machine learning. Science of the Total Environment, 776, p.145908.

Wankhade, M., Rao, A.C.S. and Kulkarni, C., 2022. A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7), pp.5731-5780.

Xie, H., Li, Y., Pu, Y., Zhang, C. and Huang, J., 2024. Evaluating airline service quality through a comprehensive text-mining and multi-criteria decision-making analysis. Journal of Air Transport Management, 120, p.102655.

Zhang, W., Deng, Y., Liu, B., Pan, S.J. and Bing, L., 2023. Sentiment analysis in the era of large language models: A reality check. arXiv preprint arXiv:2305.15005.

Zhang, Y., Lee, S.-Y. and Gu, Y. (2023). A review of air transport service quality studies: current status and future research agenda. *Journal of the Air Transport Research Society*, 1(1), pp.9–21.

Zygiaris, S., Hameed, Z., Ayidh Alsubaie, M. and Ur Rehman, S., 2022. Service quality and customer satisfaction in the post pandemic world: A study of Saudi auto care industry. Frontiers in psychology, 13, p.842141.

# Appendix

“import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from wordcloud import WordCloud

import re

import string

import warnings

warnings.filterwarnings("ignore")

from sklearn.model\_selection import train\_test\_split

from sklearn.feature\_extraction.text import TfidfVectorizer

from sklearn.linear\_model import LogisticRegression

from sklearn.naive\_bayes import MultinomialNB

from sklearn.preprocessing import LabelEncoder

from xgboost import XGBClassifier

from sklearn.metrics import confusion\_matrix, accuracy\_score

from sklearn.metrics import classification\_report

import tensorflow as tf

from tensorflow.keras.preprocessing.text import Tokenizer

from tensorflow.keras.preprocessing.sequence import pad\_sequences

from tensorflow.keras.models import Sequential

from tensorflow.keras.layers import Embedding, LSTM, Dense, Dropout

# Load the uploaded CSV file

file\_path = "Tweets.csv"

df = pd.read\_csv(file\_path)

df.head()

df.info()

df.shape

df.describe()

df.isnull().sum()

# Clean and preprocess tweets

def preprocess\_text(text):

text = text.lower()

text = re.sub(r"http\S+|www\S+|https\S+", '', text, flags=re.MULTILINE)

text = re.sub(r'\@w+|\#','', text)

text = re.sub(r"[^\w\s]", '', text)

text = text.translate(str.maketrans('', '', string.punctuation))

return text.strip()

df['clean\_text'] = df['text'].apply(preprocess\_text)

df.duplicated().sum()

df.drop\_duplicates(inplace=True)

# EDA: Sentiment distribution

plt.figure(figsize=(6, 4))

sns.countplot(data=df, x='airline\_sentiment', palette='Set2')

plt.title("Sentiment Distribution")

plt.xlabel("Sentiment")

plt.ylabel("Tweet Count")

plt.tight\_layout()

plt.show()

# EDA: Most frequent negative reasons

neg\_df = df[df['airline\_sentiment'] == 'negative']

plt.figure(figsize=(10, 5))

sns.countplot(y=neg\_df['airline'], order=neg\_df['airline'].value\_counts().index, palette='coolwarm')

plt.title("Negative Sentiment Tweets by Airline")

plt.xlabel("Count")

plt.ylabel("Airline")

plt.tight\_layout()

plt.show()

def show\_wordcloud(data, title):

text = " ".join(data)

wc = WordCloud(width=800, height=400, background\_color='white').generate(text)

plt.figure(figsize=(10, 5))

plt.imshow(wc, interpolation='bilinear')

plt.axis("off")

plt.title(title)

plt.tight\_layout()

plt.show()

show\_wordcloud(df[df['airline\_sentiment']=='positive']['clean\_text'], "Positive Tweets Word Cloud")

show\_wordcloud(df[df['airline\_sentiment']=='negative']['clean\_text'], "Negative Tweets Word Cloud")

# Split dataset

X = df['clean\_text']

y = df['airline\_sentiment']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# TF-IDF Vectorizer for ML models

tfidf = TfidfVectorizer(max\_features=5000)

X\_train\_tfidf = tfidf.fit\_transform(X\_train)

X\_test\_tfidf = tfidf.transform(X\_test)

# Logistic Regression

lr\_model = LogisticRegression(max\_iter=1000)

lr\_model.fit(X\_train\_tfidf, y\_train)

lr\_preds = lr\_model.predict(X\_test\_tfidf)

lr\_accuracy = accuracy\_score(y\_test, lr\_preds)

print(f"Accuracy: {lr\_accuracy:.4f}")

lr\_conf\_matrix = confusion\_matrix(y\_test, lr\_preds)

print(lr\_conf\_matrix)

lr\_class\_report = classification\_report(y\_test, lr\_preds)

print(lr\_class\_report)

# Naive Bayes

nb\_model = MultinomialNB()

nb\_model.fit(X\_train\_tfidf, y\_train)

nb\_preds = nb\_model.predict(X\_test\_tfidf)

nb\_accuracy = accuracy\_score(y\_test, nb\_preds)

print(f"Accuracy: {nb\_accuracy:.4f}")

nb\_conf\_matrix = confusion\_matrix(y\_test, nb\_preds)

print(nb\_conf\_matrix)

nb\_class\_report = classification\_report(y\_test, nb\_preds)

print(nb\_class\_report)

# Encode string labels into numeric

le = LabelEncoder()

y\_train\_encoded = le.fit\_transform(y\_train)

y\_test\_encoded = le.transform(y\_test)

# XGBoost Classifier

xgb\_model = XGBClassifier(use\_label\_encoder=False, eval\_metric='mlogloss')

xgb\_model.fit(X\_train\_tfidf, y\_train\_encoded)

xgb\_preds = xgb\_model.predict(X\_test\_tfidf)

# Decode predictions back to original labels

xgb\_preds = le.inverse\_transform(xgb\_preds)

xgb\_accuracy = accuracy\_score(y\_test, xgb\_preds)

print(f"Accuracy: {xgb\_accuracy:.4f}")

xgb\_conf\_matrix = confusion\_matrix(y\_test, xgb\_preds)

print(xgb\_conf\_matrix)

xgb\_class\_report = classification\_report(y\_test, xgb\_preds)

print(xgb\_class\_report)

# LSTM

# Parameters

max\_words = 5000

max\_len = 50

embedding\_dim = 64

batch\_size = 64

epochs = 5

y\_encoded = le.fit\_transform(y)

# Split dataset

X\_train\_seq, X\_test\_seq, y\_train\_seq, y\_test\_seq = train\_test\_split(df['clean\_text'], y\_encoded, test\_size=0.2, random\_state=42)

# Tokenization

tokenizer = Tokenizer(num\_words=max\_words, oov\_token="<OOV>")

tokenizer.fit\_on\_texts(X\_train\_seq)

X\_train\_pad = pad\_sequences(tokenizer.texts\_to\_sequences(X\_train\_seq), maxlen=max\_len, padding='post')

X\_test\_pad = pad\_sequences(tokenizer.texts\_to\_sequences(X\_test\_seq), maxlen=max\_len, padding='post')

# LSTM Model

model = Sequential([

Embedding(input\_dim=max\_words, output\_dim=embedding\_dim, input\_length=max\_len),

LSTM(64, return\_sequences=False),

Dropout(0.5),

Dense(3, activation='softmax') # 3 sentiment classes

])

model.compile(loss='sparse\_categorical\_crossentropy', optimizer='adam', metrics=['accuracy'])

model.summary()

# Train model

model.fit(X\_train\_pad, y\_train\_seq, epochs=epochs, batch\_size=batch\_size, validation\_split=0.1)

# Predict

y\_pred\_prob = model.predict(X\_test\_pad)

y\_pred = np.argmax(y\_pred\_prob, axis=1)

lstm\_accuracy = accuracy\_score(y\_test\_seq, y\_pred)

print(f"Accuracy: {lstm\_accuracy:.4f}")

print(confusion\_matrix(y\_test\_seq, y\_pred))

print(classification\_report(y\_test\_seq, y\_pred, target\_names=le.classes\_))

data = [ ['XGBoost Classifier', xgb\_accuracy],

['Naive Bayes', nb\_accuracy],['LSTM', lstm\_accuracy],

['Logistic Regression', lr\_accuracy]]

df = pd.DataFrame(data, columns=['Classifiers Name', 'Accuracy'])

# Set up the plot

plt.figure(figsize=(7, 4))

sns.barplot(

x='Accuracy',

y='Classifiers Name',

data=df,

palette=['#2ecc71', '#3498db', '#e74c3c', '#9b59b6'] # Green, Blue, Red, Purple

)

plt.title('Model Accuracy Comparison')

plt.xlabel('Accuracy Score')

plt.ylabel('Classifier')

plt.xlim(0, 1)

plt.tight\_layout()

plt.show()”