UNIVERSITY OF PLYMOUTH
School of Engineering, Computing and Mathematics

School of Engineering, Computing and Mathematic

PROJ518: MSc Dissertation and Research Skills

ANALYSING AIRLINE SERVICE EXCELLENCE THROUGH AI-POWERED NATURAL LANGUAGE PROCESSING OF CUSTOMER FEEDBACK: A CASE STUDY OF BRITISH AIRWAYS (BA)

Submitted by

ARINZE MARTINS EKWULUO

Student ID: 10883694

Program

ARTIFICIAL INTELLIGENCE

Submission Date

10th of September 2024

Abstract

British Airways, like many airlines, faces significant challenges in efficiently analysing and extracting actionable insights from large volumes of unstructured customer feedback. The current manual methods are time-consuming, error-prone, and do not fully leverage the potential insights that can be gained from customer reviews, leading to missed opportunities for enhancing service excellence and customer satisfaction. This study aimed to evaluate the effectiveness of various Natural Language Processing (NLP) techniques and machine learning models in predicting customer satisfaction and recommendation outcomes based on customer review data. The primary objective was to identify the best approaches for analysing customer feedback to inform data-driven decision-making. The study employed sentiment analysis tools, specifically TextBlob and VADER, to classify sentiments in customer reviews. TextBlob achieved a recall of 95% for positive sentiments but had a low overall accuracy of 67% due to a high number of false positives. In contrast, VADER provided a more balanced performance with an accuracy of 70%, reducing false positives while maintaining a low number of false negatives. Various machine learning models, including Support Vector Machine (SVM), XGBoost, and Long Short-Term Memory (LSTM), were then trained using VADER sentiment scores and TF-IDF features to predict customer satisfaction and recommendations. SVM, when combined with VADER sentiment scores, delivered the best performance with a 94% accuracy, while XGBoost also performed well, achieving a 93% accuracy. However, models using TF-IDF features showed slightly lower performance, and the LSTM model, while effective during training, exhibited potential overfitting with a validation accuracy of 87%. The findings indicated that customer satisfaction was skewed towards lower ratings, with aircraft types such as the Saab 2000 and Boeing 747 Family receiving higher ratings, and the Boeing 757 and Airbus A321 Family receiving lower ratings. The analysis also revealed that value for money was the most significant predictor of customer satisfaction and recommendations. The study recommends that airlines focus on enhancing perceived value, improving cabin staff service, and upgrading seat comfort and in-flight services to increase customer satisfaction. Also, ccontinuous monitoring and adaptation to seasonal variations in customer expectations to maintain high levels of customer satisfaction and recommendations.

Keywords: *Customer Feedback Analysis, Machine Learning Models, Natural Language Processing (NLP), Sentiment Analysis, VADER*

Table of Contents

Chapter 5

Chapter 6: Conclusion

List of Tables

List of Figures

Chapter One

1.1 Introduction

Airline services play a pivotal role in the global transportation industry. They facilitate rapid movement of people and goods across short and long distances, connecting states, countries and continents, and significantly contributing to economic development. The International Air Transport Association (IATA) reports that the airline industry supports 65.5 million jobs worldwide and contributes $2.7 trillion (USD) to the global economy, representing 3.6% of global GDP (Simon, 2022). In 2019, airlines transported more than 4.5 billion passengers. Although there was a temporary decline in numbers due to the COVID-19 pandemic, passenger traffic is expected to recover and continue growing. The air cargo sector is also crucial, with airlines moving roughly 52 million metric tons of goods each year, which represents about 35% of the global trade value (Dawod & Saeed Sharif, 2021). According to Samir et al. (2023), the airline industry operates in a hyper-competitive environment. With multiple carriers vying for passengers on similar routes, even minor service deficiencies can significantly impact customer loyalty. In this landscape, airlines that prioritize service excellence and continuously strive to improve the passenger experience stand out. For example, the (IATA,2023) report highlights the customer base and revenue of the top five airlines American Airlines, Delta Air Lines, United Airlines, Emirates, and British Airwaysleading the charge. Collectively, these airlines transported around 785 million passengers, representing approximately 15% of the global market. American Airlines topped the list with 215 million passengers and $54 billion in revenue, followed closely by Delta Air Lines and United Airlines, each with over 200 million passengers and revenues of $53 billion and $52 billion, respectively. Emirates and British Airways, while serving fewer passengers (80 million and 75 million, respectively), generated substantial revenues of $32 billion and $30 billion (IATA, 2023).

Across these airlines, passenger services accounted for the largest share of revenue, ranging from 58% to 65%. The strong performance of these top airlines in 2023 highlights the industry's resilience and the critical role of customer-centric strategies in driving growth and profitability (IATA,2023). Substandard airline services have a profound impact on customer satisfaction and the airline's reputation. For instance, frequent flight delays and cancellations, which affect as many as a quarter of passengers, often result in missed connections and disruptions to travel itineraries. Inadequate in-flight amenities, such as uncomfortable seating, poor-quality meals, and malfunctioning entertainment systems, are a concern for approximately 30% of travellers.

These issues underscore the importance of airlines maintaining high standards of service to meet passenger expectations and ensure a positive travel experience (Juliana, 2021).

Additionally, baggage handling issues, such as lost or delayed luggage, impact about 15% of customers, causing significant frustration and inconvenience. The transparency and efficiency of refund and compensation processes are also common complaints, affecting about 10% of passengers who struggle to receive timely and fair resolutions(Lane, 2014). These service deficiencies not only diminish customer satisfaction but also lead to decreased loyalty, negative reviews, and potential loss of revenue as dissatisfied customers choose competitors with better service records. Addressing these issues is crucial for airlines to maintain a positive reputation and ensure long-term customer loyalty. Even the top five airlines as stated above were not immune to the effects of poor services, which significantly impacted customer satisfaction. According to Rijitha (2023), flight delays and cancellations were a major issue, affecting around 20% of passengers across these airlines, leading to considerable frustration and logistical challenges. Inadequate in-flight amenities, such as uncomfortable seating and substandard meal options, were reported by approximately 25% of travellers, highlighting areas in need of improvement. Customer service inefficiencies, including unresponsive or unfriendly staff, were a common complaint for about 18% of passengers, contributing to negative travel experiences (Sumitha K, 2023).

People's opinions and experiences serve as essential sources of information in our daily lives. Social media platforms are widely used for sharing comments and opinions about various products and services (Ahmad et al., 2018). For large organizations and companies, understanding customer or user feedback and sentiment is crucial. Consequently, companies are increasingly recognizing the importance of incorporating social media comments into their marketing strategies. Customer feedback holds significant importance across industries, particularly in sectors like aviation. In the airline industry, where service quality and customer experience are paramount, feedback from passengers play a pivotal role in shaping service quality and passenger experiences (Kim & Kusakci, 2023). Airlines rely on feedback to gauge satisfaction levels, identify areas for improvement, and address concerns promptly. By analysing feedback, airlines can pinpoint recurring issues such as flight delays, baggage handling problems, or service deficiencies, allowing them to implement targeted solutions and enhance overall operational efficiency (Sumitha, 2023).

Moreover, customer feedback serves as a barometer of performance against industry standards and competitors, guiding strategic decisions and resource allocations. Beyond operational benefits, actively seeking and acting upon feedback demonstrates a commitment to customer-centricity, fostering trust and loyalty among passengers. Ultimately, leveraging customer feedback not only improves service delivery but also strengthens an airline's reputation, positioning it competitively in a highly demanding market. The airline industry is highly competitive, and providing excellent customer service is crucial for ensuring customer satisfaction and loyalty. Traditionally, airlines relied on surveys and basic customer service interactions to gauge customer satisfaction. However, these methods often yielded limited and subjective data. The rise of digital communication has fundamentally transformed how airlines can gather and analyse customer feedback (Ban & Kim, 2019). Social media platforms, online review sites, and in-flight entertainment systems generate a vast amount of textual data in the form of reviews, comments, and complaints. This presents a unique opportunity for airlines to leverage cutting-edge Natural Language Processing (NLP) technology to extract valuable insights and identify areas for improvement (Juliana, 2021).

The Natural Language Processing (NLP) is a field of Artificial Intelligence (AI) that deals with the interaction between computers and human language. It aims to equip computers with the ability to understand the complexities of human language. NLP focuses on processing large amounts of text data from various sources like online reviews, social media comments, customer surveys, documents and emails and chat logs. The goal for NPL is extract valuable information from textual data, these can include identifying themes and trends, sentiment analysis and entity recognition (Hyun et al., 2019).

1.2 Problem Statement

British Airways is a major airline carrier based in the United Kingdom. It is one of the world's oldest airlines, having been founded in 1974. British Airways is a full-service global airline with its main hub at London Heathrow Airport (Lane, 2014). The combination of service excellence, global connectivity, customer-centric approach, and sustainability efforts positions British Airways as a preferred choice for discerning travellers looking for reliability, comfort, and a superior travel experience. Despite the availability of customer feedback data, British Airways, like many other airlines, struggles to analyse and extract meaningful insights from the abundance of customer feedback they receive. The current manual processes are time-consuming, error-prone, and do not fully utilize the potential insights hidden in the unstructured text data. There is a pressing need to leverage AI-powered NLP techniques to efficiently process and analyse customer feedback to enhance service excellence and customer satisfaction.

To achieve the above, our dataset will be collected from various sources which include airline review websites, social media pages, travel forums, and digital interfaces. The collected data comprises textual reviews and scores related to seat comfort, staff service, in-flight entertainment, and overall customer satisfaction. Selection of data sources is based on relevance and volume to ensure a comprehensive dataset representing diverse customer experiences across geographical regions. Again, this research will explore how AI-powered NLP (NLP) can be used to analyse customer feedback for British Airways to evaluate service excellence.

1.3 Aim and Objectives

The study aims to analyse and evaluate customer feedback using AI-powered NLP techniques to enhance service excellence at British Airways (BA). The specific objectives are to:

i. Analyse customer reviews sentimentally using NLP techniques.
ii. Perform Exploratory Data Analysis (EDA)to uncover patterns, relationships, and insights within the customer review data
iii. Develop a model to evaluate and compare the performance of various machine learning models predicting customer recommendations and satisfaction based on VADER sentiment scores and TF-IDF features
iv. Develop a Deep Learning model to predict customer satisfaction and recommendation outcomes.

1.4 Research Questions

i. How effective are NLP techniques, such as VADER sentiment analysis and TextBlob, in accurately identifying and classifying the sentiment expressed in customer reviews?

ii. What key patterns, relationships, and insights can be uncovered through Exploratory Data Analysis (EDA) of customer review data, and how do these factors influence overall customer satisfaction and recommendations?

iii. How do different machine learning models, utilizing VADER sentiment scores and TF-IDF features, compare in predicting customer satisfaction and recommendation outcomes, and which model demonstrates the best performance?

iv. What is the performance of a Deep Learning model, such as an LSTM network, in predicting customer satisfaction and recommendation outcomes based on customer review data, and how does it compare to traditional machine learning models?

1.5 Obstacles and Limitations

The several obstacles and limitations faced during my research include:

i. Customer feedback often contains slang, idiomatic expressions, and varied emotional tones, making it difficult for NLP algorithms to accurately interpret the context and sentiment.
ii. Current NLP models may struggle with understanding complex linguistic nuances and multi-language feedback, which can limit the accuracy of sentiment analysis.
iii. Handling sensitive customer data requires stringent privacy and security measures to comply with regulations such as GDPR.
iv. Translating NLP-derived insights into actionable strategies can be challenging, particularly if the insights are vague or require significant changes in operations.

1.5 Structure of Work

This thesis is structured to provide a comprehensive understanding of the research topic and its significance. Chapter One introduces the topic and the rationale for the study, setting the stage for the investigation. Chapter Two offers a critical review of relevant literature, examining various methodologies and theories from different authors to justify the research aim.

In Chapter Three, the methodology is explored in detail, supported by diagrams and design elements to illustrate the approach. Chapter Four presents the research results clearly, using tables and graphs to document the findings, accompanied by interpretations and a thorough discussion. Chapter Five discussed the research findings and compared it with past literatures. Finally, Chapter Six summarizes the study and its conclusions, highlighting key results and comparing them with related works discussed in Chapter Two. This chapter also outlines potential areas for future research.

Literature Review

2.1 Introduction

This section presents the review of related literature in this study. It explains the various concepts used in this study: the airline industry, British Airways, customer feedbacks and AI-powered NLP. This chapter also discusses the theoretical review NLP as well as presenting the review of related literature by different authors, analysing what they did, method used and their results. The section also presents the gaps filled and contribution of this study based on the analysis of the reviewed literatures. Finally the conclusion detailing every subsections of this chapter will be presented.

2.2 The Airline Industry

The airline industry is highly competitive, with service excellence being a key differentiator among carriers. The airline industry is a vital component of global transportation, connecting people and goods across vast distances. As of 2019, the global airline industry was valued at approximately $838 billion, with passenger traffic reaching 4.54 billion (IATA, 2019). This growth was driven by increasing demand for air travel, economic growth, and the expansion of low-cost carriers (LCCs) which made air travel more accessible. One significant trend in the airline industry is the rise of LCCs. These carriers have expanded their market share by offering no-frills services at lower prices, appealing to cost-conscious travellers. For instance, LCCs accounted for nearly 30% of global air travel in 2019, up from 15% in 2006 (CAPA, 2019). This shift has forced traditional full-service airlines to adapt by unbundling services and offering basic economy fares to remain competitive.

Technological advancements have also played a crucial role in shaping the airline industry. Innovations in aircraft design, such as the use of lightweight materials and more efficient engines, have improved fuel efficiency and reduced operating costs. For example, the introduction of the Boeing 787 Dreamliner and the Airbus A350 has enabled airlines to operate long-haul flights more economically (Boeing, 2019). Sustainability has become a major focus for the industry, driven by increasing environmental concerns and regulatory pressures. Airlines have committed to reducing their carbon footprint through various initiatives, including investing in more fuel-efficient aircraft, implementing carbon offset programs, and exploring alternative fuels. The International Air Transport Association (IATA) set a goal for the airline industry to achieve carbon-neutral growth from 2020 onwards and to halve carbon emissions by 2050 compared to 2005 levels (IATA, 2019).

Another trend is the increasing use of data analytics and artificial intelligence (AI) to enhance operational efficiency and customer experience. Airlines are leveraging big data to optimize flight routes, improve fuel management, and offer personalized services to passengers. AI-powered chatbots and virtual assistants are being used to handle customer inquiries, streamline check-in processes, and provide real-time updates on flight status (Airlines International, 2018). However, the industry faces challenges, including economic volatility, fluctuating fuel prices, and geopolitical tensions. These factors can significantly impact airline profitability and operational stability. For instance, fuel costs, which constitute a major portion of airline expenses, are subject to market fluctuations, affecting the financial performance of airlines (ICAO, 2019).

2.2.1 British Airways

British Airways (BA), established in 1974, is one of the world's premier airlines and a major player in the global aviation industry. With its primary hub at London Heathrow Airport, BA serves over 180 destinations in more than 80 countries, offering a range of services from economy to luxury first-class travel (British Airways, 2019). In 2019, BA carried approximately 45 million passengers, marking it as one of the busiest airlines in the world (IAG, 2019). The airline has consistently focused on enhancing service quality through innovation and a strong commitment to customer satisfaction. Key areas of investment include upgrading its fleet with newer, more fuel-efficient aircraft such as the Airbus A350 and Boeing 787 and revamping its cabin interiors to provide a more comfortable and luxurious travel experience (Airbus, 2019). BA has also embraced digital transformation to improve customer service. The introduction of AI-powered chatbots and enhanced mobile applications has streamlined customer interactions, making it easier for passengers to book flights, check in, and receive real-time updates (IAG Annual Report, 2019).

Sustainability is another focus area for BA. The airline has committed to achieving net-zero carbon emissions by 2050 and has implemented various initiatives to reduce its environmental impact. These include investing in sustainable aviation fuels, improving operational efficiency, and participating in carbon offset programs (British Airways, 2019). Despite these advancements, BA faces challenges like other major carriers, including fluctuating fuel prices and economic uncertainties. However, its strategic focus on innovation and customer-centric services positions it well to navigate these challenges and maintain its competitive edge in the global airline industry (IATA, 2019).

2.3 Customer Feedback

Customer feedback is a critical asset for airlines aiming to enhance service quality and meet passenger expectations. Feedback from passengers covers various aspects such as seat comfort, staff behaviour, in-flight entertainment, punctuality, and overall travel experience. Traditionally, airlines have collected this feedback through surveys, complaint forms, and direct interactions with customers. These methods, while valuable, are limited by their scope and the effort required to gather and process the information. The proliferation of digital platforms has dramatically increased the volume and variety of customer feedback available to airlines. Passengers now share their experiences across social media, review websites, travel forums, and mobile applications.

For example, platforms like TripAdvisor and Skytrax host millions of airline reviews, while social media sites like Twitter and Facebook see a constant stream of real-time passenger feedback (Statista, 2019). This shift towards digital feedback has created a vast amount of unstructured data, presenting significant challenges for manual analysis.

Manual analysis of this unstructured data is both time-consuming and labor-intensive. Analysts must sift through countless reviews, comments, and posts to identify trends and actionable insights. This process is not only prone to human error but also often fails to capture the full spectrum of insights hidden within the data. Consequently, many airlines, including British Airways, struggle to efficiently analyse and utilize this wealth of information to drive service improvements (IATA, 2019). The limitations of manual analysis have spurred interest in advanced technologies, particularly in the realm of artificial intelligence (AI). AI-powered NLP (NLP) techniques offer a promising solution to these challenges. NLP, a subfield of AI, involves the development of algorithms and models that enable machines to understand, interpret, and generate human language. By leveraging NLP, airlines can automate the process of analysing large volumes of textual feedback, extracting meaningful insights, and identifying patterns and trends that may not be immediately apparent through manual analysis (Gartner, 2019).

The application of NLP in the airline industry has shown promising results. Sentiment analysis can gauge customer satisfaction levels and predict future behaviours, enabling airlines to proactively address issues. Topic modelling can reveal key areas of concern, allowing airlines to prioritize service improvements. These technologies not only improve the efficiency of feedback analysis but also enhance the accuracy and depth of insights obtained (Lee et al., 2019). British Airways, like many other airlines, can benefit significantly from implementing

AI-powered NLP techniques. By automating the analysis of customer feedback, BA can more effectively identify trends and areas for improvement, leading to better service quality and increased customer satisfaction. For example, if sentiment analysis reveals a recurring issue with seat comfort on specific routes, BA can take targeted actions to address this feedback.

2.4 Artificial Intelligence (AI) in Aviation

Artificial Intelligence (AI) is revolutionizing the aviation industry by enhancing operational efficiency, improving customer experience, and ensuring safety. AI technologies, such as machine learning, NLP, and computer vision, are being integrated into various aspects of airline operations. One significant application of AI in aviation is in predictive maintenance (Merlo, 2024). Airlines are utilizing AI algorithms to analyse data from aircraft sensors to predict potential mechanical issues before they occur, reducing downtime and maintenance costs (Helgo, 2023). This proactive approach enhances safety and operational reliability. AI-powered NLP is transforming customer service by automating responses to passenger inquiries. Chatbots and virtual assistants provide instant support for booking, check-in, and flight status updates, improving customer satisfaction and reducing the workload on human agents (Ceken, 2024). Airlines like KLM and Lufthansa have implemented AI-driven chatbots to handle customer interactions more efficiently (Gentsch & Gentsch, 2019).

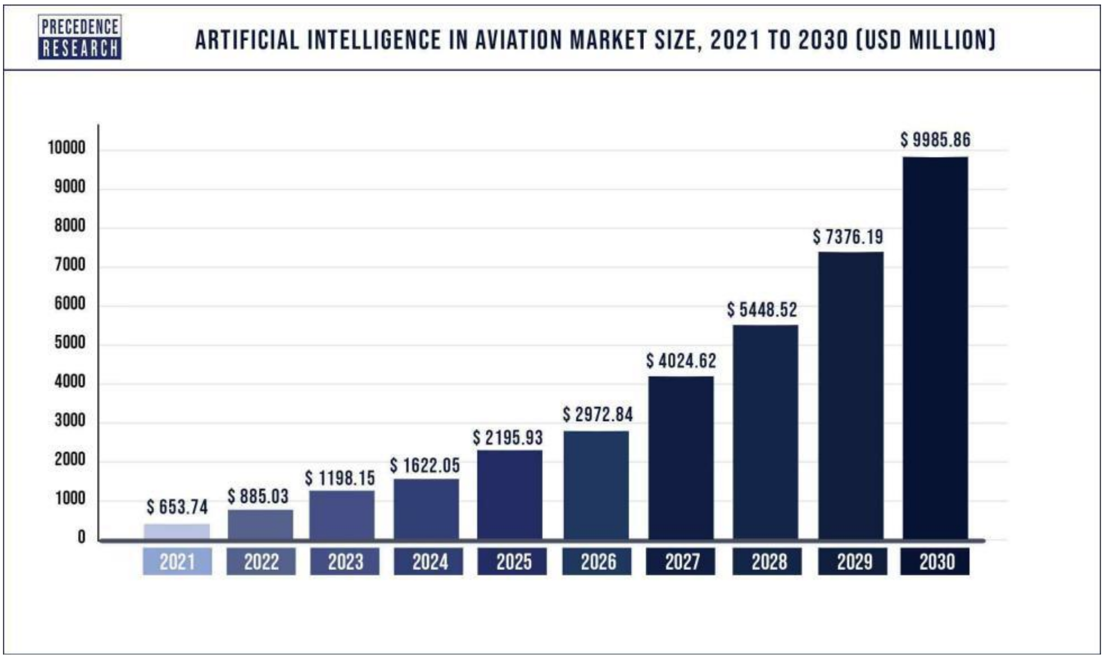


Figure 2.1: AI in Aviation Market (Precedence Research, 2022)

The Charts from Precedence Research illustrates the projected growth of the Artificial Intelligence (AI) market in the aviation industry from 2021 to 2030. Starting at $653.74 million in 2021, the market is expected to grow steadily, reaching $985.03 million in 2022 and continuing to rise each year. By 2025, the market is projected to hit $2,972.84 million, and by 2030, it is forecasted to reach $9,985.88 million. This exponential growth underscores the increasing adoption and reliance on AI technologies within the aviation sector, driven by the need for enhanced efficiency, safety, and customer service. Another trend is the use of AI in flight operations. AI systems are optimizing flight paths by analysing weather patterns, air traffic, and other variables, leading to more efficient fuel usage and reduced flight times (Zhu & Li, 2021). This not only lowers operational costs but also contributes to environmental sustainability. In terms of enhancing the passenger experience, AI is being used to personalize services. By analysing customer data, airlines can offer tailored recommendations for in-flight entertainment, meal preferences, and loyalty programs (Martins et al., 2024). This personalization enhances passenger satisfaction and loyalty. Furthermore, AI is improving security measures at airports. Facial recognition and other biometric technologies streamline the boarding process and enhance security checks, providing a smoother and more secure travel experience (Patel, 2018; Azman & Sharma, 2022).

Figure 2.2: AI Biometrics at Airport

AI is driving significant advancements in the aviation industry by improving efficiency, safety, and customer satisfaction. As AI technology continues to evolve, its integration into aviation operations is expected to deepen, bringing about more innovative solutions and transforming the industry.

2.4 Theoretical Review

AI-powered NLP

AI-powered NLP techniques offer a promising solution to the challenges of analysing large volumes of customer feedback. NLP, a subfield of AI, focuses on the interaction between computers and human language. It involves developing algorithms and models that enable machines to understand, interpret, and generate human language (Fanni et al., 2023; Parde, 2023). NLP combines computational linguistics, computer science, and artificial intelligence to process and analyse large amounts of natural language data

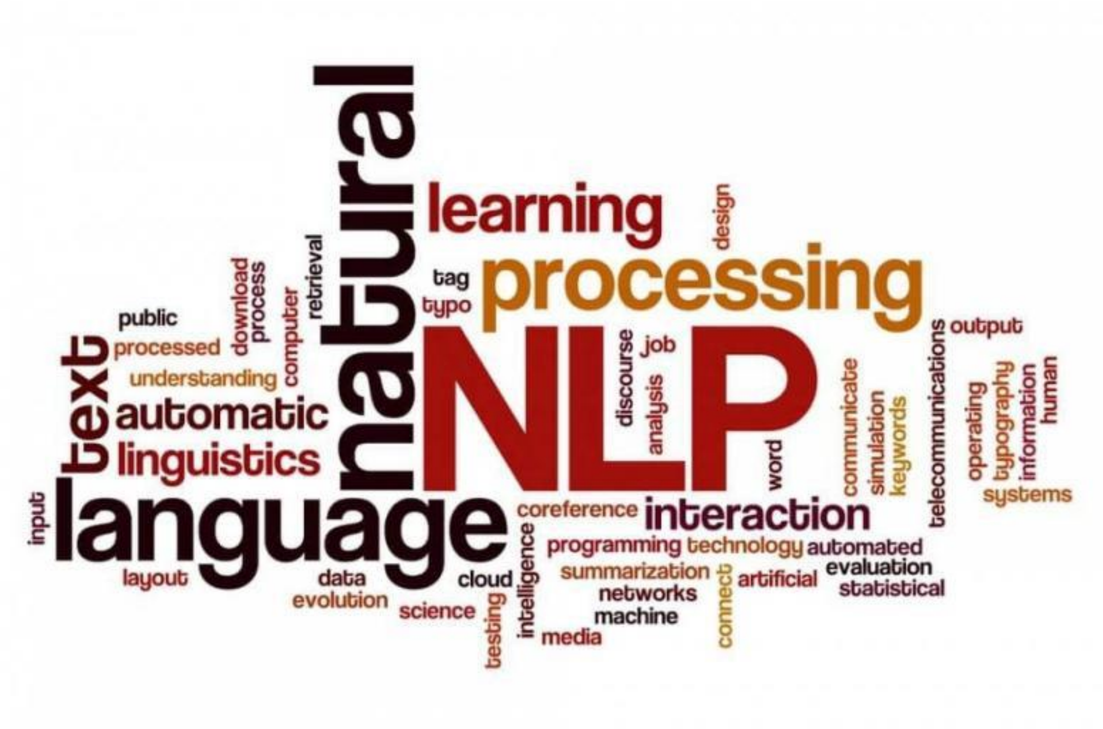


Figure 2.3: Natural Language Processing (NLP) (Jaiswal, 2022)

By leveraging NLP, airlines can automate the process of analysing vast amounts of textual feedback, extracting meaningful insights, and identifying patterns and trends that may not be immediately apparent through manual analysis (Hasib, 2022). The application of NLP in analysing customer feedback involves several key components:

i. Tokenization: This is the process of breaking down text into individual words or phrases. Tokenization is the foundational step in NLP, making it easier to analyse and process text data. For instance, breaking down a review into tokens allows the system to understand the frequency and context of specific words or phrases (Frank et al., 2024).

ii. Part-of-Speech Tagging: This technique identifies the grammatical parts of speech in a sentence, such as nouns, verbs, adjectives, etc. Understanding the parts of speech helps in grasping the context of the feedback. For example, recognizing whether a word is an adjective describing a service (e.g., "excellent" service) or a noun referring to a specific aspect (e.g., "staff") (Sodhar et al., 2019).

iii. Named Entity Recognition (NER): NER extracts entities such as names, dates, and locations from text, providing more specific insights. In the context of airline feedback, NER can identify specific flights, destinations, and staff members mentioned in reviews, which helps in pinpointing exact areas needing attention (Yadav & Bethard, 2019).

iv. Sentiment Analysis: This determines the sentiment expressed in a piece of text, categorizing it as positive, negative, or neutral. Sentiment analysis allows airlines to gauge overall customer satisfaction and identify specific aspects that elicit strong emotions, whether positive or negative (Idris & Mohammed, 2023).

v. Topic Modelling: This identifies the main topics discussed in a collection of documents, highlighting common themes and concerns among passengers. Topic modelling can reveal recurring issues like delays, staff behaviour, or seat comfort, helping airlines prioritize their service improvements (Kuhn, 2018).

2.5 Review of Related Works

Customer sentiment analysis and predictive analytics play crucial roles in enhancing customer experience and loyalty within the airline industry. Various studies have leveraged different techniques and datasets to analyse customer reviews, sentiments, and behaviours, providing valuable insights for airlines to improve their services.

Annamalai et al. (2024) utilized VADER (Valence Aware Dictionary and Sentiment Reasoner) to perform sentiment analysis on customer reviews, aiming to predict buying behaviour in the airline industry, with a specific focus on British Airways. They collected data through web scraping from the Airline Quality website, processing and cleaning the unstructured data for analysis. VADER was used to classify sentiments as positive, negative, or neutral, providing insights into customer sentiment towards British Airways. Additionally, a classification model was developed using various features such as flight routes, booking origins, trip types, number of passengers, and customer preferences. The study employed Random Forest, XGBoost, Logistic Regression, and KNN algorithms to train and evaluate the models. The predictive model helps British Airways tailor their services to meet customer expectations, improving satisfaction and increasing bookings.

Shahid et al. (2024) explored the likelihood of customers revisiting airline services using machine learning techniques, emphasizing emotional connections and loyalty. They analysed feedback comments and satisfaction ratings, employing sentiment analysis and the Linguistic Inquiry and Word Count (LIWC) methodology to categorize sentiments. Their comprehensive data collection involved an initial survey of 17,000 responses and a follow-up survey one year later. Multiple classifiers, including Decision Tree, Random Forest, and XGBoost, were evaluated using five-fold cross-validation. XGBoost achieved the highest accuracy (85%) in predicting return visits, demonstrating the predictive power of machine learning in understanding customer behaviour.

Arpita et al. (2023) conducted sentiment analysis on a dataset of 67,993 reviews from the Google Play Store and App Store, covering the ten most well-known airlines. The reviews were categorized as “Positive,” “Negative,” or “Neutral.” The study integrated word embedding with deep learning models such as CNN, LSTM, and BiLSTM to enhance accuracy. Both LSTM and BiLSTM models achieved high accuracy rates of 90% and 91%, respectively. BiLSTM outperformed other models in terms of precision (92%), recall (91%), and F1-score (91%), showcasing its effectiveness in sentiment analysis.

Nche (2023) focused on sentiment analysis of customer complaints about Brussels Airlines, analysing 2,259 TripAdvisor reviews from January 2017 to November 2022. The study only considered English reviews with ratings of "average," "poor," and "terrible." Using the Design Science Research Methodology (DSRM), the research combined the analysis of customer complaints and their sentiments. Azure Machine Learning was used for sentiment analysis, while MAXQDA software identified prominent customer complaints. The findings revealed that luggage issues, flight delays, flight cancellations, and food quality were the most frequent complaints, indicating overall customer dissatisfaction with Brussels Airlines' services. Samir et al. (2023) employed NLP, text analysis, biometrics, and computational linguistics to analyse Skytrax Airline Customers' Feedback (SACF) data. They used deep learning models to address sentiment analysis problems, applying Glove embedding models for feature extraction. A comparative study evaluated various models, including RNN, LSTM, GRU, 1D CNN, and BERT. The LSTM model outperformed others with a classification accuracy of 91%. The study highlighted the effectiveness of deep learning techniques in improving sentiment classification performance.

Nwakanma et al. (2019) conducted predictive analytics using hotels.ng as a case study, focusing on sentiment extraction from heterogeneous data sources to inform hoteliers' and tourists' decision-making processes. Using a NLP System built in JAVA, they classified respondents' opinions as positive or negative. A survey questionnaire validated customers' preferences. The results showed that room setting, good security, good customer service, cleanliness, price of hotel service, and hotel features were critical factors influencing customer satisfaction. The analysis indicated a positive effect of customer sentiments on the growth of the Nigerian hospitality industry. Hasib (2022) analysed online reviews for four major Bangladesh airlines, performing multiclass sentiment analysis and topic modelling to improve decision-making based on customer experiences. The study involved pre-processing data and balancing it using the Pegasus model's oversampling mechanism. Sentiment analysis was conducted using three machine learning classifiers (Decision Tree, Random Forest, and XGBoost) and three deep learning classifiers (CNN, LSTM, and BERT). The best accuracy of 83% was achieved using BERT. The sentiment analysis results were categorized into positive, negative, and mixed sentiments for domestic, international, and overall routes (Hasib, 2022).

Rane & Kumar (2018) analyses tweets from six major US airlines using multi-class sentiment analysis. Tweets were pre-processed and represented as vectors using Doc2vec for phrase-level analysis. Seven classification strategies, including Decision Tree, Random Forest, SVM, K-Nearest Neighbours, Logistic Regression, Gaussian Naïve Bayes, and AdaBoost, were evaluated. The results showed the overall sentiment (positive, negative, neutral) for the tweets, providing insights into customer feedback. Prabhakar et al. (2018) used a new improved Adaboost approach for sentiment analysis of US airline tweets. Various machine learning algorithms were employed, with performance analysed based on confusion matrix and accuracy. The study aimed to help customers make decisions about airlines based on sentiment analysis of online reviews and tweets.

Park et al. (2019) investigated the determinants of customer satisfaction with airline services through sentiment analysis of over 133,000 customer feedbacks. Using structural equation modelling, it was found that affective values significantly impact customer satisfaction. Differences between low-cost and full-service carriers were also analysed.

Das et al. (2017) used Naive Bayes algorithm to classify sentiments (positive, negative, neutral) of Twitter messages about airlines. Data was pre-processed and analysed using R and Rapid Miner, demonstrating the utility of Twitter data in sentiment analysis for the airline industry. Monica et al. (2019) used LSTM and RNN models to analyse sentiment polarity of tweets about six US airlines. Word embedding models (Word2Vec, Glove) were used, and the results showed significant classification accuracy, indicating the reliability of these models for future predictions. Punel et al. (2019) analysed 40,510 passenger reviews to understand variations in service quality expectations across different geographical regions and flight classes. Sentiment score analysis and path analysis methods were used, revealing regional differences in customer satisfaction and expectations

2.6 Gap and Research Contribution

Despite extensive research on sentiment analysis and customer feedback within the airline industry, significant gaps remain. Many studies have analysed various airlines, but few have specifically focused on British Airways. Although Annamalai et al. (2024) examined BA, there is a need for more comprehensive research using diverse NLP techniques. Existing studies often rely on specific NLP methods like VADER or individual machine learning algorithms, as seen in Shahid et al. (2024) and Hasib (2022). This highlights a gap in research that integrates multiple NLP techniques and machine learning algorithms for a more robust analysis of customer feedback. Additionally, most studies take a general approach to sentiment analysis without delving deeply into specific services offered by airlines, as noted in Arpita et al. (2023) and Nche (2023). There is a need for research that evaluates these services on a quality scale, particularly for British Airways. Furthermore, while some studies explored feedback for different flight routes, as seen in Hasib (2022), there is a lack of comprehensive analysis that includes customer opinions on specific flight routes and aircraft types. Few studies provide concrete recommendations based on their findings (Rane & Kumar 2018; Samir et al., 2023). This study aims to address these gaps by providing a comprehensive sentiment analysis focused exclusively on British Airways. By utilizing AI-powered NLP techniques, the study will gather and analyse customer feedback with greater precision. It will integrate various NLP techniques and machine learning algorithms, offering a comparative analysis that highlights the strengths and weaknesses of different methods. The research will also evaluate specific services offered by British Airways, providing a detailed quality scale analysis to identify areas of excellence and those needing improvement. Furthermore, the study will include a thorough analysis of customer feedback related to different flight routes and aircraft types used by British Airways, offering insights into specific operational areas. Based on these NLP-driven insights, the study will provide actionable strategies for British Airways to enhance service quality and improve customer experiences. These recommendations will be practical and data-driven, ensuring they are tailored to the airline's specific needs. This focused analysis, detailed service evaluation, and strategic recommendations will significantly contribute to enhancing service quality and customer experience at British Airways.

2.7 Conclusion

The literature review highlights critical gaps in sentiment analysis and customer feedback research within the airline industry, particularly regarding British Airways (BA). The airline industry is highly competitive, with service excellence being a key differentiator. Technological advancements, including AI-powered NLP, have revolutionized feedback analysis, offering tools like tokenization, part-of-speech tagging, named entity recognition, sentiment analysis, and topic modelling. These techniques enhance the efficiency and accuracy of analysing vast amounts of unstructured data from digital platforms. Despite the advancements, existing studies generally do not delve deeply into specific services offered by airlines or provide detailed feedback on flight routes and aircraft types. Different studies have explored various sentiment analysis techniques using NLP and machine learning. However, these studies often focus on specific airlines or limited NLP methods like VADER and LIWC, lacking comprehensive integration of multiple techniques. This study aims to fill these gaps by employing a multi-technique NLP approach to analyse BA's customer feedback comprehensively.

Research Methodology

3.1 Introduction

This research focused on the utilisation of AI-powered NLP techniques to analyse customer feedback for British Airways, deriving sentiments and preferences to guide service improvement strategies and enhance overall customer satisfaction. The dataset consists of customer reviews from various popular airlines, collected between 2011 and 2023. It includes reviews from 3,700 customers and features 19 columns of information. These columns encompass various aspects of the customer experience, providing a comprehensive view of their satisfaction and preferences.

3.2 Requirement Specification

Hardware Minimum Requirements: The minimum hardware requirements pertain to the physical features of the machine required. The following are the features: at least 500 GB HDD, Windows 10, 8 GB RAM, and Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz 2.71 GHz.

Software Requirements: These are the computer programmes and procedures needed to put the chatbot into action. The tools used include

i. Python: Programming language used.

ii. IDE is PyCharm

iii. Numpy

iv. Pandas

v. Seaborn

3.3 Research Design

i. **Reading and Exploring of data**: The dataset was read by first mounting the Google Drive to the Colab environment using the command `drive.mount('/content/drive')`. After successfully mounting the drive, the CSV file containing the airline reviews is read into a pandas DataFrame using the command `pd.read_csv('/content/drive/MyDrive/BA_AirlineReviews.csv')`. The dataset was checked for missing values using and duplicate rows.

ii. **Data Preprocessing**: This involved the removal of duplicates, treatment of missing values, Label Encoding and Feature Engineering.

iii. **Sentiment Analysis**: Sentiment analysis was conducted using two primary NLP (NLP) techniques; TextBlob and VADER (Valence Aware Dictionary and sEntiment Reasoner). TextBlob was used to analyse the sentiment polarity of each review by classifying sentiments into positive, neutral, or negative based on the polarity score. The sentiment analysis using VADER involved initializing the VADER sentiment analyser to calculate a compound polarity score for each review. The sentiment was classified into positive, neutral, or negative based on the compound score, and further converted into a binary classification (positive or non-positive).

iv. **Exploratory Data Analysis (EDA)**: Exploratory Data Analysis (EDA) was performed by creating visualizations, such as histograms, bar charts, pie charts, and box plots, to analyse the distribution of ratings, the relationship between different variables (like aircraft type, seat class, and customer satisfaction), and to assess the impact of these factors on overall customer satisfaction. The EDA also examined correlations between variables such as value for money, cabin staff service, and recommendation status, highlighting their significance in shaping customer perceptions and satisfaction levels.

v. **Model Development**: Machine learning models, Logistic Regression, Naive Bayes, Support Vector Machine (SVM), and XGBoost, wes developed and evaluated using both VADER sentiment scores and TF-IDF (Term Frequency-Inverse Document Frequency) features extracted from the customer reviews. The models were trained and tested on the dataset to predict customer satisfaction and the likelihood of recommending the airline. The performance of each model was measured using accuracy, precision, recall, and F1 scores.

vi. **Deep Learning Model Development:** Long Short-Term Memory (LSTM) network, was developed to predict customer satisfaction and recommendation outcomes based on the review data. The LSTM model was trained over multiple epochs

Chapter Four

Result

4.1 Introduction

This section presents the findings of this study based on the analysis of data, encompassing several key areas. The analysis included sentiment scoring using both TextBlob and VADER to evaluate customer reviews, followed by extensive data visualization and exploratory data analysis to uncover patterns and insights. Furthermore, the study involved the development and evaluation of various predictive models, including Logistic Regression, Naive Bayes, Support Vector Machine, XGBoost, and LSTM (Deep Learning). Each of these models was employed to assess their effectiveness in predicting customer satisfaction and recommendation likelihood, offering comprehensive insights into the factors influencing airline service quality

4.2 Sentiment scoring: Using TextBlob

In the sentiment scoring process, TextBlob was used to analyse the sentiment expressed in the reviews and quantify it. First, a function \`classify_sentiment(polarity)\` is defined to classify the sentiment score based on polarity. If the polarity is greater than 0, the sentiment is classified as positive; if it is 0, it is neutral; otherwise, it is negative. The sentiment score is categorized into positive, neutral, or negative based on the polarity using the \`classify_sentiment\` function. The overall rating is grouped into two categories to test the polarity score.

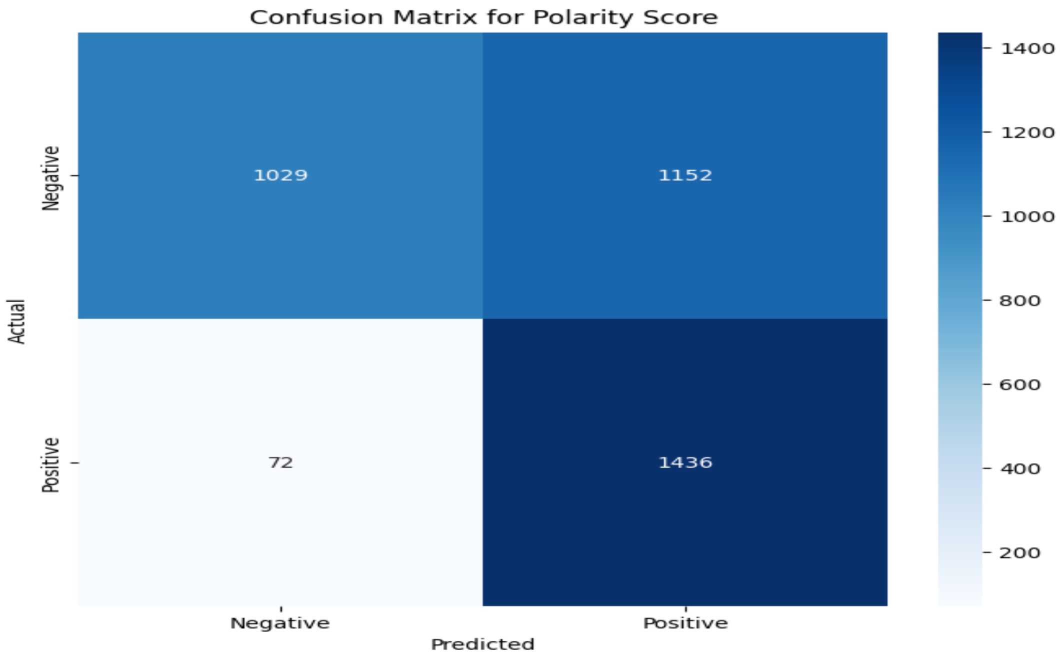


Figure 4.1: Confusion Matrix for Polarity Score

The confusion matrix evaluates the performance of the sentiment polarity score classification. It compares the actual sentiment (positive or negative) against the predicted sentiment based on the polarity score.

True Negatives (TN): 1029 instances where both actual and predicted sentiments are negative.

False Positives (FP): 1152 instances where the actual sentiment is negative, but the predicted sentiment is positive.

False Negatives (FN): 72 instances where the actual sentiment is positive, but the predicted sentiment is negative.

True Positives (TP): 1436 instances where both actual and predicted sentiments are positive.

The high number of True Positives and True Negatives indicates that the model is relatively good at predicting sentiment correctly.

Table 4.1: Classification Report for Polarity Score

	Precision	Recall	F1 Score	Support
0	0.93	0.47	0.63	2181
1	0.55	0.95	0.70	1508
Accuracy			0.67	3689
macro avg	0.74	0.71	0.66	3689
weight avg	0.78	0.67	0.66	3689

The classification report for the polarity score, above showa

- **Precision**: Precision for class 0 (negative sentiment) is 0.93, indicating that 93% of the predicted negative sentiments were negative. For class 1 (positive sentiment), precision is 0.55, meaning that 55% of the predicted positive sentiments were correct.
- **Recall**: Recall for class 0 is 0.47, indicating that the model correctly identified 47% of all actual negative sentiments. For class 1, recall is 0.95, showing that the model identified 95% of all actual positive sentiments.
- **F1-Score**: The F1-score for class 0 is 0.63, and for class 1, it is 0.70.
- **Accuracy**: The overall accuracy of the model is 0.67, indicating that 67% of the predictions were correct.

Although the TextBlob model accurately predicts many reviews, it also produces a significant number of false positives. The model tends to over-predict positive sentiments, as indicated by the high number of false positives (1152). This discrepancy suggests that while the model is effective in identifying positive sentiments (high recall), it struggles with precision, leading to many incorrect positive predictions.

4.3 Sentiment scoring: Using VADER

The VADER sentiment analyser was initialized with the command vader_analyser = SentimentIntensityAnalyzer(). The necessary NLTK resources are downloaded, including the VADER lexicon. Two functions were defined: classify_vader_sentiment(text) classifies sentiment based on the compound polarity score. A score ≥ 0.05 is considered positive, ≤ -0.05 is negative, and scores in between are neutral get_vader_polarity(text) returns the compound polarity score from VADER using the snippet code below. A copy of the DataFrame is created for VADER analysis to ensure the original data remains unchanged. The compound polarity scores for each review are calculated and stored in the vader_polarity column. The sentiment scores are classified into positive, neutral, or negative and stored in the vader_sentiment_score column using the classify_vader_sentiment function. The sentiment scores are further converted into a binary classification. Reviews with a positive sentiment are classified as 1, and all other sentiments (neutral and negative) are classified as 0. This binary classification is stored in the vader_sentiment_binary column.

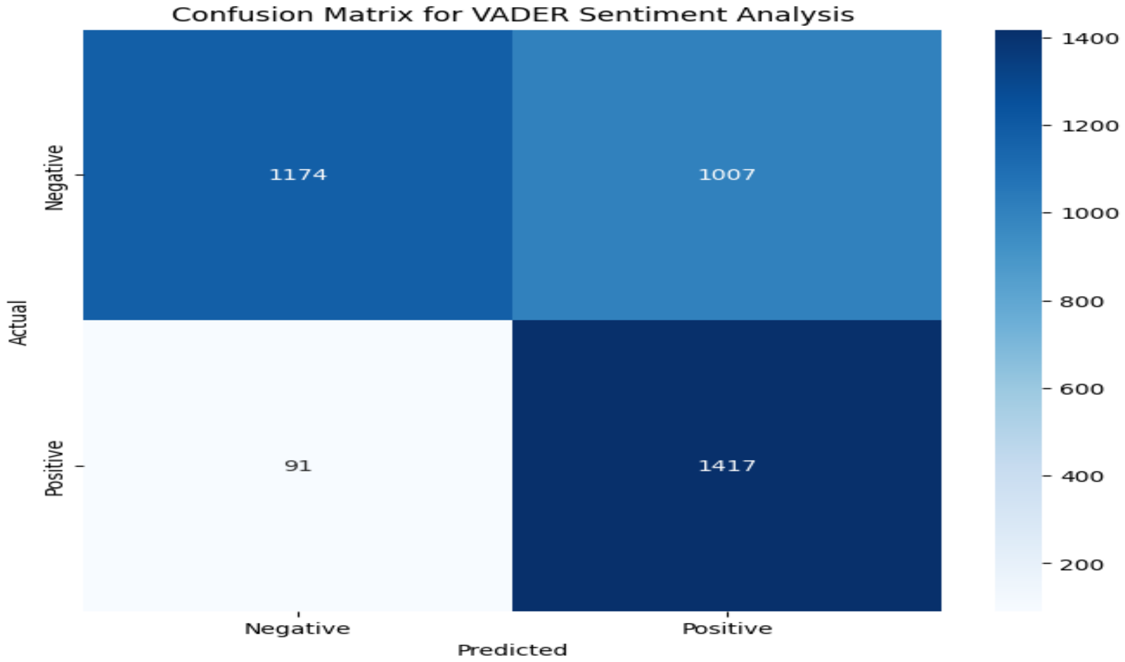


Figure 4.2: Confusion Matrix for VADER Sentiment Analysis

From the confusion matrix for the VADER sentiment shows 1174 true negatives, indicating instances where both actual and predicted sentiments are negative. It also shows 1007 false positives, where actual sentiments are negative, but the model predicted positive sentiments. Additionally, there are 91 false negatives, where actual sentiments are positive, but the model predicted negative sentiments, and 1417 true positives, where both actual and predicted sentiments are positive. The high number of true positives (1417) suggests that the model effectively identifies positive reviews. The true negatives (1174) indicate the model's ability to correctly identify negative reviews. However, the false positives (1007) indicate a tendency to overestimate positive sentiment, where negative reviews are incorrectly classified as positive. The false negatives (91) are relatively low, showing that the model misses only a small number of positive reviews.

Comparing this to the TextBlob confusion matrix, VADER shows an improvement in reducing false positives from 1152 to 1007, while maintaining a low number of false negatives, increasing slightly from 72 to 91. This suggests that VADER offers a better balance in sentiment prediction, reducing the tendency to over-predict positive sentiments. VADER demonstrates high accuracy in predicting positive reviews while maintaining reasonable accuracy in predicting negative reviews. Despite the overestimation of positive sentiment, the model's overall performance indicates it as a reliable tool for sentiment analysis.

Table 4.2: Classification Report for VADER Sentiment Analysis

	Precision	Recall	F1 Score	Support
0	0.93	0.54	0.68	2181
1	0.58	0.94	0.72	1508
Accuracy			0.70	3689
macro avg	0.76	0.74	0.70	3689
weight avg	0.79	0.70	0.70	3689

From the classification report.

Class 0 (Negative Sentiment): Precision of 0.93 indicates that 93% of the reviews predicted as negative were negative.

Recall of 0.54 indicates that the model correctly identified 54% of all actual negative reviews. F1-Score: 0.68 is the harmonic mean of precision and recall, balancing the two metrics. Class

1 (Positive Sentiment): Precision of 0.58 indicates that 58% of the reviews predicted as positive were positive. Recall of 0.94 indicates that the model correctly identified 94% of all actual positive reviews. F1-Score of 0.72 balances precision and recall for positive reviews.The accuracy of 0.70 means that 70% of the reviews were correctly classified by the model. Macro Average of 0.76 (precision), 0.74 (recall), and 0.70 (F1-score) average the metrics across both classes without considering class imbalance.

Weighted Average: 0.79 (precision), 0.70 (recall), and 0.70 (F1-score) average the metrics across both classes while considering class imbalance.

True Positives (TP): The model effectively identifies positive reviews with a high recall of 0.94, meaning it captures most of the positive sentiments.

True Negatives (TN): The high precision of 0.93 for negative sentiment indicates that when the model predicts negative, it is usually correct.

False Positives (FP): The lower precision for positive sentiment (0.58) suggests that the model often predicts positive when the actual sentiment is negative.

False Negatives (FN): The lower recall for negative sentiment (0.54) indicates that the model misses a significant number of negative reviews.

Compared to the TextBlob results, VADER has a slightly better overall accuracy (0.70 compared to TextBlob's 0.67). VADER shows an improvement in recall for positive sentiment (0.94 vs. 0.95) and maintains a balance between precision and recall for both positive and negative sentiments. VADER provides a more balanced performance in sentiment analysis compared to TextBlob, particularly in reducing false positives and maintaining high recall for positive sentiments. The overall accuracy of 70% and balanced precision and recall metrics suggest that VADER is a reliable tool for sentiment analysis.

4.4 Data Visualization & Exploratory Data Analysis

Word Cloud for Negative Reviews

The word cloud serves to highlight common issues and themes present in reviews where customers have provided a negative recommendation. Such insights are invaluable for pinpointing areas requiring improvement. Negative reviews are extracted and concatenated into a single string, with spaces separating individual reviews. This comprehensive string is then cleaned to remove any newline characters that could disrupt the visualization process. The resulting word cloud visually represents the most frequent words used in negative reviews. This visualization aids in identifying prevalent issues and recurring themes among dissatisfied customers.

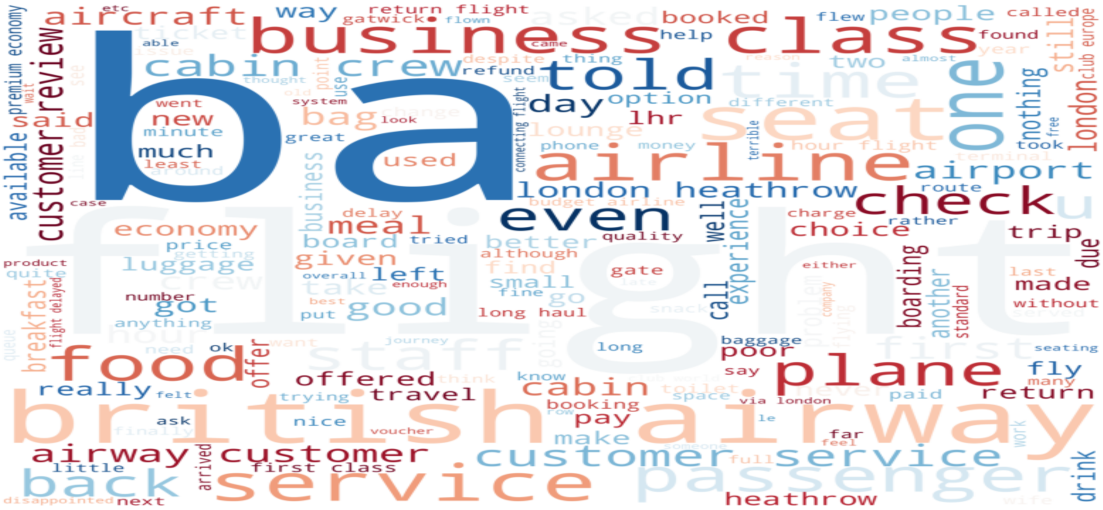


Figure 4.3: Word Cloud

4.4.1 Exploratory Data Analysis

1. Distribution of Over All Rating

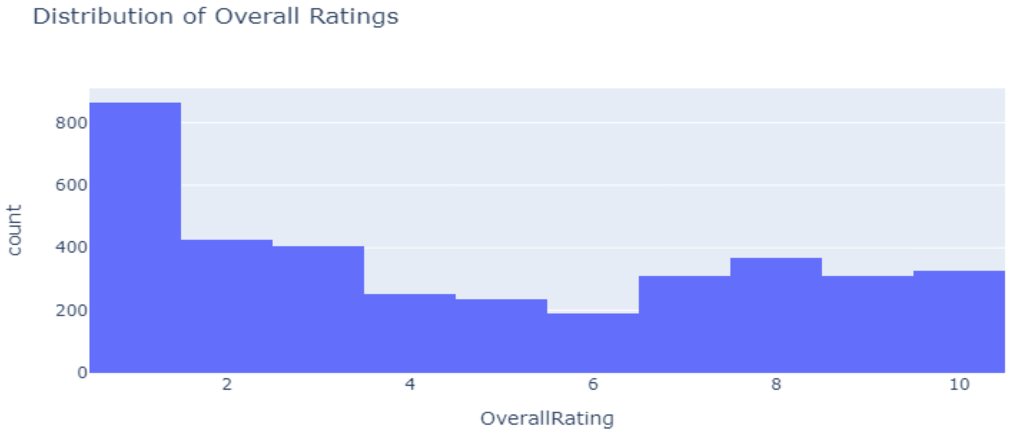


Figure 4.4: Distribution of Over All Rating

From the histogram, the distribution is highly skewed towards the lower end of the rating scale. The highest concentration of reviews is seen at the rating of 1, with over 800 reviews. This suggests that a significant number of customers were highly dissatisfied with their experience. As the rating increases, the number of reviews steadily decreases, indicating fewer occurrences of higher satisfaction levels. Ratings around the mid-point, particularly between 4 and 6, show a noticeable dip in the number of reviews, suggesting that moderate satisfaction is less common among the respondents. Towards the higher end of the scale, ratings from 8 to 10, the frequency of reviews rises slightly but remains relatively low compared to the lowest rating. This pattern implies that while there are some highly satisfied customers, they are significantly outnumbered by those who rated their experience poorly.

2. Average Overall Rating by Aircraft Company

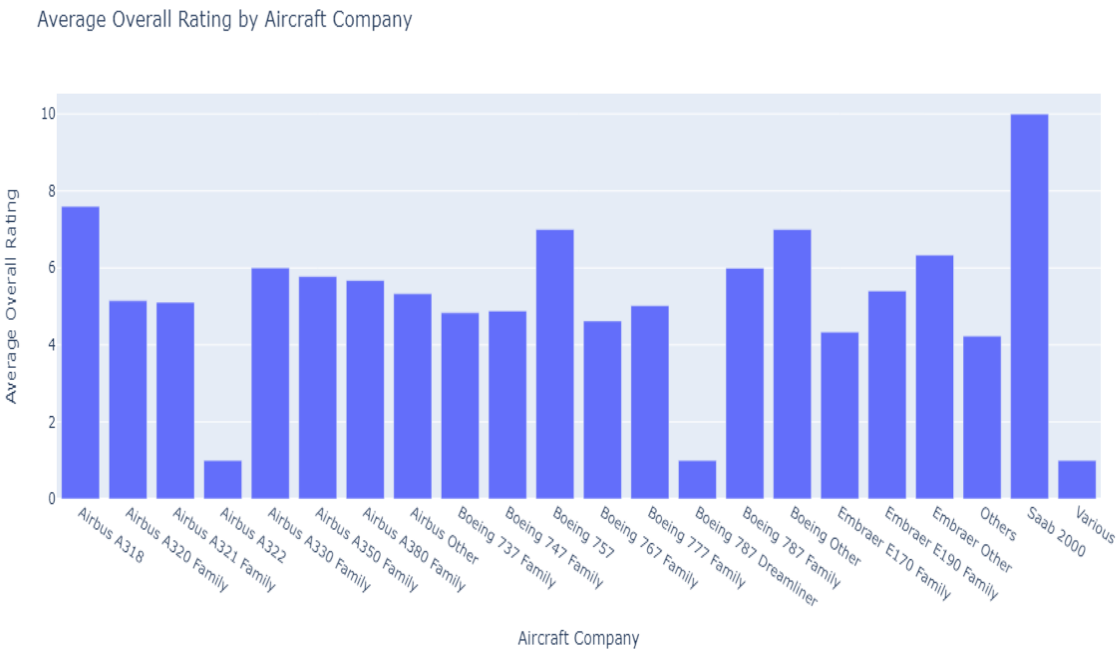


Figure 4.5: Average Overall Rating by Aircraft Company

From the chart, there is significant variation in customer satisfaction across different aircraft types. The Saab 2000 and the Boeing 747 Family stand out with the highest average ratings, indicating higher customer satisfaction for these aircraft. In contrast, the Boeing 757 and the Airbus A321 Family have the lowest average ratings, suggesting lower satisfaction levels among customers for these aircraft. The Airbus A318 and Boeing 737-900 also show relatively lower ratings compared to other models. On the other hand, the Boeing 787 Dreamliner and the Airbus A350 have moderate to high ratings, indicating a generally favourable customer experience. The variation in ratings suggests that different aircraft types significantly impact customer satisfaction. Factors contributing to this variation could include comfort, space, in-flight services, and overall flight experience associated with each aircraft type.

3. Proportion of Seat Types

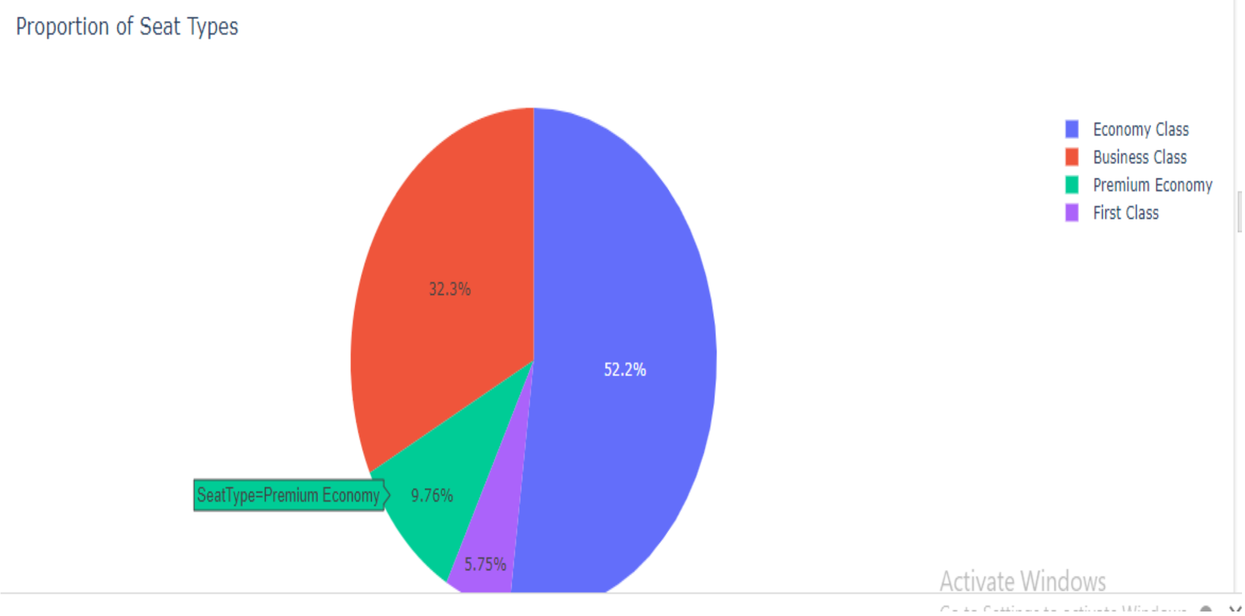


Figure 4.6: Proportion of Seat Types

The pie chart illustrates the proportion of passengers flying in different classes: Economy Class, Business Class, Premium Economy, and First Class. The largest segment, representing over half of the passengers (52.2%), flew in Economy Class, indicating it as the most popular choice among travellers. This is followed by Business Class, which accounts for 32.3% of the passengers. Premium Economy and First Class make up smaller portions of the total, with 9.76% and 5.75%, respectively. The significant dominance of Economy Class suggests that cost-effective travel options are a priority for the majority of passengers. Business Class also has a substantial share, indicating a significant demand for enhanced comfort and services that come at a higher cost. The relatively lower percentages for Premium Economy and First Class highlight these as niche segments, likely chosen by passengers seeking additional comfort and services but representing a smaller portion of the market.

4. Verified Review Count by Type of Traveller

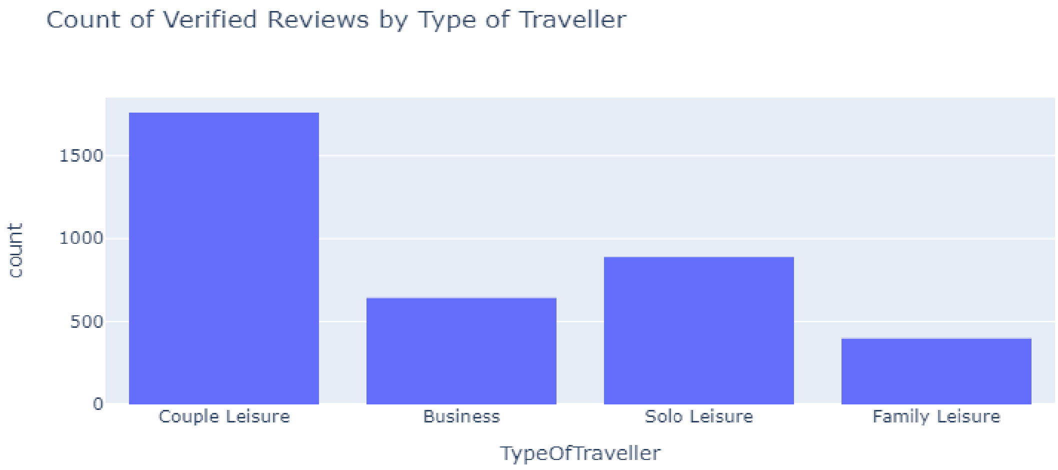


Figure 4.7: Verified Review Count by Type of Traveller

The bar chart displays the count of verified reviews categorized by different types of travellers: Couple Leisure, Business, Solo Leisure, and Family Leisure. The x-axis represents the type of traveller, while the y-axis indicates the count of verified reviews for each category. From the chart, it is clear that the majority of verified reviews come from travellers on leisure trips as couples, with the highest count at approximately 1759 reviews. This is followed by solo leisure travellers, who contribute around 891 reviews. Business travellers provide a moderate number of reviews, with a count of about 641. The smallest number of reviews comes from family leisure travellers, with around 398 reviews. The dominance of reviews from couple leisure (1759) and solo leisure (891) travellers suggests that these groups are more likely to provide feedback about their travel experiences.

5. Recommended vs Value for Money

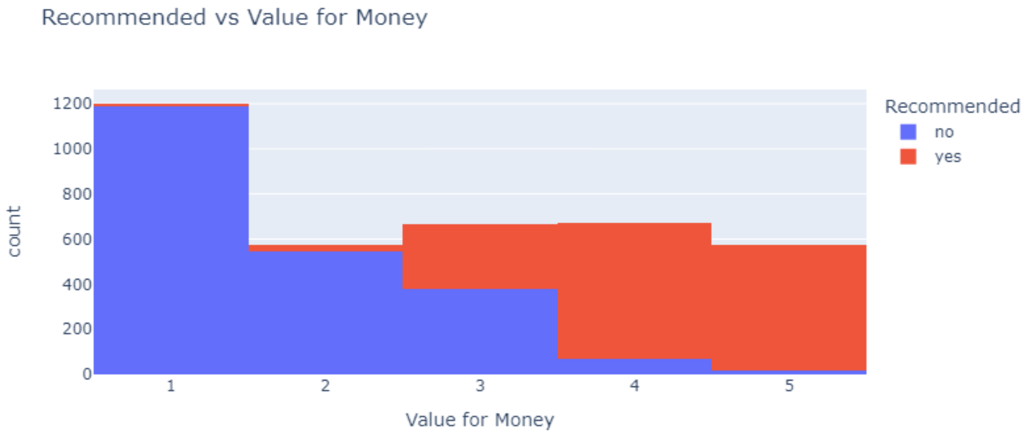


Figure 4.8: Recommended vs Value for Money

From the chart, the higher value for money ratings are associated with a higher likelihood of recommending the airline. At the lowest value for money rating of 1, the majority of reviews indicate that customers do not recommend the airline, with the count reaching up to approximately 1186 for non-recommendations and very few recommendations. As the value for money rating increases to 2 and 3, the count of non-recommendations decreases, with counts around 546 and 380 respectively, while the count of recommendations starts to increase, reaching up to 30 for rating 2 and around 285 for rating 3. The higher ratings of 4 and 5, the pattern shifts significantly. For a value for money rating of 4, there is a noticeable increase in the count of recommendations, approximately 606. This continues and reduces at the highest value for money rating of 5, where recommendations is 558 . The histogram clearly shows a positive correlation between the perceived value for money and the likelihood of recommending the airline. Improving the value for money aspect of the airline's services could thus be an effective strategy to increase customer recommendations and overall satisfaction.

6. Average Recommendation Status by Year

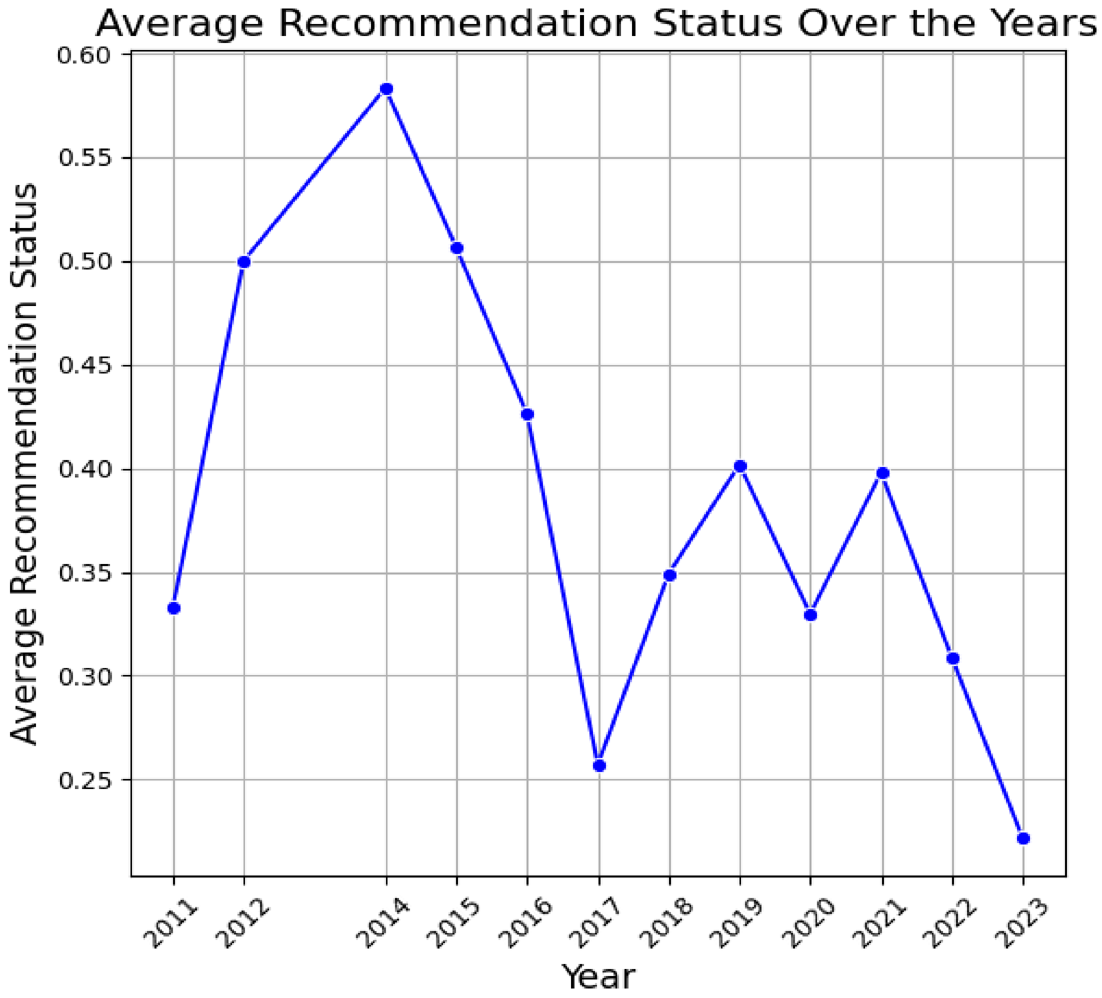


Figure 4.9: Average Recommendation Status by Year

The line graph displays the average recommendation status of the airline over the years, from 2011 to 2023. The graph shows significant fluctuations in the average recommendation status over the analysed period. Starting in 2011, the average recommendation status is around 0.30. This figure rises steadily, reaching a peak of approximately 0.58 in 2014, indicating a high level of customer satisfaction and recommendation during that year. After 2014, there is a noticeable decline in the average recommendation status, dropping to about 0.40 in 2016 and further declining to a low point of around 0.25 in 2017. This sharp decrease suggests a period of declining customer satisfaction and fewer recommendations. From 2017 onwards, the recommendation status begins to recover, rising to about 0.35 in 2018. The following years show a pattern of fluctuations: the status increases to around 0.40 in 2019, drops slightly in 2020, and sees another rise and fall through 2021 and 2022, respectively. In 2023, the average recommendation status drops to its lowest point around 0.25. The peak in 2014 suggests a time of excellent customer experiences, while the dips in 2017 and 2023 indicate times when customer satisfaction was significantly lower.

7. Average Recommendation Status by Year

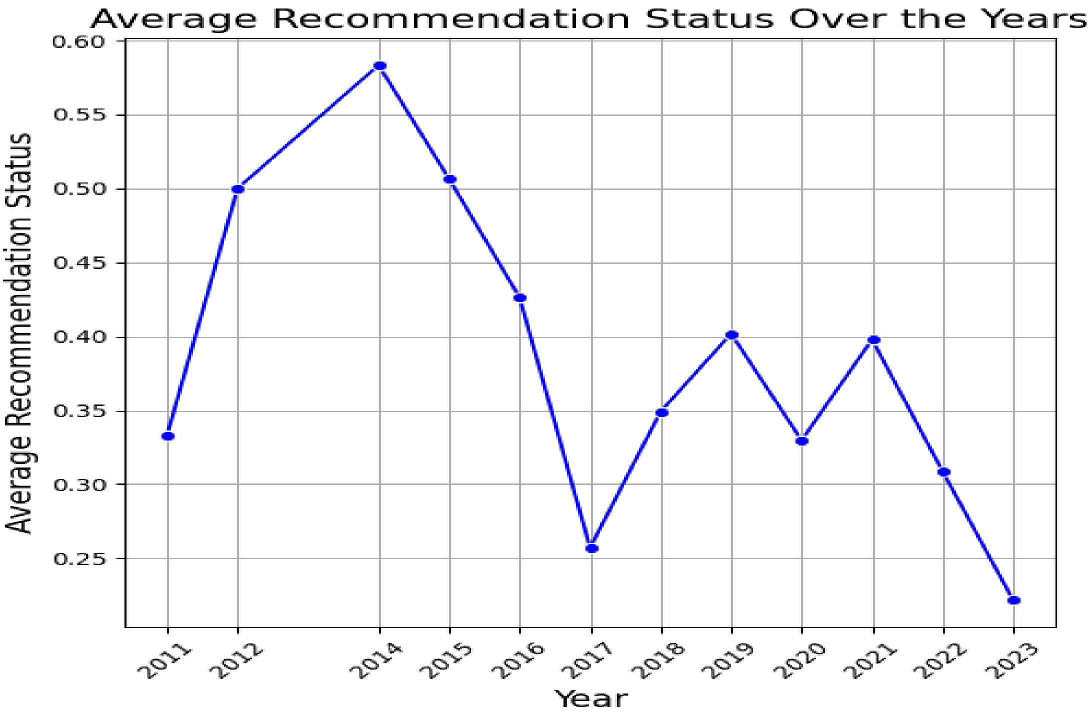


Figure 4.10: Average Recommendation Status by Year

The line plot Average Recommendation Status Over the Years shows the trend of recommendation status from 2011 to 2023. In 2014, the recommendation status reached its peak at around 0.58. However, following this peak, there was a significant decline, with the status dropping sharply by 2017. After 2017, the trend became more variable, with the status fluctuating but never returning to the heights observed in 2014. The data for 2023 indicates that the recommendation status has fallen to its lowest point on the graph, around 0.25. This overall trend suggests that while there was a period of high recommendation status in the mid-2010s, it has generally declined in the years since, with notable fluctuations in the later years.

8. Average Recommendation Status by Month

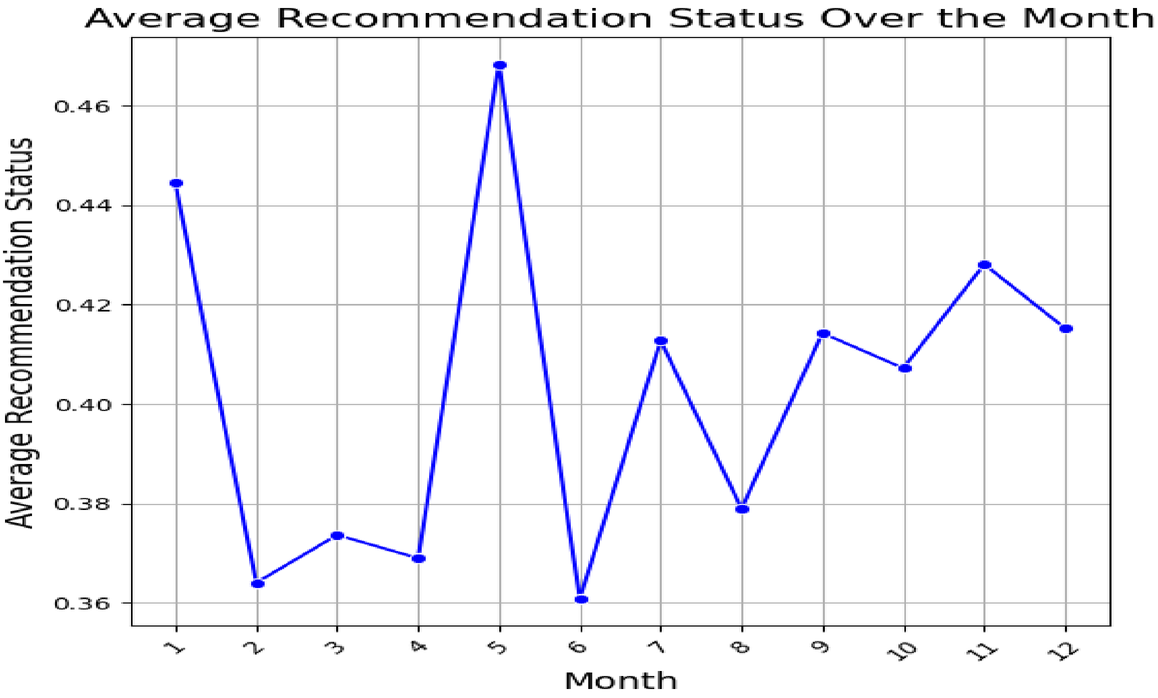


Figure 4.11: Average Recommendation Status by Month

The plot Average Recommendation Status Over the Month shows how the recommendation status changes across the months of a year. In the beginning of the year, the recommendation status is relatively high, starting at around 0.44 in January. However, this status quickly drops in February, reflecting a sharp decline. The status stabilizes somewhat in March and April, followed by a dramatic increase in May, where it peaks at its highest value of around 0.46. June sees a steep drop, with the recommendation status falling to its lowest point of approximately 0.36. As the year progresses, the recommendation status starts to rise again from July onward, although this upward trend is marked by some fluctuations. November shows a local peak before the status slightly decreases again in December. The plot indicates that the recommendation status is quite volatile throughout the year, with significant variations between months. The reasons for these fluctuations could be tied to seasonal effects, specific events, or other time-sensitive factors that influence recommendation status differently at various points in the year.

9. Distribution of Type of Travellers

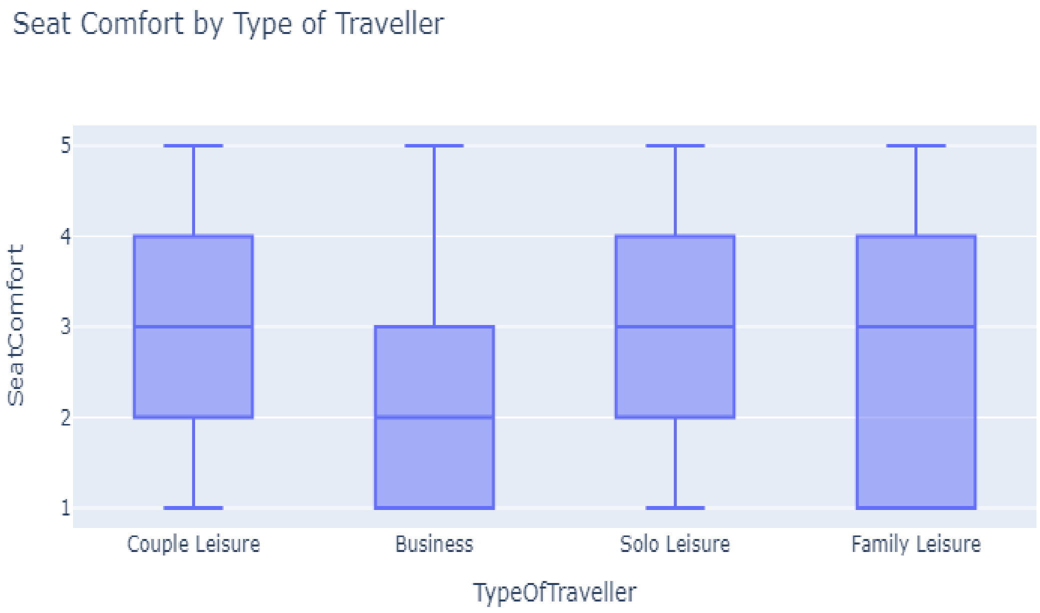


Figure 4.12: Distribution of Type of Travellers

The box plot illustrates the distribution of seat comfort ratings across different types of travellers. Couple Leisure and Family Leisure travellers exhibit a wide range of seat comfort ratings, from 1 to 5, indicating a high variability in their experiences. The median comfort rating for these groups hovers around 3, with some travellers rating their seat comfort as very high (5) or very low (1).Business travellers also show a wide range of seat comfort ratings, but the median is slightly lower than for leisure travellers, suggesting that business travellers might have a slightly less favourable experience with seat comfort overall. Solo Leisure travellers show a similar pattern to business travellers, with ratings ranging from 1 to 5 and a median rating around 3. The variability (indicated by the length of the boxes and the whiskers) is substantial across all traveller types, suggesting diverse experiences within each group. However, no significant outliers are visible in the plot, indicating that extreme ratings are not unusually frequent.

The plot suggests that seat comfort experiences are quite variable regardless of the type of traveller. While there are differences in medians and the range of ratings, all types of travellers experience the full spectrum of seat comfort, from very poor to excellent. This variability in seat comfort ratings across different traveller types could be influenced by several factors, including the specific airline services, the type of aircraft, or personal expectations of comfort.

10. Overall Rating Distribution by Recommendation Status

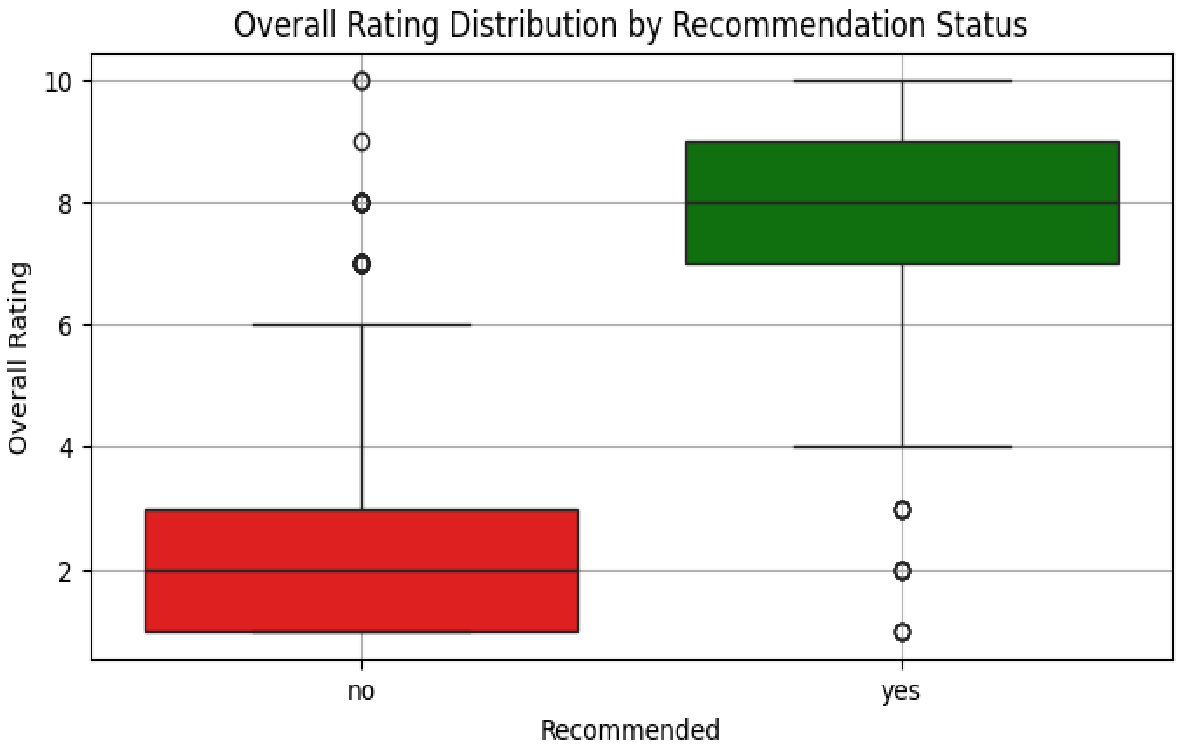


Figure 4.13: Overall Rating Distribution by Recommendation Status

The box plot compares the overall ratings between two groups: those who would recommend the service ("yes") and those who would not ("no").The group that recommended the service ("yes") has a higher overall rating, with the interquartile range (IQR) mostly between 7 and 9. This suggests that most people who recommend the service rate it quite high. In contrast, the group that did not recommend the service ("no") has a significantly lower overall rating, with the IQR ranging between 2 and 4. This indicates that those who do not recommend the service generally rate it poor. The box plot for the "yes" group shows a relatively compact distribution, with few outliers, indicating consistent positive ratings among those who recommend the service. The "no" group shows a more dispersed distribution, with several outliers indicating that some people who do not recommend the service still gave it higher ratings. Further, there is a clear distinction between the overall ratings of those who would recommend the service and those who would not. The "yes" group generally has high ratings, suggesting satisfaction with the service, while the "no" group has much lower ratings, indicating dissatisfaction. This plot effectively highlights the relationship between recommendation status and overall satisfaction, showing that those who are satisfied with the service are much more likely to recommend it to others.

11. Overall Rating by Cabin Staff Service

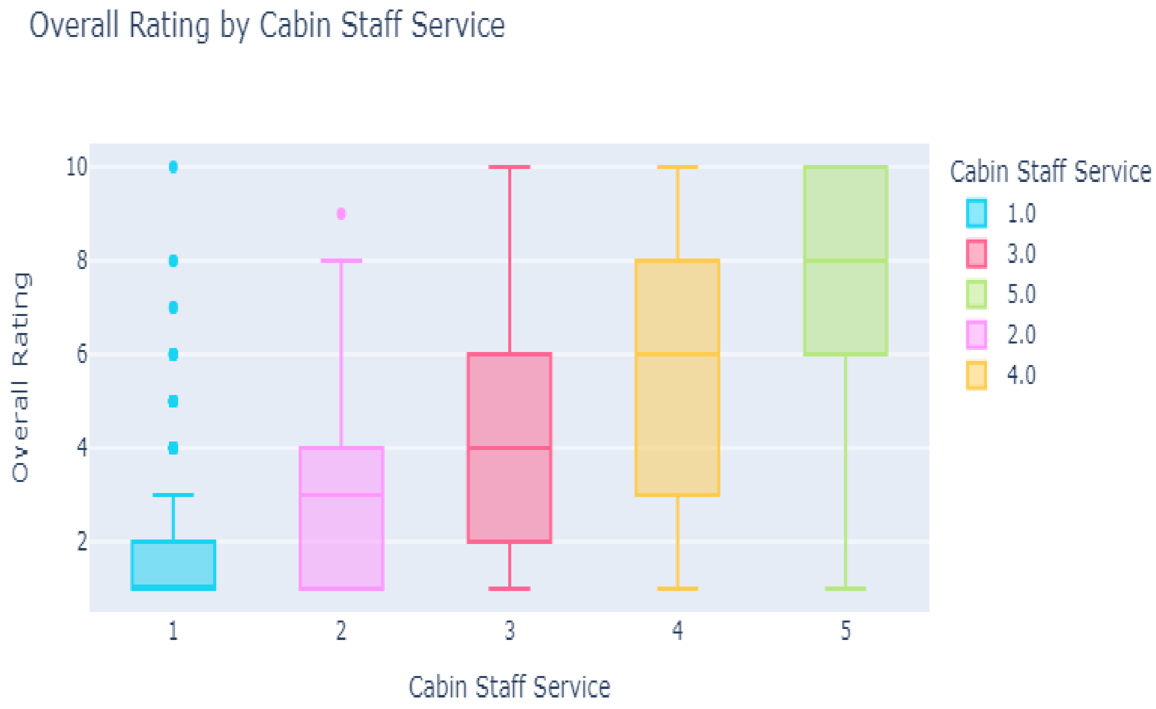


Figure 4.14: Overall Rating by Cabin Staff Service

The box plot illustrates the relationship between the cabin staff service quality and the overall ratings given by customers. The plot reveals a clear trend where higher cabin staff service ratings are associated with higher overall ratings. For instance, when cabin staff service is rated as 1, the overall ratings are consistently low, mostly clustered around 2 to 3, with very little variability and several outliers indicating particularly poor experiences. As the cabin staff service rating increases to 2 and 3, the overall ratings start to improve, with median ratings rising and the range of ratings broadening. However, the median still remains in the lower half of the scale. Significant improvement is evident when the cabin staff service is rated at 4 or 5. At these levels, the overall ratings are notably higher, with median ratings approaching the upper end of the scale, and the range of ratings extending from around 5 to 10. Particularly, a cabin staff service rating of 5 corresponds with some of the highest overall ratings, indicating that excellent service by the cabin staff is closely linked with overall customer satisfaction. This analysis suggests a strong positive correlation between the quality of cabin staff service and the overall rating, with better service consistently leading to higher customer satisfaction. The variability in ratings also decreases as the service quality improves, highlighting that excellent service tends to produce more consistently high ratings.

12. Recommendation Status by Food & Beverage Service

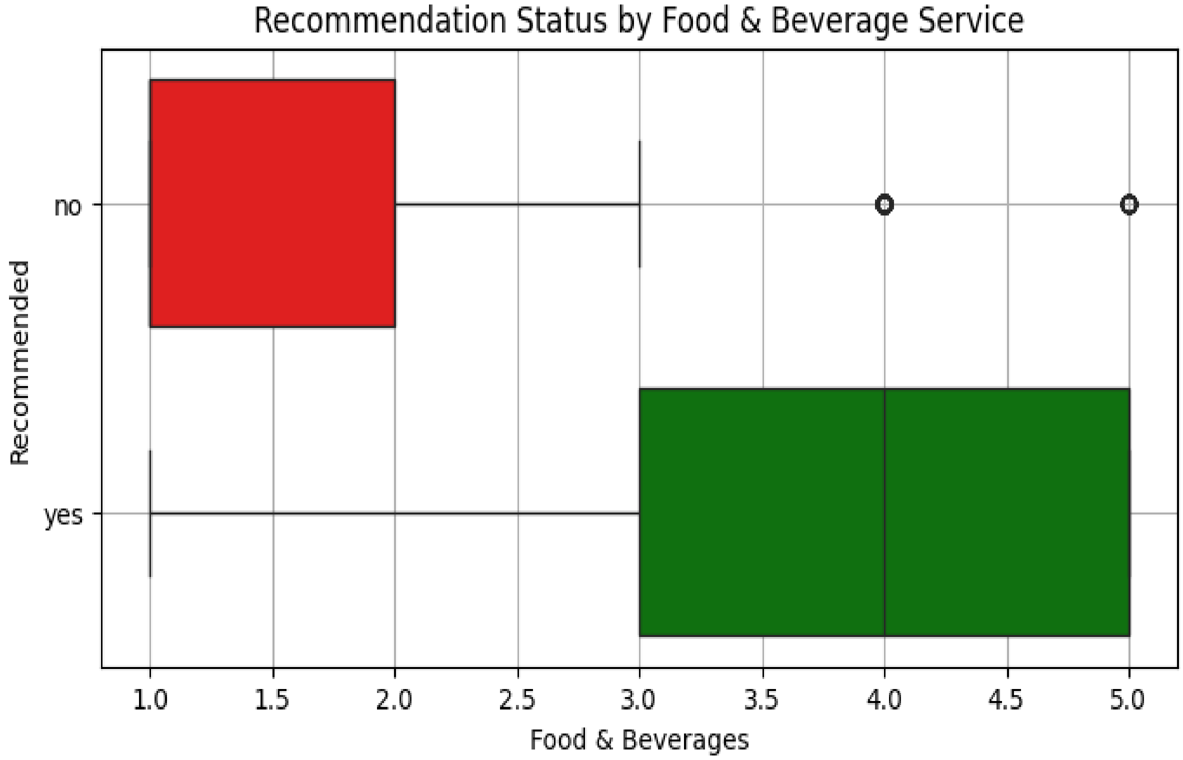


Figure 4.15: Recommendation Status by Food & Beverage Service

The plot reveals a clear pattern: customers who rated the food and beverage service lower (1 or 2) are much less likely to recommend the service. This is indicated by the red box corresponding to the "no" category, where most ratings for food and beverages are clustered around 1 to 2. As the ratings for food and beverage service increase to 3, the recommendation status shifts more favourably, with more customers in the "yes" category, represented by the green box. For higher food and beverage ratings of 4 and 5, the majority of customers fall into the "yes" category, indicating they would recommend the service. The green box here is larger, suggesting that higher satisfaction with food and beverages strongly correlates with a positive recommendation. There are also some outliers in the higher rating categories, indicating that a few customers still chose not to recommend the service despite rating the food and beverage service highly, though these cases are exceptions. The analysis demonstrates that better food and beverage service is associated with a higher likelihood of customers recommending the service, with a clear divide between low and high ratings and their corresponding recommendation statuses.

13. Sentiment Score Distribution

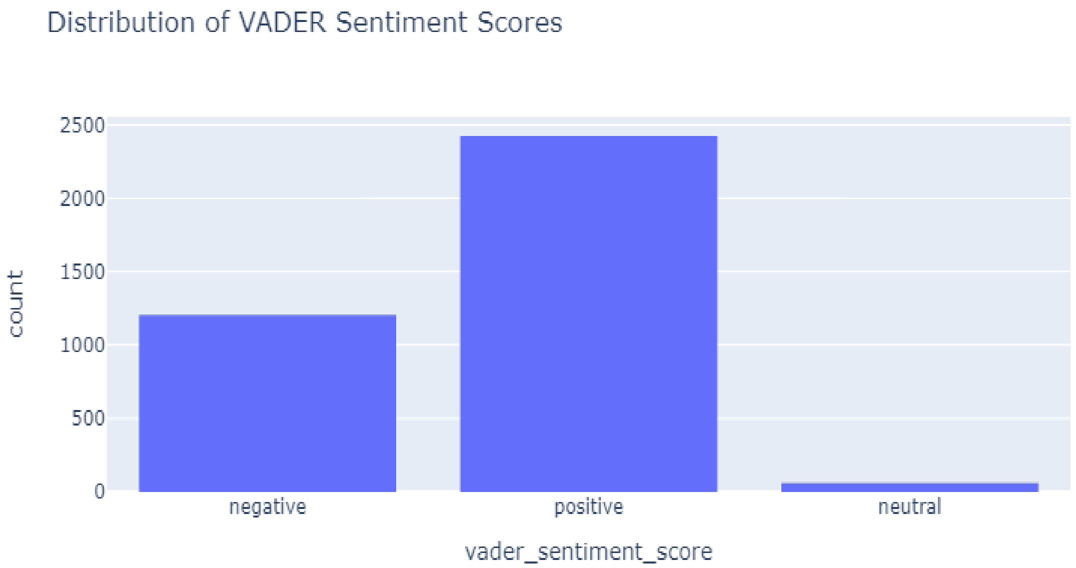


Figure 4.16: Sentiment Score Distribution

The majority of the sentiments analysed are classified as "positive," with a count of 2424 instances. This indicates that most of the textual data or reviews assessed have a favourable sentiment. The "negative" sentiment category is also significant but considerably lower, with around 1204 instances, suggesting that while there are some unfavourable sentiments, they are less frequent compared to the positive ones. The "neutral" sentiment category has an extremely low count (61), almost negligible compared to the other two categories. This suggests that most of the sentiments expressed in the analysed data tend to be polarized towards either positive or negative, with very few neutral sentiments. The data reveals that positive sentiments dominate the dataset, with negative sentiments being present but less prevalent, and neutral sentiments being rare.

14. Heat Map

The heatmap visualizes the correlation between various features related to airline service and customer satisfaction. The color scale represents the strength and direction of these correlations, with values close to 1 indicating a strong positive correlation and values close to -1 indicating a strong negative correlation. The strongest correlation observed is between the overall rating and the value for money (0.87). This indicates that customers who perceive they received good value for money are more likely to rate the overall service highly. This suggests that pricing and the perceived fairness of costs play a significant role in overall customer satisfaction. The correlation between the likelihood of recommending the service and the overall rating is also very high (0.86). This demonstrates that customers who rate their experience highly are very likely to recommend the airline to others, emphasizing the importance of maintaining high service standards across all touchpoints.

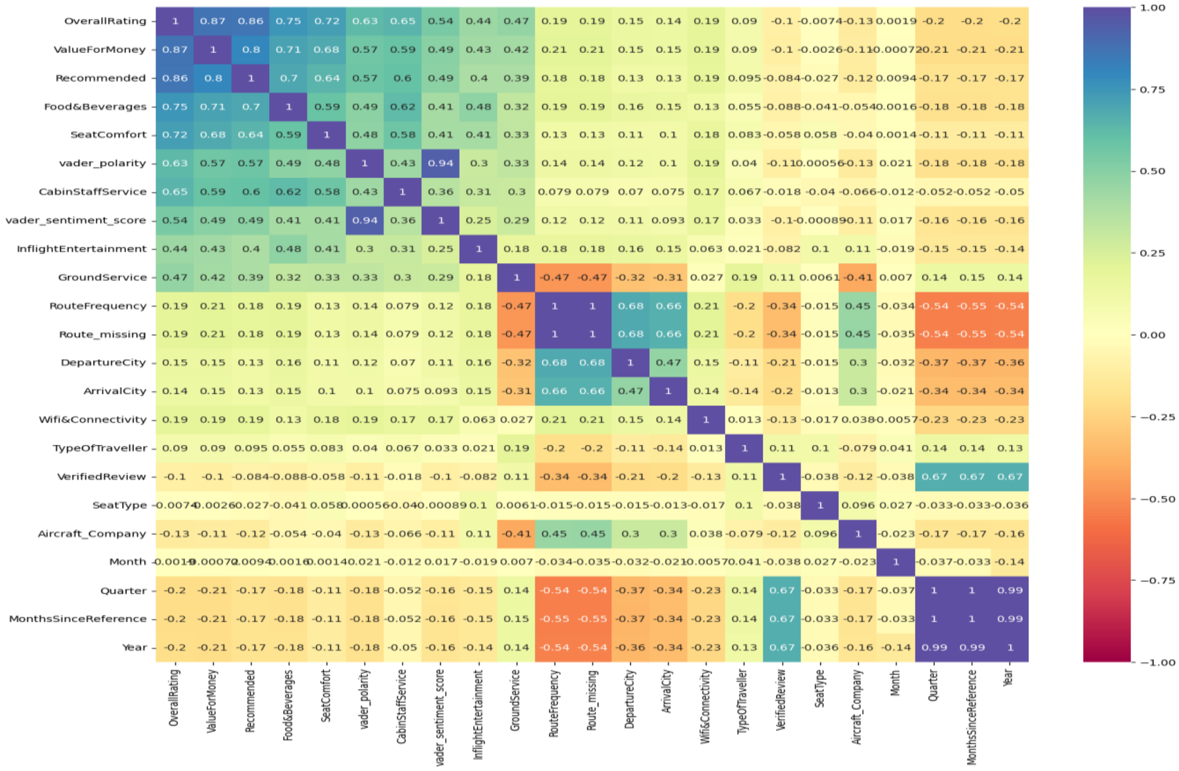


Figure 4.17: Heat Map

Food and beverages (0.72) and seat comfort (0.72) both show strong positive correlations with the overall rating. This suggests that the quality of in-flight meals and the comfort of seating are critical factors in shaping the overall satisfaction of passengers. The correlation between cabin staff service and overall rating is also significant (0.65), indicating that the behaviour and efficiency of cabin staff contribute greatly to the passenger experience. Enhancing staff training and ensuring consistent, high-quality service can positively impact customer perceptions.

The VADER sentiment score, which measures the sentiment expressed in customer reviews, has a notable positive correlation with both the overall rating (0.54) and the recommendation status (0.49). This highlights the value of analysing customer feedback to predict overall satisfaction and the likelihood of customers recommending the airline. Both inflight entertainment (0.47) and ground service (0.48) show moderate correlations with overall satisfaction. While they are important, their impact is slightly less compared to factors like food, seat comfort, and value for money. Route frequency and Wi-Fi connectivity show weaker correlations with overall satisfaction, with values of 0.19 and 0.12, respectively. Temporal factors like the year and month of the flight have very low correlations with the overall rating, indicating that customer satisfaction is relatively stable over time and not significantly affected by the time of travel. The heatmap analysis suggests that the most critical factors influencing customer satisfaction are value for money, overall service quality (including food, beverages, and seat comfort), and the likelihood of recommending the airline to others. Enhancing these aspects can significantly boost customer satisfaction and loyalty. Additionally, analysing customer sentiment through reviews provides valuable insights into their experiences and can guide service improvements.

4.5 Model Development

The following models were used in this project

- Logistic Regression
- Naive Bayes
- Support Vector Machine
- XGBoost
- LSTM (Deep Learning)

4.5.1 Classification Models Using VADER Sentiment Score for Text Data

1. Logistic regression

Table 4.3 Logistic Regression

Models	Accuracy	Precision	Recall	F1 Score
Logistic regression	0.92	0.92	0.92	0.92
Naive Bayes	0.90	0.89	0.90	0.90
Support Vector Machine	0.94	0.94	0.94	0.94
XGBoost Model	0.93	0.93	0.93	0.93

The logistic regression performance demonstrates a 92% performance across all metrics. Naive Bayes shows no improvement of Logistic Regression. The support vector machine performance demonstrates a 94% performance accross all metrics which is a slight improvement and shows that the training set performs slight better. XGBoost Model has overfitting on the training set with 100% in all metrics. The ensemble model is overfitting which is a common limitation of tree-based models. Pre-pruning by adjusting the max-depth was applied to address the overfitting. After the pre-pruning, the model shows high performance with 99% accuracy on training data and 93% on testing data. This suggests strong learning and good generalization, but the slight drop in test accuracy indicates minor overfitting.

Since XGBoost is a tree-based model, the researcher capitalized on its feature importance functionality to identify the top features that contributed to the prediction of user recommendation/ satisfaction.

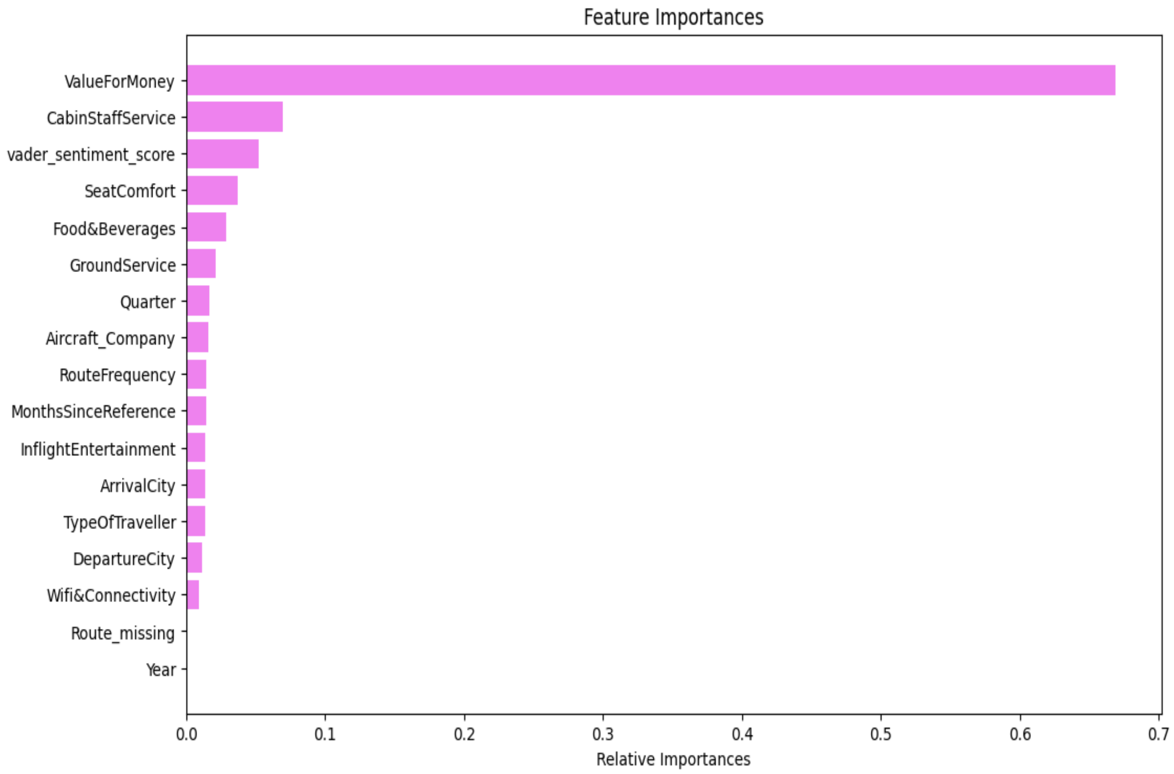


Figure 4.18: Feature Importance

The feature importance chart derived from the XGBoost model provides shows which factors most influence customer recommendations for an airline.

1. Value for Money feature is the most important, far surpassing all other factors. It suggests that the primary determinant of whether customers recommend the airline is their perception of the value they receive for the price they pay. This underscores the importance of pricing strategies and ensuring that customers feel they are getting good value for their money, whether through competitive fares, inclusive services, or added benefits.
2. Cabin Staff Service is the second most important feature. This indicates that passengers place a high value on their interactions with the staff, and positive experiences in this area significantly boost the likelihood of a recommendation. This highlights the need for continuous investment in staff training and maintaining high standards of customer service.
3. VADER Sentiment Score which reflects the tone of customer reviews, also ranks highly in importance. This suggests that the way customers express their experiences in written reviews—whether positive or negative—strongly correlates with their likelihood of recommending the airline. Positive sentiments in reviews are likely to be associated with higher

recommendations, making sentiment analysis a useful tool for understanding customer satisfaction.

4. Seat Comfort during the flight is another critical factor. Comfortable seating contributes significantly to overall passenger satisfaction, influencing whether they would recommend the airline to others. Ensuring that seats are comfortable and meet the expectations of passengers can have a direct impact on recommendations.

5. The quality and variety of food and beverage offerings also play an important role. Passengers who are satisfied with the in-flight dining experience are more likely to recommend the airline, suggesting that attention to catering and menu quality is an essential aspect of service.

6. Ground Service, such as check-in and boarding, are also significant but less so than the in-flight experience. Nevertheless, smooth and efficient ground services contribute positively to the overall travel experience and can enhance the likelihood of a recommendation.

7. The time of year (quarter) has a noticeable impact, possibly due to seasonal variations in service or changes in passenger expectations during different times of the year. This indicates that airlines may need to adjust their strategies and service levels depending on the season to maintain high customer satisfaction.

8. The airline itself, represented by the "Aircraft_Company" feature, shows some importance, suggesting that different airlines may have varying levels of service quality that affect customer recommendations.

4.5.2 Classification Models Using TF-IDF for Text Data

Table 4.4: Classification Models

Models	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.89	0.89	0.89	0.89
Naive Bayes	0.84	0.85	0.84	0.84
Support Vector Machine	0.87	0.88	0.87	0.87
XGBoost	0.91	0.91	0.91	0.91

The logistic regression performance demonstrates a 89% performance accross all metrics. The NB performance demonstrates a 84% lower performance on accuracy, recall, F1 score and 85% on precision. The SVM performance demonstrates a slightly improved performance of

87% on accuracy, recall, F1 score and 88% on precision. The XGBoost models developed using both VADER sentiment analysis and TF-IDF stages outperformed other classification models. Overall, the VADER sentiment models performed better than the TF-IDF models, indicating that VADER sentiment was more effective in extracting relevant features from customer reviews. Additionally, the classification models built with VADER sentiment scores and other categorical variables showed relatively less overfitting compared to those using TF-IDF.

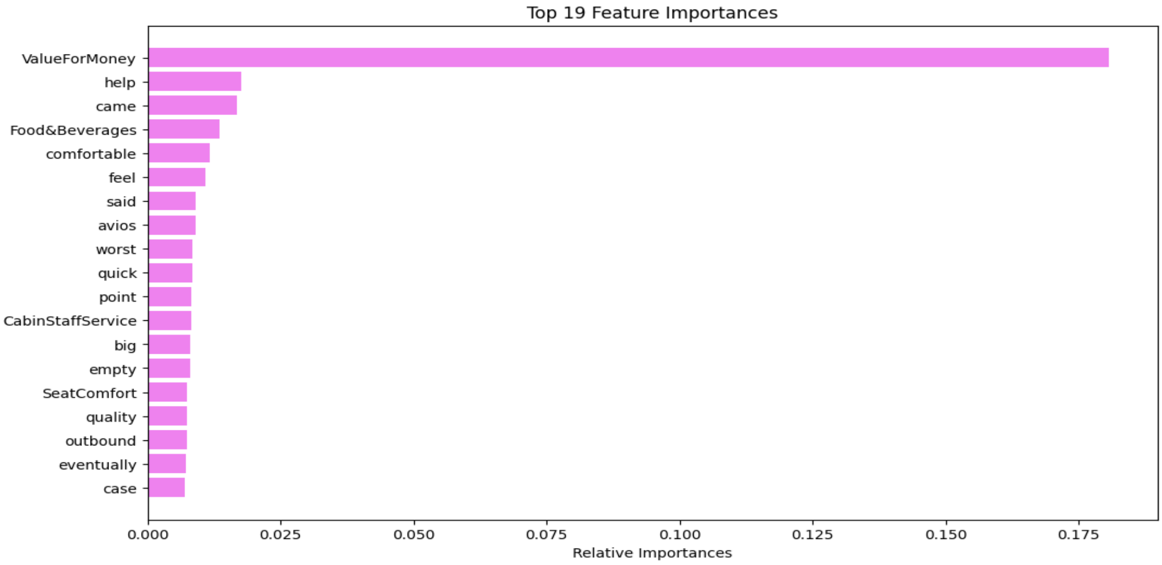


Figure 4.19: Top 19 Feature Importance

The visualization of the top 19 contributing features in the XGBoost model using TFIDF vectors from the written reviews reveals show that both texts from the user reviews and categorical features play crucial roles in determining the target outcome.

i. ValueForMoney feature consistently stands out as the most important predictor. This aligns with previous models, indicating that customers' perception of value significantly impacts their recommendations.

ii. Several words from the text review s emerged as highly influential features. Notable examples include:

   a) help: Indicates that assistance and customer service are critical.

   b) came: Might be related to the timing or arrival aspects of the service.

   c) comfortable: Reflects passengers' comfort levels, a key aspect of their overall experience.

   d) feel: Suggests that emotional responses and feelings about the service are pivotal.

   e) said: Could relate to communication and statements made by the staff or passengers.

Other 3 categorical cariables that contributed to user satisfaction include

i. CabinStaffService: The quality of cabin staff service is a significant factor, emphasizing the importance of staff interactions.
ii. Food&Beverages: The availability and quality of food and beverages contribute notably to the overall experience. SeatComfort: The comfort of the seats is crucial for passenger satisfaction.

4.6 Hyperparameter Tuning

Since the best performing model so far is the XGBoost model built using vader sentiment score and other correlated features. The model will undergo hyperparameter tuning to further enhance it.

Table 4.5: XGBoost

Accuracy	Precision	Recall	F1 Score
0.94	0.94	0.94	0.94

The XGBoost model, after undergoing hyperparameter tuning, demonstrates strong performance with an accuracy of 94% across all metrics on the test sets. This high level of accuracy indicates that the model is highly effective at making predictions with relatively low signs of overfitting, making it the best-performing model so far.

The confusion matrix provides further insights into the model’s performance. The model correctly predicted 316 instances as "Not Recommended," which are referred to as True Negatives (TN). It also accurately predicted 204 instances as "Recommended," known as True Positives (TP). However, there were 20 instances where the actual recommendation status was "Recommended," but the model incorrectly predicted them as "Not Recommended"; these are False Negatives (FN). Additionally, there were 14 instances where the model incorrectly predicted a "Not Recommended" service as "Recommended," which are False Positives (FP). The model’s ability to make accurate predictions, coupled with the low number of misclassification, underscores its robustness and reliability for predicting whether an airline service will be recommended.

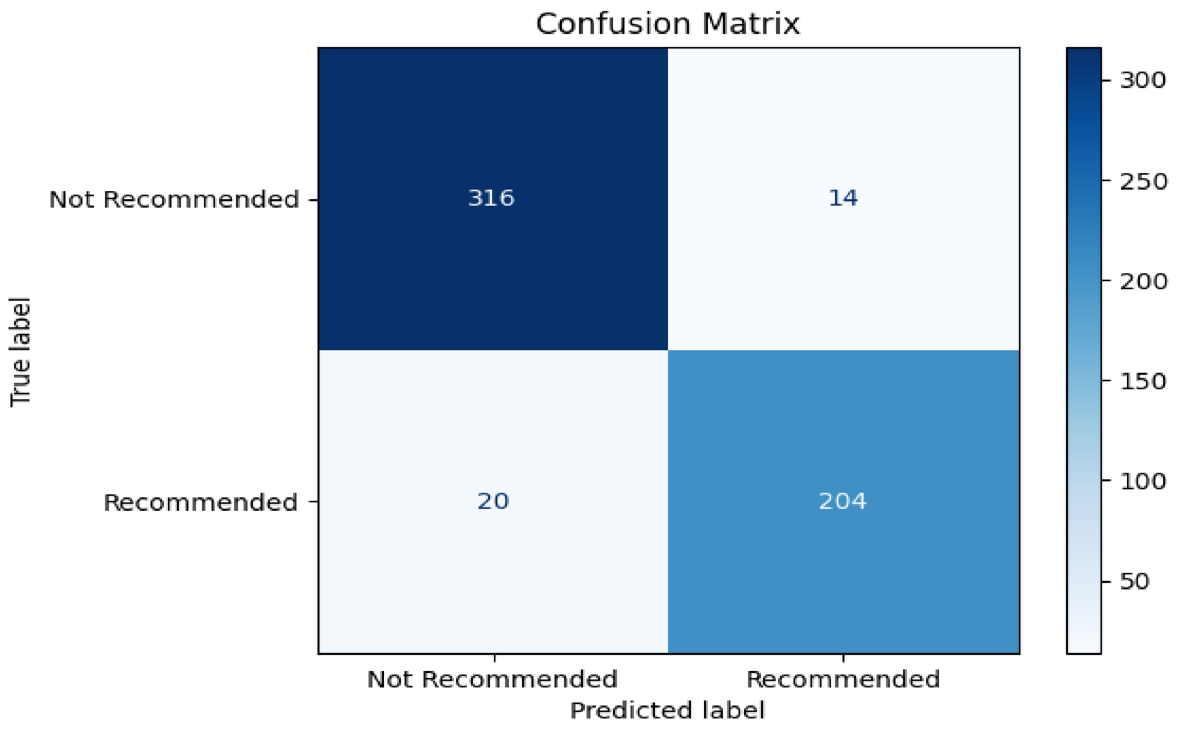


Figure 4.20: Confusion Matrix for XGBoost

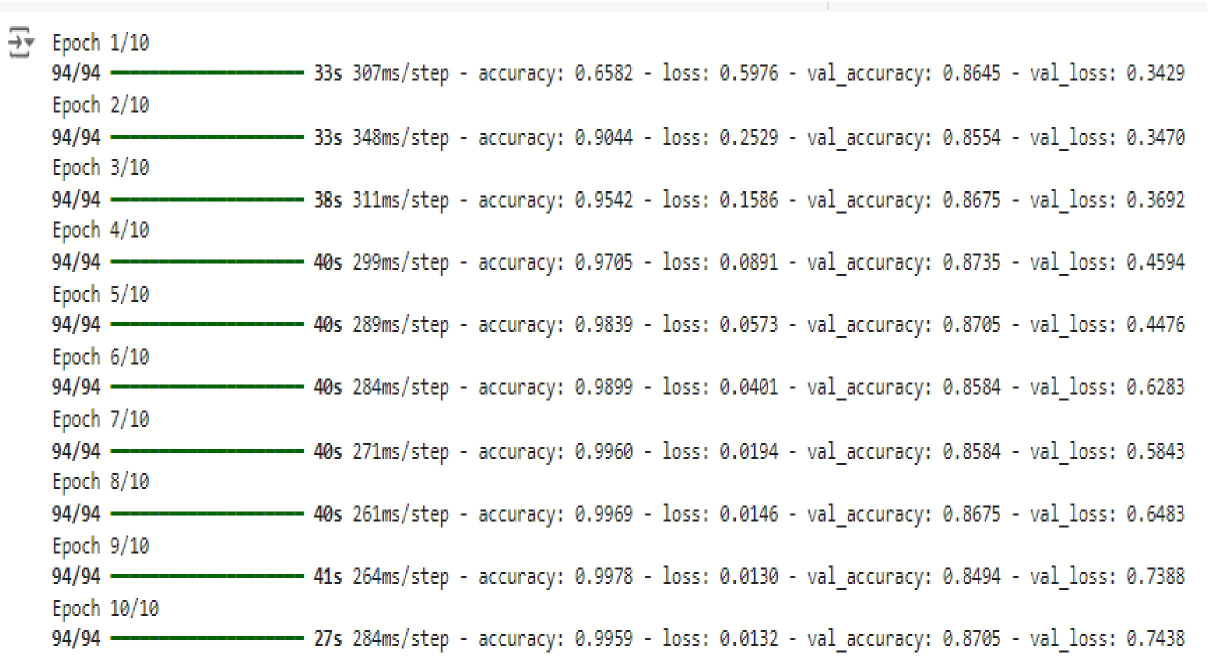
4.6.1 Cross-validation of XGboost using K-Fold

Cross-validation on the training set helps assess the model's performance and generalizability during training. It provides insights into how well the model might perform on unseen data by dividing the training data into multiple folds.

The cross-validation results of the XGBoost model, using a 5-fold cross-validation method, provide valuable insights into the model's performance and generalizability. Cross-validation is a technique that helps to evaluate the model's performance on unseen data by dividing the training set into multiple folds, thereby simulating different scenarios where the model is tested on different subsets of the data.

The XGBoost model achieved a mean accuracy of 92.49% across the five folds. This high accuracy indicates that the model performs well on the training data and is likely to make accurate predictions when applied to new, unseen data. The consistency of this accuracy across multiple folds further validates the model's reliability. The standard deviation of the accuracy scores is 0.38%, which is very low. A low standard deviation implies that the model's performance is stable across different folds of the training data. This means that the model's accuracy is not heavily influenced by which subset of data it is trained on, indicating good generalizability and a low likelihood of overfitting. The accuracy scores for the individual folds range from 91.86% to 93.02%, which are all very close to the mean accuracy. This uniformity across the folds suggests that the model is not only effective but also robust, as it performs consistently across various splits of the data.

4.7 Deep Learning Using LSTM



```
Epoch 1/10
94/94 ━━━━━━━━━━━━━━━━━━━━ 33s 307ms/step - accuracy: 0.6582 - loss: 0.5976 - val_accuracy: 0.8645 - val_loss: 0.3429
Epoch 2/10
94/94 ━━━━━━━━━━━━━━━━━━━━ 33s 348ms/step - accuracy: 0.9044 - loss: 0.2529 - val_accuracy: 0.8554 - val_loss: 0.3470
Epoch 3/10
94/94 ━━━━━━━━━━━━━━━━━━━━ 38s 311ms/step - accuracy: 0.9542 - loss: 0.1586 - val_accuracy: 0.8675 - val_loss: 0.3692
Epoch 4/10
94/94 ━━━━━━━━━━━━━━━━━━━━ 40s 299ms/step - accuracy: 0.9705 - loss: 0.0891 - val_accuracy: 0.8735 - val_loss: 0.4594
Epoch 5/10
94/94 ━━━━━━━━━━━━━━━━━━━━ 40s 289ms/step - accuracy: 0.9839 - loss: 0.0573 - val_accuracy: 0.8705 - val_loss: 0.4476
Epoch 6/10
94/94 ━━━━━━━━━━━━━━━━━━━━ 40s 284ms/step - accuracy: 0.9899 - loss: 0.0401 - val_accuracy: 0.8584 - val_loss: 0.6283
Epoch 7/10
94/94 ━━━━━━━━━━━━━━━━━━━━ 40s 271ms/step - accuracy: 0.9960 - loss: 0.0194 - val_accuracy: 0.8584 - val_loss: 0.5843
Epoch 8/10
94/94 ━━━━━━━━━━━━━━━━━━━━ 40s 261ms/step - accuracy: 0.9969 - loss: 0.0146 - val_accuracy: 0.8675 - val_loss: 0.6483
Epoch 9/10
94/94 ━━━━━━━━━━━━━━━━━━━━ 41s 264ms/step - accuracy: 0.9978 - loss: 0.0130 - val_accuracy: 0.8494 - val_loss: 0.7388
Epoch 10/10
94/94 ━━━━━━━━━━━━━━━━━━━━ 27s 284ms/step - accuracy: 0.9959 - loss: 0.0132 - val_accuracy: 0.8705 - val_loss: 0.7438
```

Figure 4.21: Deep Learning

During the training process, which spans 10 epochs, the model shows a significant increase in accuracy. Initially, in the first epoch, the model achieves a training accuracy of 65.82%. As the epochs progress, this accuracy improves dramatically, reaching an impressive 99.95% by the final epoch. This indicates that the model has effectively learned the patterns in the training data. However, when we examine the model's performance on validation data, which measures its ability to generalize to new, unseen data, the results are slightly different. The validation accuracy starts at 86.45% and stabilizes around 87.05% by the end of the training. This pattern indicates that while the model performs exceptionally well on the training data, it is likely overfitting.

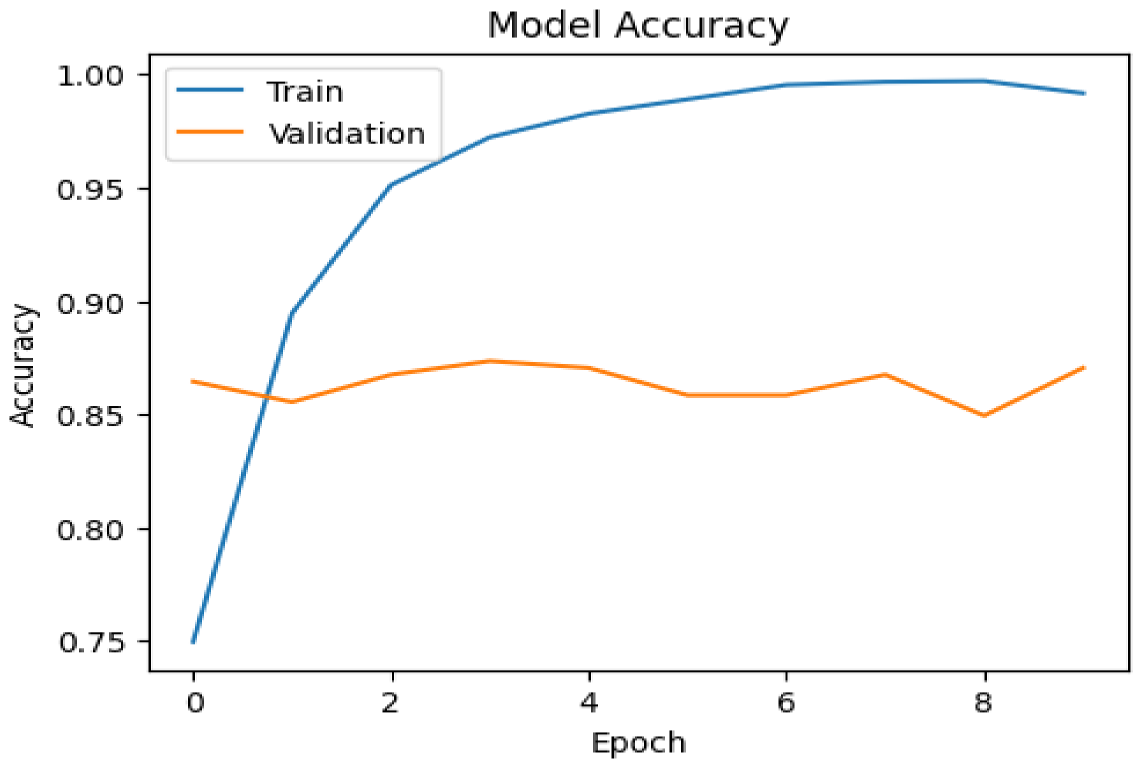


Figure 4.22: Model Accuracy

The Model Accuracy graph indicates that while the LSTM model performs exceptionally well on the training data, achieving near-perfect accuracy, its performance on the validation data did not improve.

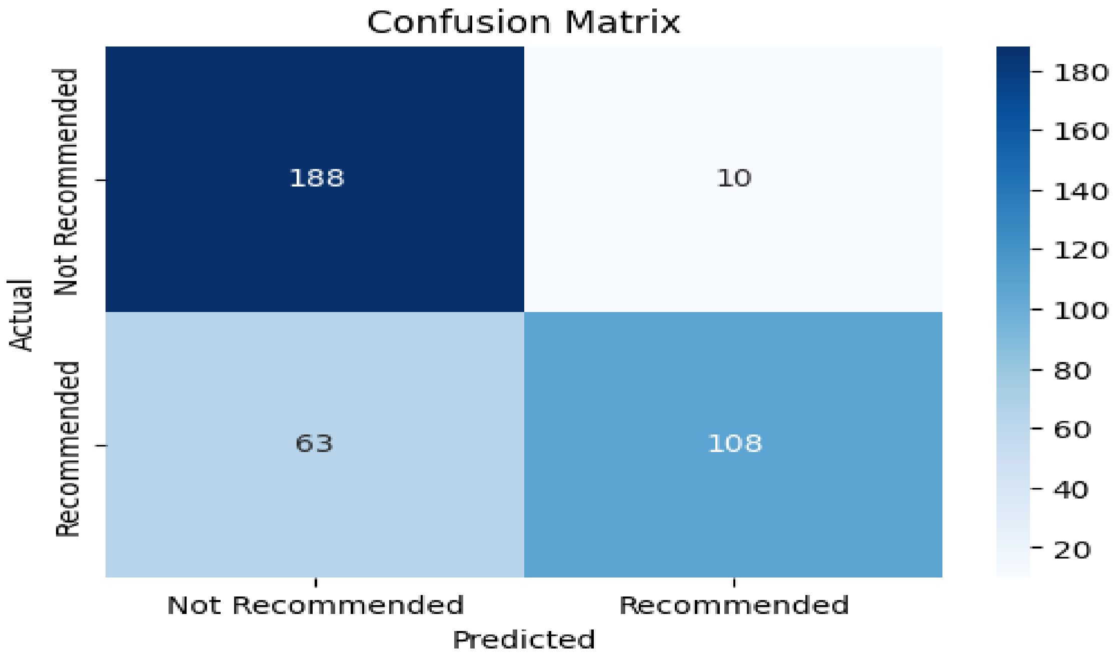


Figure 4.23: Confusion Matrix

Table 4.6 Classification Report

	Precision	Recall	F1 Score	Support
Not Recommended	0.75	0.95	0.84	198
Recommended	0.92	0.63	0.75	171
Accuracy			0.80	369
macro avg	0.83	0.79	0.79	369
weight avg	0.83	0.80	0.80	369

For "Not Recommended" reviews, the LSTM model achieved a high precision of 0.75 and an impressive recall of 0.95, leading to an F1-score of 0.84. For "Recommended" reviews, the model had a precision of 0.92 and a lower recall of 0.63, resulting in an F1-score of 0.75. Overall accuracy is 0.80.

The confusion matrix shows that the model correctly classified 108 "Not Recommended" reviews and misclassified 10, while it correctly identified 108 "Recommended" reviews and misclassified 68. This suggests the deep learning model is more effective at identifying "Not Recommended" reviews.

Chapter 5

Discussion of Findings

5.1 Introduction

This section presents the findings of this research based on the research questions and compared it with similar studies to corroborate of oppose the findings.

5.2 Discussion of Findings

The first research question investigates the effectiveness of NLP techniques, specifically VADER sentiment analysis and TextBlob, in accurately identifying and classifying sentiment expressed in customer reviews. TextBlob demonstrated a solid performance in identifying positive sentiments, achieving a recall of 95% for positive sentiments, indicating that the model correctly identified the vast majority of actual positive sentiments. However, its lacking with precision, particularly in classifying negative sentiments, resulting in a significant number of false positives (1152). This suggests that while TextBlob is effective in capturing positive sentiments, it has a tendency to over-predict them, leading to a lower precision for positive sentiment predictions (55%) and an overall accuracy of 67%. On the other hand, VADER exhibited a more balanced performance. It reduced the number of false positives to 1007, a notable improvement over TextBlob, and maintained a relatively low number of false negatives. VADER's accuracy was higher at 70%, with a more balanced precision and recall across both positive and negative sentiments. This indicates that VADER provides a more reliable and accurate sentiment classification, particularly by reducing the overestimation of positive sentiments and offering a more nuanced analysis of customer feedback. These findings align with a study by Hutto & Gilbert (2014), who developed VADER, demonstrated that VADER performs well in social media contexts and is capable of providing sentiment analysis, particularly in environments with short, informal text. Also, Ribeiro et al. (2016), which found that VADER outperforms TextBlob in scenarios requiring a more balanced approach to sentiment analysis, especially in reducing false positives. Loria (2018) noted that TextBlob is particularly useful for simpler applications where high recall for positive sentiment is desired, though it may not always provide the precision necessary for more critical sentiment classification tasks.

The second research question tends to address the effectiveness of Exploratory Data Analysis (EDA) in identifying key patterns, relationships, and insights within customer review data, and how do these factors influence overall customer satisfaction and recommendations. Findings shows that in distribution of overall rating, customer satisfaction is highly skewed towards the lower end of the rating scale. This show that a significant portion of customers were extremely dissatisfied with their experiences. The gradual decrease in the number of reviews as the rating increases indicates that fewer customers felt highly satisfied. This pattern underscores the importance of addressing the causes of dissatisfaction to shift the distribution towards higher satisfaction levels.

The analysis of average overall ratings by aircraft type shows significant variations in customer satisfaction across different models. The Saab 2000 and Boeing 747 Family, received the highest average ratings, indicate that customers tend to have more favourable experiences on these aircraft. Conversely, aircraft like the Boeing 757 and Airbus A321 Family have the lowest average ratings, suggesting that these models may contribute to lower customer satisfaction. The variability in ratings across different aircraft types highlights the importance of factors such as comfort, space, and in-flight services in shaping customer perceptions. For airlines, this insight emphasizes the need to consider aircraft type when evaluating and improving service quality.

Further, the distribution of passengers across different seat types, reveals that over half of the travellers opt for Economy Class, making it the most popular choice. This suggests that cost-effective travel options are a priority for the majority of customers. Business Class, which accounts for a significant portion of travellers, indicates a substantial demand for enhanced comfort and services. The relatively smaller percentages for Premium Economy and First Class suggest these are niche segments, appealing to customers seeking additional luxury and comfort. Understanding this distribution is crucial for airlines when tailoring their service offerings to meet the needs and preferences of different customer segments. The bar chart in figure 4.7 shows the verified review count by different traveller types which highlights that couples on leisure trips are the most likely to provide feedback, followed by solo leisure travellers. Business and family leisure travellers contribute fewer reviews. This distribution suggests that leisure travellers, particularly couples, are more engaged in sharing their experiences, possibly because they are more focused on the overall experience and have higher expectations. For airlines, this indicates a need to pay particular attention to the service quality provided to leisure travellers, as they are more likely to influence the airline's reputation through reviews.

Additionally, the relationship between value for money and the likelihood of recommending the airline, as depicted in figure 4.8, shows a clear positive correlation. Customers who perceive they received good value for money are more likely to recommend the airline. At the lowest value for money rating of 1, the majority of customers do not recommend the airline. However, as the value for money rating increases, the number of recommendations rises significantly.

This finding emphasizes that perceived value for money is a critical determinant of customer satisfaction and recommendations. Airlines can enhance customer loyalty by focusing on pricing strategies and ensuring that customers feel they are receiving good value for the services provided.

The line graph in figure 4.9 displaying the average recommendation status over the years highlights significant fluctuations in customer satisfaction. The peak in 2014 suggests a period of high customer satisfaction, possibly due to improvements or innovations in service during that year. However, the sharp decline after 2014, reaching a low point in 2017 and again in 2023, indicates periods of declining customer satisfaction. The variability in recommendation status over the years suggests that external factors, such as changes in service quality, industry trends, or economic conditions, can significantly impact customer satisfaction. Understanding these trends can help airlines identify and address the causes of declining satisfaction and work towards sustaining higher levels of customer satisfaction over time. Also, the monthly fluctuations in recommendation status indicate that customer satisfaction varies significantly throughout the year. The sharp decline in February and June, followed by peaks in May and November, suggests that seasonal factors, specific events, or operational changes may influence customer experiences at different times of the year. For BA, this insight highlights the importance of maintaining consistent service quality year-round and being mindful of potential seasonal impacts on customer satisfaction. By anticipating and addressing these fluctuations, airlines can improve overall satisfaction and reduce the variability in customer recommendations.

The box plot in figure 4.10 analysing seat comfort ratings across different traveller types reveals high variability in customer experiences. While couple leisure and family leisure travellers show a wide range of ratings, business travellers tend to rate their comfort slightly lower. This suggests that seat comfort is a critical factor influencing satisfaction across all traveller types, but particularly for business travellers who may have higher expectations for comfort. This finding underscores the importance of ensuring consistent seat comfort across all traveller types to enhance overall satisfaction. The box plot in figure 4.11 comparing overall ratings between customers who would recommend the service and those who would not clearly shows that higher overall ratings are associated with positive recommendations. This clear distinction highlights the strong relationship between overall satisfaction and the likelihood of recommending the service. This insight reinforces the importance of striving for high overall satisfaction to encourage positive word-of-mouth and customer loyalty.

Additionally, the positive correlation between cabin staff service quality and overall ratings, suggests that the behaviour and efficiency of cabin staff play a significant role in shaping customer satisfaction. Higher ratings for cabin staff service are consistently associated with higher overall ratings, indicating that excellent service by cabin staff is closely linked to customer satisfaction. This finding highlights the importance of investing in staff training and maintaining high standards of customer service, as this can directly impact the overall customer experience and lead to higher satisfaction and recommendations. The analysis of recommendation status based on food and beverage service ratings shows a clear divide between low and high ratings. Customers who rate the food and beverage service highly are much more likely to recommend the airline, while those who give low ratings are less likely to do so. This finding emphasizes the importance of in-flight dining experiences in shaping customer perceptions and influencing their likelihood of recommending the service. Airlines can enhance customer satisfaction and recommendations by focusing on the quality and variety of food and beverage offerings.

The sentiment score distribution reveals that most customer reviews are classified as positive, with a significant portion of reviews expressing negative sentiments and very few neutral sentiments. This polarization of sentiments suggests that customers tend to have strong opinions about their experiences, either positive or negative, with little middle ground. For airlines, this insight highlights the importance of addressing both positive and negative feedback to understand the factors driving extreme sentiments and to work towards balancing customer experiences. The heatmap analysis provides a comprehensive view of the correlations between various features related to airline service and customer satisfaction. The strongest correlation observed is between overall rating and value for money, indicating that customers who perceive good value for money are more likely to rate the service highly. The high correlation between overall rating and the likelihood of recommending the service further emphasizes the importance of maintaining high service standards. Factors like food and beverage quality, seat comfort, and cabin staff service also show strong correlations with overall satisfaction, indicating that these aspects are critical in shaping customer perceptions. The relatively weaker correlations with inflight entertainment, ground service, and Wi-Fi connectivity suggest that while these factors contribute to satisfaction, their impact is less significant compared to the core aspects of service quality.

Research question four compares the effectiveness of various machine learning models in predicting customer satisfaction and recommendation outcomes using VADER sentiment scores and TF-IDF features. Among the models evaluated using VADER sentiment scores, Support Vector Machine (SVM) demonstrated the highest performance with an accuracy, precision, recall, and F1 score of 94% across all metrics. This suggests that SVM is highly effective in utilizing the sentiment analysis data to predict customer satisfaction and recommendation outcomes. The XGBoost model, which also performed well with a 93% accuracy across all metrics, showed signs of overfitting during training but maintained strong performance after applying pre-pruning techniques to mitigate this issue. The high performance of both SVM and XGBoost indicates that these models are particularly well-suited for handling sentiment analysis in predicting customer outcomes. This is consistent with the findings of Jindal et al. (2019), who found that SVM often excels in text classification tasks due to its ability to handle high-dimensional data effectively, such as TF-IDF vectors and sentiment scores. However, the overfitting observed in XGBoost aligns with study of Chen & Guestrin (2016), who noted that tree-based models like XGBoost can be prone to overfitting without careful tuning.

When using TF-IDF features, the XGBoost model again performed well, achieving a 91% accuracy, slightly outperforming the other models, including Logistic Regression and SVM. However, the overall performance of the models using TF-IDF features was slightly lower compared to those using VADER sentiment scores. The logistic regression model demonstrated a consistent performance of 89% across all metrics, but it did not surpass the performance of models trained on VADER sentiment scores. The lower performance of TF-IDF models suggests that while TF-IDF is effective in capturing the importance of words within the text, it may not be as powerful as sentiment scores when it comes to predicting customer satisfaction and recommendation outcomes. This is corroborated by the work of Zhang et al. (2023), who found that sentiment scores often provide a more direct and interpretable measure of customer sentiment, leading to better predictive performance in models.

Further, the feature importance analysis in the XGBoost models highlights that "Value for Money" is the most significant predictor of customer satisfaction and recommendations, far surpassing other factors. This suggests that customers' perceptions of receiving good value for their money are the primary determinants of whether they will recommend the airline. This finding is supported by the research of Ye et al. (2019), who emphasized that perceived value is a critical factor in customer satisfaction across various service industries, including airlines. Other important features include "Cabin Staff Service," which indicates that high-quality interactions with staff greatly influence customer recommendations, and "VADER Sentiment

Score," reflecting that the tone of customer reviews strongly correlates with their likelihood of recommending the airline. The significance of sentiment scores is in line with the findings of Jain et al. (2019), who highlighted the effectiveness of sentiment analysis in predicting customer behaviour in the service industry.

After hyperparameter tuning, the XGBoost model showed further improvement, achieving an accuracy of 94% across all metrics on test sets, demonstrating strong generalization capabilities with minimal overfitting. The cross-validation results, with a mean accuracy of 92.49% and a low standard deviation of 0.38%, further confirm the robustness of the XGBoost model. These findings are consistent with the research by Prokhorenkova et al. (2019), who found that well-tuned XGBoost models can achieve high accuracy and stability across different datasets.

Finally the fourth research question seeks to explore the performance of a Deep Learning model, such as an LSTM (Long Short-TerMemory) network, in predicting customer satisfaction and recommendation outcomes based on customer review data. The result showed that the LSTM model showed a significant increase in accuracy, starting at 65.82% and reaching an accuracy of 99.95% by the final epoch during the training process. This rapid improvement in training accuracy indicates that the LSTM model is highly effective at learning the patterns within the training data. However, the performance on the validation data, which measures the model's ability to generalize to unseen data, was considerably lower, stabilizing around 87.05%.

Yin et al. (2019), pointed out that while LSTM models are powerful in capturing complex patterns in sequential data, they are also prone to overfitting, particularly when not adequately regularized or when the dataset is not sufficiently large to support the model's complexity. When comparing the LSTM model’s performance to traditional machine learning models like Support Vector Machine (SVM) and XGBoost, the overall accuracy of the LSTM model on the validation set was 80%, which is lower than 93% accuracy achieved by XGBoost using VADER sentiment scores. Overall, the XGBoost model using VADER for text feature extraction (sentiment score) performed better than the LSTM deep learning model. This superior performance may be attributed to the dataset's structure, which included categorical variables more effectively handled by tree-based models like XGBoost. The XGBoost model demonstrated an impressive 94% accuracy across all metrics on the unseen/test data after hyperparameter tuning.

Additionally, the LSTM model showed a precision of 0.75 for "Not Recommended" reviews and 0.92 for "Recommended" reviews, with an F1-score of 0.84 and 0.75, respectively.

Although these metrics demonstrate that the LSTM is capable of making accurate predictions, especially in terms of recall for negative reviews (0.95), its overall performance does not surpass that of the traditional models. The lower validation accuracy and signs of overfitting in the LSTM model suggest that while deep learning models like LSTM are powerful, they require careful tuning and large datasets to outperform traditional machine learning models. This observation is supported by the findings of Young et al. (2018), who noted that while LSTMs can excel in tasks involving sequential data, their performance can be inconsistent when applied to tasks where traditional models, such as SVM and XGBoost, have already proven to be highly effective. The classification report for the LSTM model shows a high recall of 0.95 for "Not Recommended" reviews, indicating that the model is particularly effective at identifying negative reviews. However, the lower recall of 0.63 for "Recommended" reviews suggests that the model struggles to identify positive recommendations accurately, potentially missing many positive cases. This imbalance in recall highlights a potential area for improvement in the LSTM model, particularly in its application to real-world customer review data.

Conclusion

6.1 Introduction

This section concludes the research by summarizing the study's findings. It also highlights areas for future research, reflects on the study's limitations, and suggests possible improvements.

6.2 Summary of Findings

This study aimed to investigate the effectiveness of various NLP (NLP) techniques and machine learning models in predicting customer satisfaction and recommendation outcomes based on customer review data. TextBlob and VADER sentiment analysis were evaluated for their ability to classify sentiments in customer reviews. TextBlob demonstrated strong recall for positive sentiments (95%) but struggled with precision, particularly in classifying negative sentiments, leading to a high number of false positives. This resulted in an overall accuracy of 67%, indicating a tendency to over-predict positive sentiments. VADER, on the other hand, provided a more balanced performance with an accuracy of 70%. It reduced the number of false positives and maintained a relatively low number of false negatives, offering a more reliable sentiment classification. This finding aligns with prior studies that suggest VADER's effectiveness in scenarios requiring balanced sentiment analysis.

The distribution of overall ratings revealed a significant skew towards lower ratings, indicating widespread customer dissatisfaction. Higher satisfaction levels were less frequent, highlighting the importance of addressing key pain points in service delivery. Variations in customer satisfaction were observed across different aircraft types, with the Saab 2000 and Boeing 747 Family receiving higher ratings, whereas the Boeing 757 and Airbus A321 Family had lower ratings. This suggests that factors like comfort, space, and in-flight services play crucial roles in shaping customer perceptions. The analysis of seat type preferences showed that Economy Class was the most popular, followed by Business Class, indicating a demand for cost-effective travel and enhanced services. Understanding these preferences is essential for tailoring service offerings to meet customer needs. Also, the correlation between value for money and the likelihood of recommending the airline underscored the importance of perceived value in driving customer recommendations. Higher value-for-money ratings were strongly associated with increased recommendations. Among the models evaluated using VADER sentiment scores, the Support Vector Machine (SVM) outperformed other models with a 94% accuracy, precision, recall, and F1 score. XGBoost also performed well with a 93% accuracy, although it showed signs of overfitting, which were mitigated through pre-pruning techniques.

When using TF-IDF features, XGBoost again demonstrated strong performance with a 91% accuracy, slightly surpassing the other models, including Logistic Regression and SVM. However, the overall performance of models using TF-IDF features was slightly lower compared to those using VADER sentiment scores, indicating that sentiment scores may provide more direct and interpretable insights into customer sentiment. The LSTM model showed significant improvement during training, reaching nearly perfect accuracy on the training data. However, its validation performance was lower, stabilizing around 87%, indicating potential overfitting. Compared to XGBoost, the LSTM's performance was lower, with an overall accuracy of 80% on the validation set. The findings suggest that while LSTM is effective at capturing complex patterns, traditional models like XGBoost may offer more robust and reliable performance, particularly in structured datasets with categorical variables. Overall, the study found that XGBoost using VADER sentiment scores was the most effective model for predicting customer satisfaction and recommendation outcomes, outperforming both traditional machine learning models and the LSTM deep learning model. This highlights the importance of model selection and feature engineering in achieving optimal predictive performance in sentiment analysis tasks.

6.3 Recommendation

Based on the findings of this research, the following recommendations are made:

i. Airline service companies should prioritize strategies that improve customers' perceptions of value for money, as this was identified as the most significant predictor of customer satisfaction and recommendations.

ii. The quality of cabin staff service was the second most important factor influencing customer recommendations. Airlines should invest in continuous training and development programs for their staff to ensure high standards of customer service. This focus on improving customer interactions can lead to higher satisfaction and increased likelihood of positive recommendations.

iii. Given the importance of seat comfort, food, and beverage quality in influencing customer satisfaction, airlines should focus on upgrading these aspects of their service. Ensuring that seats are comfortable and that the quality of in-flight meals meets or exceeds customer expectations can significantly enhance the overall travel experience and boost recommendations.

iv. The analysis showed that customer satisfaction and recommendation status fluctuate throughout the year. Airlines should continuously monitor these trends and adapt their service strategies to address seasonal variations in customer expectations and experiences.

6.4 Suggestion for Future Study

i. Future research should explore the use of more advanced NLP (NLP) techniques, such as BERT (Bidirectional Encoder Representations from Transformers), to improve sentiment analysis accuracy.

ii. Future research can use the application of ensemble learning techniques or the development of hybrid models that combine the strengths of both traditional machine learning algorithms and deep learning models.

Reference

Airbus. (2019). British Airways receives its first A350. Retrieved from https://www.airbus.com/newsroom/

Airlines International. (2018). The Role of AI in the Airline Industry. Retrieved from https://airlines.iata.org/news/the-role-of-ai-in-the-airline-industry

Annamalai, R., Rasool, S. A., Deena, S., Venkatraman, K., & Soundaram, Y. (2024, February). Sentiment Analysis using VADER: Unveiling Customer Sentiment and Predicting Buying Behavior in the Airline Industry. In *2024 IEEE International Conference for Women in Innovation, Technology & Entrepreneurship (ICWITE)* (pp. 277-282). IEEE.

Arpita, H. D., Al Ryan, A., Hossen, M. S., & Rahman, M. S. (2023, December). Airline Sentiments Unplugged: Leveraging Deep Learning for Customer Insights. In *2023 5th International Conference on Sustainable Technologies for Industry 5.0 (STI)* (pp. 1-6). IEEE.

Azman, M., & Sharma, K. (2022). Smart Boarding System with e-Passports for Secure and Independent Interoperability. *SN Computer Science*, *3*(1), 52.
Boeing. (2019). Boeing 787 Dreamliner. Retrieved from https://www.boeing.com/commercial/787/

British Airways. (2019). Sustainability Report. Retrieved from https://www.britishairways.com/en-gb/information/about-ba/community-and-environment
CAPA - Centre for Aviation. (2019). Low-Cost Carriers: Global Market Share Analysis. Retrieved from https://centreforaviation.com/

Ceken, S. (2024). Cleared for Takeoff: Exploring Digital Assistants in Aviation. In *Harnessing Digital Innovation for Air Transportation* (pp. 1-24). IGI Global.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).
Das, D., Sharma, S., Natani, S., Khare, N., & Singh, B. (2017). Sentimental

Analysis for Airline Twitter data. *IOP Conference Series: Materials Science and Engineering*, 263. https://doi.org/10.1088/1757-899X/263/4/042067.
Delta Air Lines. (2019). "Personalizing Passenger Experience with AI." Retrieved from Delta\

Fanni, S. C., Febi, M., Aghakhanyan, G., & Neri, E. (2023). Natural language processing. In *Introduction to Artificial Intelligence* (pp. 87-99). Cham: Springer International Publishing.

Frank, E., Oluwaseyi, J., & Olaoye, G. (2024). Data preprocessing techniques for NLP in BI.

Gentsch, P., & Gentsch, P. (2019). AI best and next practices. *AI in Marketing, Sales and Service: How Marketers without a Data Science Degree can use AI, Big Data and Bots*, 129-247.

Hasib, K. M. (2022). *Sentiment analysis on Bangladesh airlines review data using machine learning* (Doctoral dissertation, Brac University).

Helgo, M. (2023). Deep Learning and Machine Learning Algorithms for Enhanced Aircraft Maintenance and Flight Data Analysis. *J. Robotics Spectrum*, *1*, 90-99.
Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 216-225).

IAG - International Airlines Group. (2019). Annual Report and Accounts. Retrieved from https://www.iairgroup.com/en/investors-and-shareholders

IATA - International Air Transport Association. (2019). Industry Statistics. Retrieved from https://www.iata.org/en/publications/statistics/

ICAO - International Civil Aviation Organization. (2019). Fuel Cost and Airline Profitability. Retrieved from https://www.icao.int/

Idris, S. L., & Mohamad, M. (2023). A Study on Sentiment Analysis on Airline Quality Services: A Conceptual Paper. *Information Management and Business Review*, *15*(4 (SI) I), 564-576.

Jain, P. K., Pamula, R., & Srivastava, G. (2021). A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews. *Computer science review*, *41*, 100413.

Jaiswal, A. (2022, January 13). *NLP Tutorials Part -I from Basics to Advance*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2022/01/nlp-tutorials-part-i-from-basics-to-advance/

Kuhn, K. D. (2018). Using structural topic modelling to identify latent topics and trends in aviation incident reports. *Transportation Research Part C: Emerging Technologies*, *87,* 105-122.

Liu, B. (2020). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press.
Loria, S. (2018). textblob Documentation. *Release 0.15*, *2*(8), 269.

Martins, M., e Quadros, R. C. C., & Barqueiro, A. (2024). Personalization Strategies and Passenger Satisfaction Analysis in Full-Service Airlines: A Study of Lisbon Airport's Leading Carriers. In *Strategic Management and Policy in the Global Aviation Industry* (pp. 173-202). IGI Global.

Merlo, T. R. (2024). Emerging Role of Artificial Intelligence (AI) in Aviation: Using Predictive Maintenance for Operational Efficiency. In *Harnessing Digital Innovation for Air Transportation* (pp. 25-41). IGI Global.
Monika, R., Deivalakshmi, S., & Janet, B. (2019). Sentiment Analysis of US Airlines Tweets Using LSTM/RNN. *2019 IEEE 9th International Conference on Advanced Computing (IACC)*, 92-95. https://doi.org/10.1109/IACC48062.2019.8971592.

Nche, R. E. C. (2023). A Sentiment Analysis of Customers' Complaints in the Airline Industry: The Case of Brussels Airlines.

Nwakanma, C. I., Ogbonna, A. C., Etus, C., Nwifor, E. U., Onyebuchi, J. E., & Ugwueke, E. C. Predictive analytics of customer sentiments towards Nigerian hospitality industry: Case study approach. In *Proc. 3rd International Conference on Intelligent Computing and Emerging Technologies (ICET 2019)* (pp. 60-68).

p.c, Shilpa & Shereen, Rissa & Jacob, Susmi & Vinod, P.. (2021). Sentiment Analysis Using Deep Learning. 930-937. 10.1109/ICICV50876.2021.9388382.

Parde, N. (2023). Natural language processing. *The SAGE Handbook of Human–Machine Communication*, 318.

Park, E., Jang, Y., Kim, J., Jeong, N., Bae, K., & Pobil, A. (2019). Determinants of customer satisfaction with airline services: An analysis of customer feedback big data. *Journal of Retailing and Consumer Services*. https://doi.org/10.1016/J.JRETCONSER.2019.06.009.

Patel, V. (2018). Airport passenger processing technology: a biometric airport journey.

Prabhakar, E., Santhosh, M., Krishnan, A., Kumar, T., & Sudhakar, R. (2019). Sentiment Analysis of US Airline Twitter Data using New Adaboost Approach. *International journal of engineering research and technology,* 7.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems, 31.*

Punel, A., Hassan, L., & Ermagun, A. (2019). Variations in airline passenger expectation of service quality across the globe. *Tourism Management,* 75, 491-508. https://doi.org/10.1016/J.TOURMAN.2019.06.004.

Rane, A., & Kumar, A. (2018). Sentiment Classification System of Twitter Data for US Airline Service Analysis. *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC),* 01, 769-773. https://doi.org/10.1109/COMPSAC.2018.00114.

Research, P. (2022, May 23). *Artificial Intelligence in Aviation Market Report 2022-2030*. https://www.precedenceresearch.com/artificial-intelligence-in-aviation-market

Ribeiro, F. N., Araújo, M., Gonçalves, P., André Gonçalves, M., & Benevenuto, F. (2016). Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science, 5,* 1-29.

Samir, H. A., Abd-Elmegid, L., & Marie, M. (2023). Sentiment analysis model for Airline customers' feedback using deep learning techniques. *International Journal of Engineering Business Management, 15*, 18479790231206019.

Shahid, R., Mozumder, M. A. S., Sweet, M. M. R., Hasan, M., Alam, M., Rahman, M. A., ... & Islam, M. R. (2024). Predicting Customer Loyalty in the Airline Industry: A Machine Learning Approach Integrating Sentiment Analysis and User Experience. *International Journal on Computational Engineering, 1*(2), 50-54.

Sodhar, I. N., Jalbani, A. H., Channa, M. I., & Hakro, D. N. (2019). Parts of speech tagging of Romanized Sindhi text by applying rule based model. *IJCSNS, 19*(11), 91.

Yadav, V., & Bethard, S. (2019). A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470.*

Ye, H., Yang, X., Wang, X., & Stratopoulos, T. C. (2021). Monetization of digital content: Drivers of revenue on Q&A platforms. *Journal of Management Information Systems, 38*(2), 457-483.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, *13*(3), 55-75.

Zhang, Wenkang & Chen, Shaofan & Li, Caimao & Liu, Guanzhong & Jing, Defu. (2023). Sentiment Analysis Method Based on Improved Feature Vector. 1332-1338. 10.1109/ACAIT60137.2023.10528512.

Zhu, X., & Li, L. (2021). Flight time prediction for fuel loading decisions with a deep learning approach. *Transportation Research Part C: Emerging Technologies*, *128,* 103179.