Sentiment Analysis of Airport Google Reviews: A Comparative Study of Natural Language Processing Techniques and Machine Learning Models

MSc Research Project
Master of Science in Artificial Intelligence

Niall Kierans
Student ID: x23311550

School of Computing
National College of Ireland

Supervisor: Dr. Devanshu Anand

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Niall Kierans
Student ID:	x23311550
Programme:	MSCAITOP
Year:	2024
Module:	MSC Artificial Intelligence
Supervisor:	Dr. Devanshu Anand
Submission Due Date:	12th Aug 2024
Project Title:	Sentiment Analysis of Airport Google Reviews: A Comparative Study of Natural Language Processing Techniques and Machine Learning Models
Word Count:	6470
Page Count:	25

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

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Niall Kierans
Date:	12th August 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

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

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

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

Sentiment Analysis of Airport Google Reviews: A Comparative Study of Natural Language Processing Techniques and Machine Learning Models

Niall Kierans
x233115501

Abstract

Airports play a central role in the global travel network, and their efficiency and service quality significantly impact passenger satisfaction. In this study, we explore the potential of Natural Language Processing (NLP) and Machine Learning (ML) techniques to analyse customer sentiments reflected in Google reviews of various airports. A sample of airport reviews was collected from Google, and NLP techniques like tokenisation, stop-word removal, and lemmatisation were applied to prepare the data. Sentiment classification was performed using dictionary-based lexicons (Vader, NRCLex, TextBlob) and then with ML models (SVM, Random Forest, Naïve Bayes). Finally, NLP and ML models were blended for further experimentation.

The analysis revealed key insights into the aspects of airport services that influence passenger sentiment, such as cleanliness, staff behaviour, waiting times, and facilities. The results indicate that the blending of NLP with ML models provides a strong framework for sentiment analysis, offering reliable predictions and valuable insights. The insights gained can guide airport management in making informed decisions to elevate service quality and boost passenger satisfaction.

Keywords *- Airport, NLP, Machine Learning, Google Reviews, SVM, Random Forest, Naive Bayes, Vader, Textblob and NRCLex*

1 Introduction

In our globalised society, European airports play a vital role as key hubs in the international travel network, directly impacting the experiences of millions of passengers each day. The quality of service provided by these airports influences passenger satisfaction, loyalty, and overall travel experience. Traditionally, Airport Service Quality measures (SQMs) have been evaluated through structured surveys and standardised metrics set by the aviation authorities in each country. However, these conventional methods often fall short of capturing the nuanced and real-time sentiments of passengers. The rise of online review platforms like Google Reviews has created new opportunities for analysing customer feedback using advanced techniques such as artificial intelligence sentiment analysis.

Google Map reviews are voluntary opinions shared by the public about places they have visited. These reviews include ratings and written comments, guiding future users make better decisions about businesses or services. Google enforces a strict violation policy to prevent deceptive or inappropriate content which makes it a reliable source of information for sentiment analysis Google (2024b). This makes Google reviews a trustworthy location for users compared to many known websites. Each review includes a five-star rating, with one star being the lowest and five stars the highest. Additionally, you can write detailed reviews about your experiences with a business or service Google (2024a)

Sentiment analysis is a Natural Language Processing (NLP) method designed to evaluate and classify emotions within text. Sentiment analysis reveals whether a piece of writing carries positive, negative, or neutral opinions. This method is greatly used in various fields, including analysing customer reviews, monitoring social media posts, and conducting market research. Sentiment analysis automates the interpretation of emotional tones in large amounts of text data, providing companies with valuable insights into customer attitudes and preferences. These insights are vital for improving decision-making processes and ensuring that business strategies align with customer expectations. This study will compare NLP methods along with Machine Learning (ML) techniques to identify the most effective approach for text mining. The evaluation will consider Vader, Textblob, and NRCLex on the NLP side and from the ML methods use Support Vector Machines (SVM), Random Forest, and Naive Bayes. Then evaluate each ML method with the best lexicon to see if the blended approach yields better results

1.1 Background and Motivation

Airports are now fully operational after low activity during the COVID-19 years and given that airports are profitable again, it would be wise to invest some profits to improve the passenger experience. The Airports Council International of Europe (ACI) is the only professional airport association in the world and covers 500 airports in 45 countries throughout Europe. In a paper published in 2018 by the ACI Europe Europe (2018), An increase of 1 percent in passenger satisfaction generates, on average, growth of 1.5 percent in non-aeronautical revenues. In other words, happier customers are more likely to use airport car park, airport duty-free shops, and eat in the food and beverage units within the airport. A targeted approach to spending is needed, but this is only possible if airports know which services to target.

Some airports in Europe have a regulation obligation and an example of this is the Irish Aviation Authority (IAA) which sets the SQM for Dublin Airport. These SQMs are based on a ten-point scale, where ten is the highest possible score and one is the lowest. The surveys are conducted independently of both the IAA and Dublin Airport and can be administered at any time. For departing passengers, the surveys are typically conducted at the departure gate, while arriving passengers are usually asked for their feedback in the baggage or arrival hall. This data is compiled monthly, and each service quality measurement yields an average score out of ten. The IAA establishes the required passing score for each SQM, and if the target is not met, financial penalties are imposed. The latest IAA termination raised the pass score from 8.0 to 8.5 in some areas of concern. Additionally, the IAA introduced a bonus scheme, where achieving a higher recorded score can reduce any penalties accumulated for missed targets during the same calendar year IAA (2024). A table with examples of the service quality measures can be found in the appendix section for reference.

Monthly and quarterly SQM scores are provided to Dublin Airport operations teams with no context for further information about the score. If a failure happens there is no way to determine where the failures occurred and the associated score for that area. In some cases, the airport fills the gaps using customer survey products from external partners like Happy or NotHappyOrNot (2024). These products are then used to provide passenger feedback so they can identify improvement areas and work towards increasing the SQM scores. While this study does not look to replace any aspects of the current passenger feedback at Dublin Airport it does aim to enhance it by providing additional feedback from passengers using personal reviews and comments. Sentiment analysis is beneficial for assessing airports as it provides deep insights into passenger experiences by analysing reviews and feedback. This information helps airports identify strengths and weaknesses, benchmark against other airports, and make data-driven improvements to services. Additionally, sentiment analysis assists in engaging stakeholders, complying with regulations, and tailoring services to meet passenger needs, ultimately enhancing overall customer satisfaction and travel experience.

2 Research Question & Objectives

This research is based on Google reviews for 6 airports of similar passenger numbers, from April 2023 to March 2024. This includes the most recent data and covers the seasonality aspect of the airport business.

Here are the research questions to be answered during this project:

Can sentiment analysis of Google Reviews provide accurate insights into customer satisfaction for airport services?

1. Which NLP methods returns the best evaluate sentiment from Google reviews?
   (a) Evaluate Textblob model against Google reviews.
   (b) Evaluate Vader model against Google reviews.
   (c) Evaluate NRCLex model against Google reviews.
2. How do NLP methods perform in sentiment analysis at the aspect and sentence level of Google reviews?
   (a) Evaluate Textblob at sentence and aspect level.
   (b) Evaluate Vader at sentence and aspect level.
   (c) Evaluate NRCLex at sentence and aspect level.
3. How do different ML models perform in predicting review ratings from Google reviews?
   (a) Evaluate SVM, Random Forest and Naive Bayes ML methods against Google review?
4. Can combining the best NLP method with ML methods improve the sentiment evaluation of Google reviews?

3 Related Work

Sentiment analysis has recently gained popularity, moving from research to widespread industry use. The capability to convert large volumes of textual data into valuable insights about customer emotions and experiences is a must in world which is data focused. This section delves into the methodologies of sentiment analysis, exploring its different levels and approaches. Additionally, it examines the specific context of airport service quality measures, and the various methods used in sentiment analysis, from NLP approaches to sophisticated ML methods.

3.1 Airport Service Quality Measures

Airports are seen as the start point or destination end point in any journey and therefore also part of the overall travel and tourism experience. The airport experience significantly influences travelers' attitudes and behaviors toward a destination; a positive airport experience enhances the destination's image and increases tourism. Prentice et al. (2021)

While most studies have typically used normal research methods to examine airport service quality measures, only a handful of researchers have delved into internet content analysis. For example, some academics have performed sentiment analysis to evaluate passengers' views on airport service quality. Also content analysis has been used to identify factors influencing airport service quality and passenger satisfaction, evaluate airport service quality and analyse the airport experience comprehensively. Lee and Yu (2018) Nghiêm-Phú and Suter (2018)

The ACI is well established and experienced in the aviation sector so provides excellent guidance on the aviation industry. International (2024) Airports typically rely on two streams of revenue:- **Aeronautical** - This normally involves fees paid by the airlines and passengers for the use of airport facilities. This would typically consist of departing passenger charges, landing fees, and aircraft parking charges to name a few. **Non-Aeronautical** - This would usually fall into the commercially generated revenue from items like car parking, retail concessions, advertising, lounge facilities, and hotel operations.

Using the ACI airport customer experience method below this project aims to enhance customer understanding and provide a new measurement for customer sentiment using text feedback from passengers. The output of the research will feed into the airport strategy, service design, and innovation as well as guiding operational improvements.

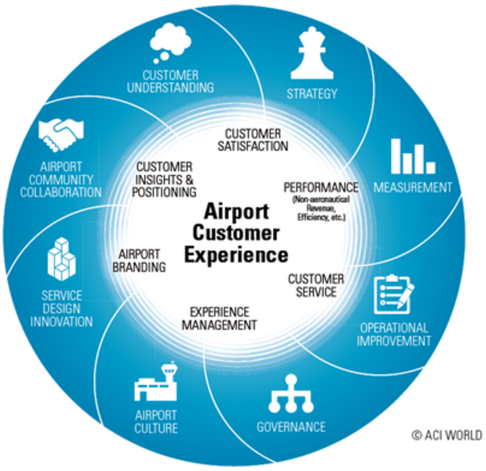


Figure 1: ACI Airport Customer Experience Method

Passengers' perceptions of their airport service experience are shaped by interactions with the airport environment and staff. For the departure journey, the first interaction starts with access which can be by car, either parking or drop-off, or by public transport like bus or taxi. The next steps are check-in, passport control, security, way-finding, food and beverage, airport facilities and finally boarding. For the arrival journey, it is a different process which starts with, disembarking the aircraft, baggage reclaim, customs and immigration, airport facilities, signage, availability of park, and public transport.

3.2 Sentiment Analysis Understanding

Sentiment analysis using NLP and online reviews has become very popular and is now used widely in a whole host of different sectors. Large amounts of text can be transformed to understand customer feelings and experiences of a specific product or service. Sentiment analysis can be examined at different levels as outlined below. These levels include document, sentence, phrase, and aspect.

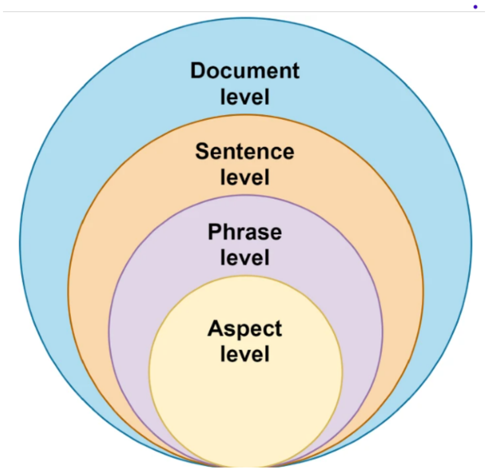


Figure 2: Level of Sentiment Analysis

3.2.1 Document Level: -

Document-level sentiment analysis evaluates the overall sentiment of a full text, assigning a single polarity score. It is often used for product reviews, articles, or emails, and involves methods like aggregating sentence-level scores, applying machine learning, and using NLP to assess tone.

3.2.2 Sentence Level: -

Sentence-level analysis reviews each sentence's polarity, necessary for texts with mixed sentiments. The main challenge is effectively modeling long texts to capture semantic relationships for accurate document-level sentiment classification. Rao et al. (2018) This is useful when the document-level sentiment is insufficient for specific uses. Behdenna et al. (2018) In the case of online Google Reviews, this approach works well as you can pick up several types of sentiments across a review with several paragraphs.

3.2.3 Phrase Level: -

Phrase-level sentiment analysis evaluates specific parts of a sentence to capture nuanced opinions. For example, in "The camera quality is poor, but the battery life is excellent," it identifies both positive and negative tones. This method is effective for multi-line reviews where a single aspect is expressed in a phrase. Thet et al. (2010) From a Google reviews standpoint this approach would be very useful for reviews with short feedback provided in a few sentences.

3.2.4 Aspect Level: -

The Aspect-Based Sentiment Analysis (ABSA) focal point is to identify and analyse sentiment features within a text, whether it's a single aspect or multiple aspects with their associated dependency relationships. These relationships can then be used to determine if the data is positive, negative, or neutral. Sann and Lai (2020) This will be very useful to understand specific features of a service or product, allowing for targeted improvements.

3.3 Natural Language Processing Approach

Positive and negative sentiments are determined in a NLP approach by using words from predefined dictionaries. These methods are popular for use in social media analysis and mining opinions while remaining computationally inexpensive and easy to understand. This methodology sums the values using polarity scores and can provide the sentiment output from as high a level as a whole document down to the aspect level. There is a downside to this approach, and they often fail to handle context, cynicism, and domain-specific language nuances. Regardless of the downsides to this approach, NLP methods are a good foundation and the first step in sentiment analysis, particularly when using a combination of advanced methods. Taboada et al. (2011)

Social media content presents challenges for sentiment analysis, but Vader, an easy rule-based model, proves highly useful. Vader outclasses eleven benchmark methods, including Linguistic Inquiry and Word Count (LIWC), Affective Norms for English Words (ANEW), and various machine learning algorithms (Naive Bayes and SVM), achieving an F1 classification accuracy of 0.96 compared to 0.84 for human raters. It is generally better across different contexts, making it one of the best and easiest approaches to use for online content. Hutto and Gilbert (2014)

TextBlob is a Python library for processing textual data, offering a straightforward API for various NLP tasks. These tasks include part-of-speech tagging, noun phrase extraction, sentiment analysis, text classification, translation, language detection, tokenisation, and word inflection and lemmatisation. By leveraging the strengths of underlying libraries like the Natural Language Toolkit (NLTK) and Pattern, Textblob simplifies complicated NLP operations, making it accessible for both developers and researchers to evaluate and control text data efficiently. Loria et al. (2018)

NLP and text analytics are widely used to understand a person's feelings about any given issue. This information is widely available in multiple social media and online locations across the web. NRCLex which is available in Python also and has a library that contains around 27,000 words based on the National Research Council Canada (NRC). Like Vader and Textblob, NRCLex is simple to use and understand but is limited to positive and negative sentiments. It also has a distinct advantage over the other two approaches with libraries which can measure the emotional impact of words and categorise them into eight primary emotions - anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. Mohammad and Turney (2013) These categorizations would come in useful for more in-depth analysis of any text or reviews.

3.4 Machine Learning Approach

Machine learning procedures involve training algorithms on datasets with text that has prelabeled sentiments. Commonly used methods include Support Vector Machines (SVM), Naive Bayes, and neural networks. This classification process employs machine learning principles to create a model using both training and testing data. Typically, 75% of the dataset is designated as training data, while the remaining 25% is used for testing the model. Dhini and Kusumaningrum (2018) There are many pros and cons to using a machine learning approach for sentiment analysis. The main advantages are:- **Accuracy and precision** - Advanced machine learning models, including neural networks and deep learning architectures, can attain remarkable accuracy and precision in sentiment analysis by detecting difficult patterns within data. Zhang et al. (2018)

Scalability - Large-scale datasets empower machine learning models to efficiently analyze extensive amounts of text data. This scalability is pivotal for applications such as social media monitoring and customer feedback analysis. Ghiassi et al. (2013)

Automation - After training, machine learning models can independently analyse fresh text data without human involvement, restructuring the process and reducing the requirement for human intervention. Cambria et al. (2016)

Adaptability - Machine learning models can enhance their flexibility and expand their application range by fine-tuning and adapting to certain domains or languages through retraining with relevant datasets. Medhat et al. (2014)

There are also some downsides to the machine-learning approach, here is some of them:- **Data Dependency** - The performance of machine learning models relies considerably on the quality and size of the training dataset. If the data is poor or biased, it can result in inaccurate or skewed predictions. Sun et al. (2017)

Resource Intensity and Complexity - Creating and training machine learning models, especially deep learning ones, demands sizable computational resources. These include powerful hardware and considerable time investment. Young et al. (2018)

Interpretability - Machine learning models, specifically deep learning models, often function as black boxes, concealing the logic behind their conclusions. This lack of interpretability can be difficult in applications where it is necessary to identify the decision-making process. Guidotti et al. (2018)

Need for Continuous Updating - Sentiment analysis and NLP models indeed need ongoing maintenance and adjustment. As language evolves, staying accurate and relevant requires regular updates and retraining. Mäntylä et al. (2018)

Sentiment analysis is a versatile and robust tool providing insights into customer emotions and feedback. This research explored different levels of sentiment analysis, from full review to aspect level, highlighting their distinctive methods and applications. Furthermore, the discussion extended to the significance of airport service quality, showcasing how sentiment analysis can be used to enhance passenger experiences and operational efficiencies. Different approaches to sentiment analysis were investigated, including NLP methods and machine-learning techniques, each with its benefits and limitations. The comprehensive overview provided here sets the stage for understanding the blend of these approaches, ultimately aiming to achieve a more nuanced and effective sentiment analysis framework.

4 Methodology

In this research, the use of CRISP-DM methodology is deployed as this framework provides a well-planned and structured approach to sentiment analysis. The below graphic show how the CRISP-DM steps are deployed to the project design.

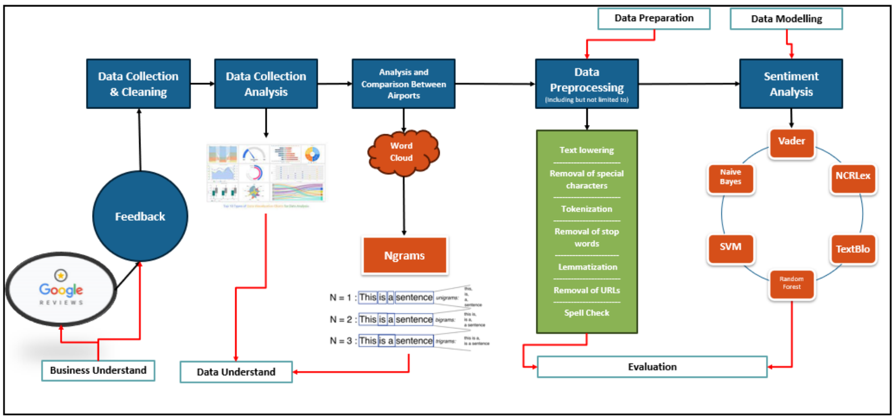

Data Preparation
Data Modelling
Data Collection & Cleaning
Data Collection Analysis
Analysis and Comparison Between Airports
Data Preprocessing (including but not limited to)
Sentiment Analysis
Feedback
Google REVIEWS
Word Cloud
Ngrams
N = 1 : This is a sentence unigrams: this, is, a, sentence
N = 2 : This is a sentence bigrams: this is, is a, a sentence
N = 3 : This is a sentence trigrams: this is a, is a sentence
Text lowering
Removal of special characters
Tokenization
Removal of stop words
Lemmatization
Removal of URLs
Spell Check
Vader
Naive Bayes
NCRLex
SVM
TextBlo
Random Forest
Business Understand
Data Understand
Evaluation

Figure 3: CRISP-DM Methodology

The structure provides key tasks that are completed in numerical order, with six major steps:

1. **Business Understanding:-** From a business perspective document the project objectives and requirements.
2. **Data Understanding:-** Collect and understand the data by describing, exploring, and checking for quality.
3. **Data Preparation:-** Before modeling any datasets, clean, prepare, integrate, and format all data.
4. **Modeling:-** Use and apply the appropriate modeling techniques by, selecting, building, and assessing methods.
5. **Evaluation:-** Evaluate the models ensuring the best results have been achieved. Review the process and determine the next steps.
6. **Deployment:-** Plan the deployment, supervise, and preserve the system. Evaluate the project and deliver the final report.

4.1 Data Input

In this project, Google reviews are used as the primary text source for sentiment analysis modeling. There are two main methods for extracting this information - Google API or third-party services. Due to the high cost and complexity of using the Google API, the third party service called Apify was selected. Apify is a web scraper that facilitates the extraction of data from Google reviews, offering customisation and automation features. Additionally, it is more affordable compared to many mainstream web scraping tools available in the market. All data extracted from Apify is provided in CSV file format.

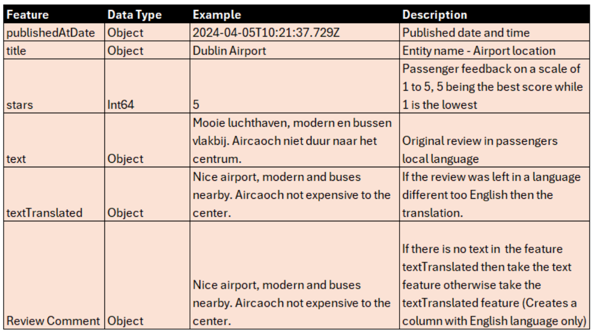
For this research, I have taken data for 6 airports within Europe with a data range of the 1st April 2023 to 31st March 2024. One year's data covers reviews across both summer and winter seasons. The airports covered in this project are as follows:-

- Dublin Airport - Based in Ireland and processed 33.5 million passengers in 2023. Airport (2024c)
- Athens International Airport - Based in Greece and processed 28.9 million passengers in 2023. Airport (2024a)
- Copenhagen Airport - Based in Denmark and processed 26.7 million passengers in 2023. Airport (2024b)
- Manchester Airport - Based in the northeast of England and processed 28.1 million passengers in 2023. MAG (2024)
- London Stansted Airport - Based outside England's capital city and processed 28.0 million passengers in 2023. MAG (2024)
- Zurich Airport - Based in Switzerland and processed 26.8 million passengers in 2023. Airport (2024d)

This list of airports offers a comprehensive view of European passenger opinions across various locations, setups, and ownerships, with similar annual passenger volumes.

4.2 Data Preprocessing and Transformation

The original data extracted from Google reviews contains a significant amount of data columns. For this project and to keep within the General Data Protection Regulation (GDPR) the data used was reduced to 6 columns of data outlined in the figure below.



Feature	Data Type	Example	Description
publishedAtDate	Object	2024-04-05T10:21:37.729Z	Published date and time
title	Object	Dublin Airport	Entity name - Airport location
stars	Int64	5	Passenger feedback on a scale of 1 to 5, 5 being the best score while 1 is the lowest
text	Object	Mooie luchthaven, modern en bussen vlakbij. Aircaoch niet duur naar het centrum.	Original review in passengers local language
textTranslated	Object	Nice airport, modern and buses nearby. Aircaoch not expensive to the center.	If the review was left in a language different too English then the translation.
Review Comment	Object	Nice airport, modern and buses nearby. Aircaoch not expensive to the center.	If there is no text in the feature textTranslated then take the text feature otherwise take the textTranslated feature (Creates a column with English language only)

Figure 4: Metadata list for Sentiment Analysis

From the initial review of the data, we can see 46100 rows of the data, containing the following features - published date, title, and stars review score all populated for all 46100 rows. When moving to the remaining features the number of rows populated is much lower. The text feature contains 21526 rows and the text translation feature contains 7369 rows of data which was translated from various languages into English.

Here are the steps to transforming the data before use:

1. A new feature called 'char count' which counts the number of characters from the review information in 'Review Comment'. This indicates the length of the review left by the passenger.
2. Using the 'publishedAtDate' feature to create several new pieces of data. These new features are:
   (a) 'Date' in a format of year - month - date (yyyy-mm-dd)

(b) Year number (yyyy)
(c) Month number (mm)
(d) Day of week name (Monday)
(e) Time by hour and minutes (hh:mm)
(f) Hour of the day in 24 hour clock (hh)

All of these additional features will give the ability to analyse the data in more depth.

4.3 Data Collection Analysis

4.3.1 Google Star Ratings Analysis

Using the features generated in the last section there is an opportunity to review the data in a lot of detail. An exploratory analysis in Python was conducted to gain deeper insights into the collected text.

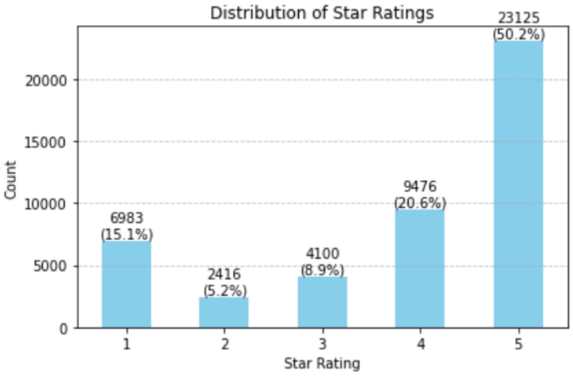


Figure 5: Overall Google Star Ratings for All Airports

Looking at the spread of the Google review star ratings for all airports, 70.6 percent of passengers gave a positive rating (4 and 5) followed by 20.3 percent giving a negative rating (1 and 2).

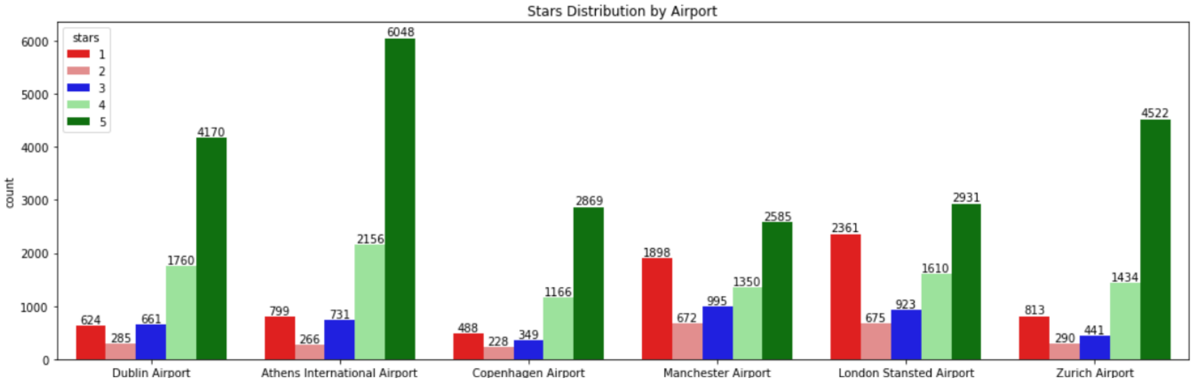


Figure 6: Google Star Rating Distribution by Airport

Star distribution by each airport for the year in review is outlined in the figure above. Green colours representing a positive score, blue representing a neutral score and finally the red colours representing a negative score.

The chart shows that Manchester and London Stansted airports have the most negative ratings, while Athens International Airport receives the most 5-star ratings, followed by Zurich and Dublin

Break down of the percentages of positive Google star ratings across the 6 airports:

1. Athens Airport = 82 percent
2. Dublin Airport = 79 percent *
3. Copenhagen Airport = 79 percent *
4. Zurich Airport = 79 percent *
5. Stansted Airport = 53 percent
6. Manchester Airport = 52 percent

* Dublin Airport had the lowest negative rate at 12 percent followed by Copenhagen at 14 percent and Zurich at 15 percent.

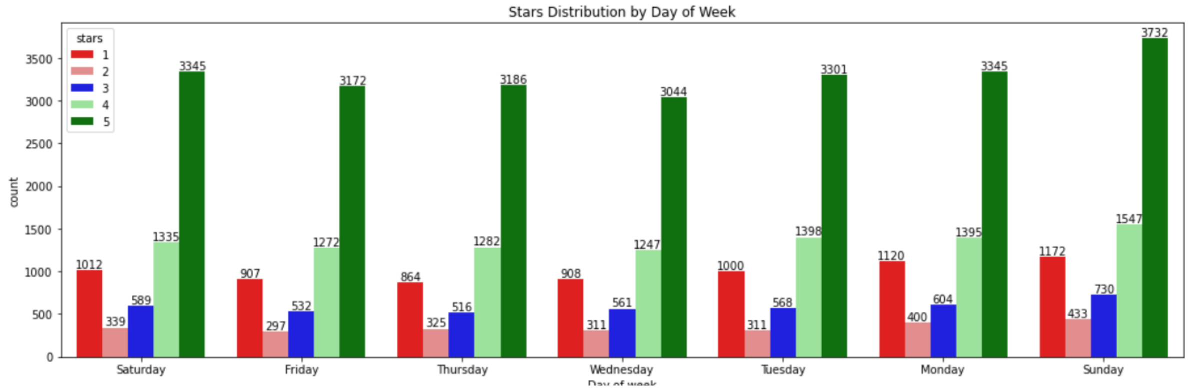


Figure 7: Google Star Rating Distribution for All Airports by Day of the Week

This chart shows Google star ratings distribution by day of the week, reflecting higher review feedback around weekends, with Sunday (7605), Monday (6864), and Saturday (6620) being the busiest.

Monday and Sunday have lower positive ratings (69%) compared to Thursday (72%). The decline suggests busier airports may have lower ratings due to increased pressure on facilities

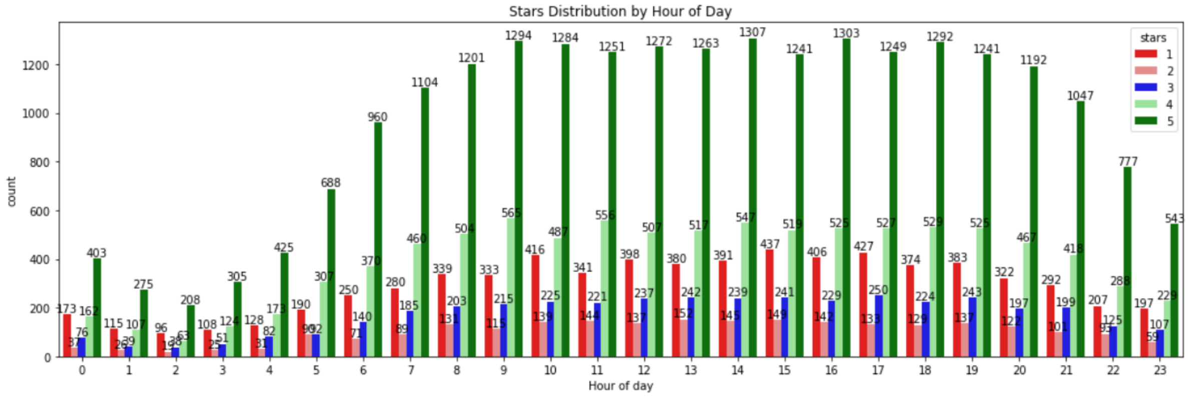


Figure 8: Google Star Ratings Distribution by Hour of the Day

Figure 8 shows Google Star ratings across 24 hours. Busy European airports operate from 03:00 to 00:00, with delays often extending arrivals into the early morning.

61% of Google reviews are posted between 09:00 and 19:00, with the peak at 14:00, aligning with the busiest times for passenger arrivals and departures. The hours early in the morning (05:00 to 08:00) are mainly reserved for departure-only flights, while the late hours (22:00 to 00:00) see the most arrival flights. Positive ratings are higher between 04:00 and 10:00 (73%), but drop to 67% from 23:00 to 03:00.

4.3.2 Text Analysis

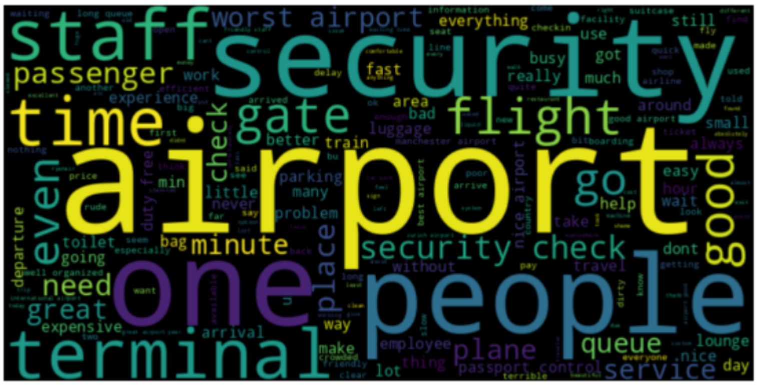


Figure 9: Word Cloud of the most frequent words

The word cloud output from Figure 9 shows the most frequent words used by passengers . The top 5 words are 'airport' (13,884 mentions), 'security' (4,189), 'staff' (3,557), 'flight' (3,125), and 'time' (3,124). 'Airport' is mentioned three times more than 'security.'

Ngrams - Top 5 Words

The top 5 words don't really give us too much information so the use of Ngrams is being used to enhance the details. All the charts for the Ngrams can be viewed in the Appendix section of the configuration manual.

Airport - The data reveals mixed sentiments about airports, with frequent negative phrases like "worst airport" indicating widespread dissatisfaction. However, positive mentions such as "nice airport" suggest some airports are well-regarded, showing significant variation in passenger experiences across the six airports.

Security - The data shows "security check" as the most frequent negative bigram, along with others like "security staff" and "security control." Trigrams such as "long queue security" and "waiting time security" highlight frustrations with lengthy security processes, indicating significant dissatisfaction with security efficiency and wait times.

Staff - Sentiment towards airport staff is generally positive, with "friendly staff" and "helpful staff" frequently mentioned. However, there are also mentions of "rude staff," indicating occasional negative experiences. Overall, most interactions are positive, though some issues impact passenger satisfaction.

Flight - The data reveals significant passenger concerns about missing flights, with frequent phrases like "missed flight" and "miss flight". Issues with "connecting flight" and "flight delayed" also emerge, indicating that delays and missed connections are of concern. Overall, flight-related disruptions are a prevalent source of passenger frustration.

Time - The analysis shows "waiting time" as the top concern for travelers, with frequent mentions like "long waiting time" and "waiting time security." These issues, especially at security checks, are recurring problems, indicating widespread dissatisfaction with airport delays and inefficiencies.

Ngrams - By Airport

Analysing the top 10 bigrams and trigrams per airport provides insights into positive and negative issues. Here is a summary for each location:

Athens - Key features include efficient passport control, excellent duty-free shops, and strong organisation. Phrases like "nice airport," "well organized," and "best airport" reflect its reputation for quality and ease of navigation, earning it recognition as one of Europe's best.

Dublin - Key features include "security check," "friendly staff," "duty free," and "well organized airport." Phrases like "nice airport" and "easy navigate" indicate a positive experience, with praise for its organization, helpful staff, and clean facilities.

Copenhagen - Key features include "security check," "nice airport," and "passport control." Phrases like "good airport" and "easy find way" reflect a positive experience, with praise for efficiency, cleanliness, and easy navigation.

Manchester - The data shows significant dissatisfaction, with frequent complaints like "worst airport ever," "long queue," and "security check." Passengers consistently describe the airport as one of the worst, citing poor security and customer service.

Stansted - The data reveals widespread dissatisfaction, with frequent complaints about "security check," "worst airport," and "long queue." Travelers often describe it as "worst airport ever," citing long wait times, poor security, car parking, and overall inefficiency.

Zurich - review highlights Zurich Airport's strengths in cleanliness and organisation, with phrases like "one best airport" and "clean well organized." However, concerns about "long waiting time" and "waiting time security" suggest issues with delays. Overall, the airport is praised for its high standards but faces challenges with wait times.

In summary, the analysis shows mixed sentiment towards airports, with significant dissatisfaction over security checks and wait times. Positive staff interactions are common, but missed flights and delays are frequent frustrations. Athens, Dublin, Copenhagen, and Zurich are praised for efficiency, while Manchester and Stansted face criticism for long waits and poor service. Overall, improvements in security and time management are needed to boost passenger satisfaction.

4.4 Sentiment Analysis Methods

Google reviews from six airports offer enough data for detailed sentiment analysis. Using a blend of NLP and ML techniques to achieve the best evaluation results and apply the top-performing model.

TextBlob is processes textual data with simple APIs for NLP tasks like sentiment analysis, part-of-speech tagging, and noun phrase extraction, often using a lexicon-based approach.

Vader is a sentiment analysis tool for social media text that uses a lexicon and rule-based approach to assess sentiment polarity.

NRCLex is a sentiment analysis tool that uses the NRC Emotion Lexicon to categorize text by eight emotions and two sentiments, providing detailed emotional analysis.

SVM is a supervised machine learning algorithm for sentiment analysis that classifies text by finding the optimal hyperplane to separate different sentiment classes in a high-dimensional space.

Naïve Bayes is a probabilistic classifier based on Bayes' Theorem that predicts sentiment by calculating the likelihood of each sentiment class based on specific words in the text.

4.5 Evaluation

Evaluation is key to selecting the best model and ensuring research accuracy. This project uses the following evaluation techniques:

1. **Accuracy Score**:- Measures the percentage of correct predictions out of the total, useful for balanced class distributions.

$$\text{Accuracy} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i+FP_i+FN_i)}$$

Figure 10: Formula for Accuracy Score Multi-Class Classification

2. (a) TPi (True Positives for class i) are the instances correctly predicted for class i.
   (b) FPi (False Positives for class i) are the instances incorrectly predicted as class i.
   (c) FNi (False Negatives for class i) are the instances of class iii incorrectly predicted as another class.
   (d) n is the number of classes (in this case, n=3)
3. **Precision Score**:- Evaluates the ratio of correctly predicted positives to total predicted positives, useful for minimizing false positives.

For a given class i:

$$\text{Precision}_i = \frac{TP_i}{TP_i+FP_i}$$

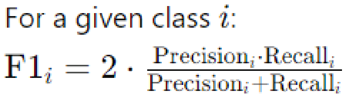
Figure 11: Formula for Precison Score Multi-Class Classification

4. **Recall Score**:- Measures the ratio of correctly predicted positives to total actual positives, important for minimizing false negatives.

For a given class i:

$$\text{Recall}_i = \frac{TP_i}{TP_i+FN_i}$$

Figure 12: Formula for Recall Score Multi-Class Classification

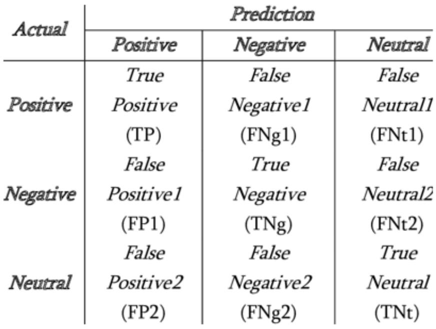


For a given class i:

$$\mathbf{F1}_i = 2 \cdot \frac{\text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

Figure 13: Formula for F1 Score Multi-Class Classification

5. **F1 Score**:- Balances precision and recall, accounting for both false positives and false negatives.
6. **Confusion Matrix**:- Compares model predictions to actual labels, detailing true positives, true negatives, false positives, and false negatives.



Actual	*Prediction*		
	Positive	*Negative*	*Neutral*
Positive	*True Positive* (TP)	*False Negative1* (FNg1)	*False Neutral1* (FNt1)
Negative	*False Positive1* (FP1)	*True Negative* (TNg)	*False Neutral2* (FNt2)
Neutral	*False Positive2* (FP2)	*False Negative2* (FNg2)	*True Neutral* (TNt)

Figure 14: Confusion Matrix for the Sentiment Classification Model

5 Design Specification

Here is the architectural design for the project on sentiment analysis:

1. The data is retrieved from Google Reviews via a third-party service provider.
2. Data is reduced to a few required columns with additional features added to enhance the date and time information for data collection analysis
3. Exploratory Data Analysis (EDA) is performed on the data to gain a better understanding and uncover deeper insights and patterns.
4. To generate sentiment analysis and create prediction models from text reviews the following approach was adopted:
   (a) **Textblob** - Compute sentiment polarity and subjectivity.
   (b) **Vader** - Analyze sentiment polarity scores (positive, negative, neutral).
   (c) **NRCLex** - Assign emotions to text (joy, anger, sadness, etc.).
   (d) **Random Forest** - Ensemble learning method using multiple decision trees.
   (e) **SVM** - Classification technique that finds the hyperplane separating different classes.
   (f) **Naive Bayes** - Probabilistic classifier based on Bayes' theorem.

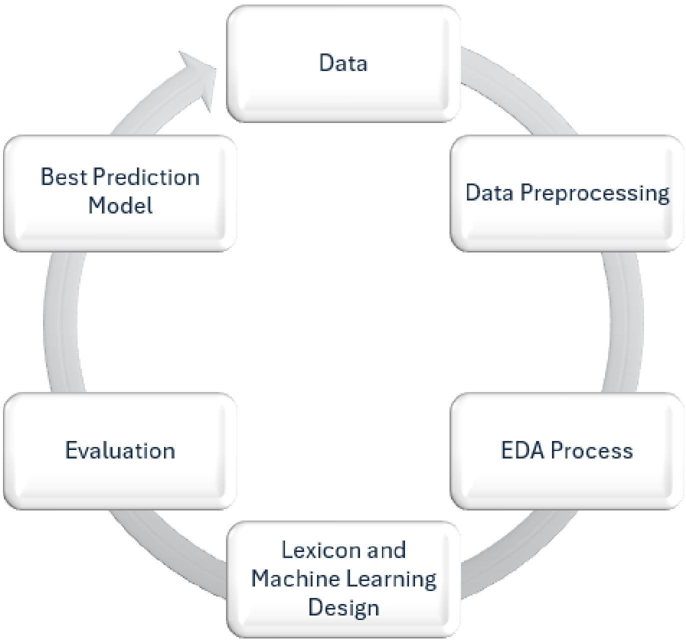


Figure 15: Project Design

6 Implementation

6.1 Data Collection

The first step involves retrieving Google review data from a third-party provider, downloading it in CSV format. Automation and API functionality are used for web scraping, with only the English text review, date/time stamp, star rating, and location included in the models. All other data is discarded.

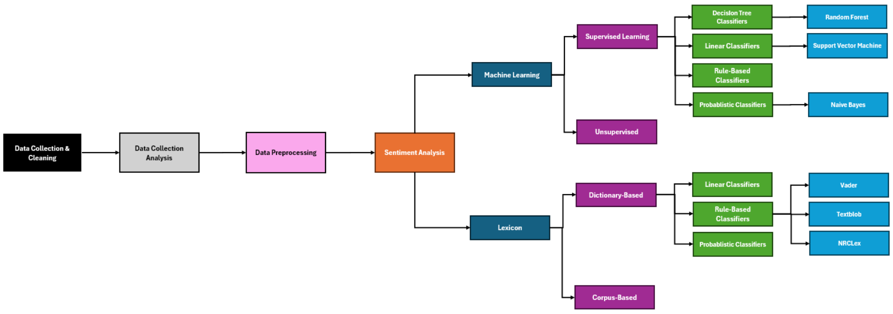


Figure 16: Sentiment Analysis Design Diagram

6.2 Data Analysis

Using Seaborn, Matplotlib, and WordCloud in Python, we performed initial data analysis and aspect-level review. N-grams were used to gain deeper insights, revealing patterns like ”worst airport” and ”nice airport” in bigrams, and ”worst airport ever” and ”one best airport” in trigrams. This approach provided better context for identifying language patterns.

6.3 Data Cleaning and Preprocessing

This step is important, cleaning and preparing the text data before applying any lexicon or machine learning techniques. These steps included:-

- Convert the text to lowercase
- Remove any URLs
- Remove any emails
- Remove punctuation
- Remove numbers
- Tokenise the text - Breaking down text into individual words or phrases.
- Remove stop words - Removing common words that do not contribute to sentiment.
- Lemmatisation - Reducing words to their base or root form.
- Rejoin tokens into strings
- Remove extra white spaces
- Remove leading and trailing white spaces

6.4 Sentiment Analysis

The sentiment analysis employed two approaches: **NLP** with dictionary-based methods (VADER, NRCLex, TextBlob) and **ML** with classifiers (Naive Bayes, Random Forest, SVM).

6.4.1 Natural Language Processing

VADER, NRCLex, and TextBlob were integrated using Python's NLTK library. VADER and TextBlob used pre-built sentiment analyzers for scores, while NRCLex offered detailed emotion-based analysis. Each text received a compound sentiment score, classified into positive, negative, or neutral categories based on predefined thresholds.

6.4.2 Machine Learning

The second phase focused on developing ML models for sentiment classification using supervised learning with labeled Google review star ratings. The dataset was split into 80% training and 20% testing for SVM, Random Forest, and Naive Bayes.

.

SVM - The text data is vectorized using `CountVectorizer`, and the dataset is split into training and testing sets. A SVM classifier with a linear kernel is then trained on the training data to predict sentiment labels.

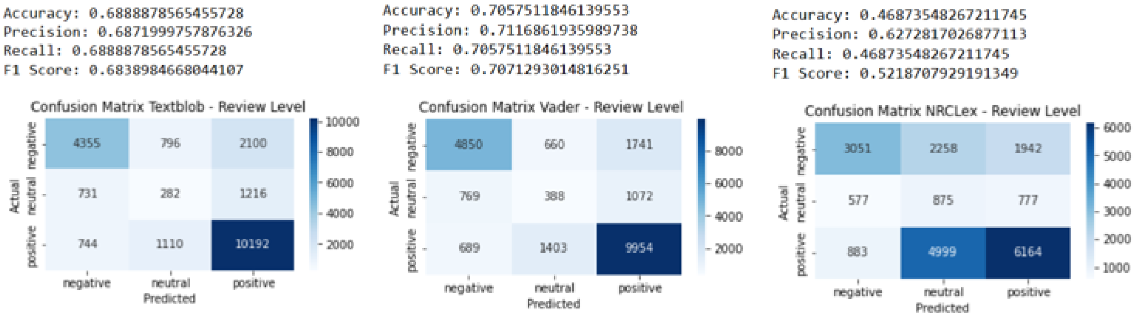
Random Forest - The text data is converted into numerical features using Count Vectorizer, and the dataset is split into training and testing sets. A Random Forest classifier with multiple trees is then trained on the training data to classify the sentiment labels.

Naive Bayes - The text data is transformed into numerical features using Count Vectorizer, and the dataset is divided into training and testing sets. A Naive Bayes classifier (`MultinomialNB`) is then trained on the training data to predict sentiment labels.

7 Evaluation and Results

7.1 Which NLP Methods Returns the Best Evaluate Sentiment from Google Reviews?

This experiment compares Google star ratings (1-5) with polarity scores from TextBlob, VADER, and NRCLex. Ratings of 1 and 2 stars, along with negative polarity scores, indicate negativity; ratings of 4 and 5 stars, along with positive polarity scores, indicate positivity; and a 3-star rating with a zero polarity score represents neutrality. .



```
Accuracy: 0.6888878565455728
Precision: 0.6871999757876326
Recall: 0.6888878565455728
F1 Score: 0.6838984668044107
```

Confusion Matrix Textblob - Review Level

Actual \ Predicted	negative	neutral	positive
negative	4355	796	2100
neutral	731	282	1216
positive	744	1110	10192

```
Accuracy: 0.7057511846139553
Precision: 0.7116861935989738
Recall: 0.7057511846139553
F1 Score: 0.7071293014816251
```

Confusion Matrix Vader - Review Level

Actual \ Predicted	negative	neutral	positive
negative	4850	660	1741
neutral	769	388	1072
positive	689	1403	9954

```
Accuracy: 0.46873548267211745
Precision: 0.6272817026877113
Recall: 0.46873548267211745
F1 Score: 0.5218707929191349
```

Confusion Matrix NRCLex - Review Level

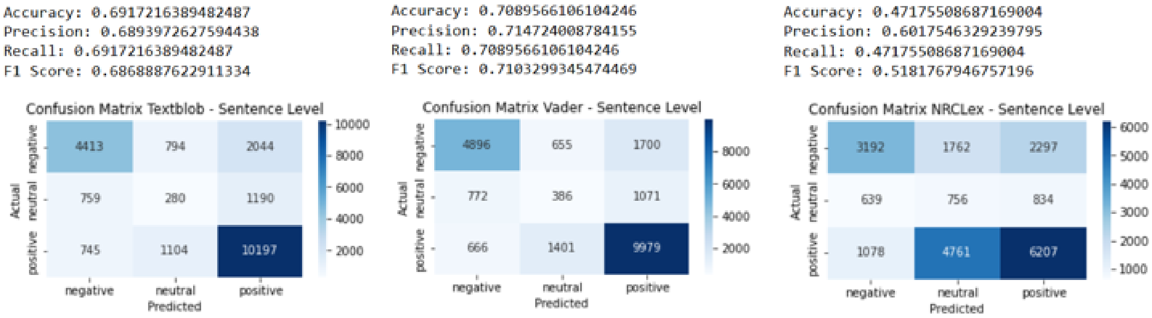
Actual \ Predicted	negative	neutral	positive
negative	3051	2258	1942
neutral	577	875	777
positive	883	4999	6164

Figure 17: Evaluation Scores for Total Review

VADER outperforms TextBlob and NRCLex in accuracy, precision, recall, and F1 score. TextBlob is slightly less effective than VADER, while NRCLex performs the worst across all metrics.

7.2 How do NLP Methods Perform in Sentiment Analysis at the Sentence Level of Google Reviews?

Using the same parameters from the full review, here is the results from the sentence level.



```
Accuracy: 0.6917216389482487
Precision: 0.6893972627594438
Recall: 0.6917216389482487
F1 Score: 0.6868887622911334
```

Confusion Matrix Textblob - Sentence Level

Actual \ Predicted	negative	neutral	positive
negative	4413	794	2044
neutral	759	280	1190
positive	745	1104	10197

```
Accuracy: 0.7089566106104246
Precision: 0.714724008784155
Recall: 0.7089566106104246
F1 Score: 0.7103299345474469
```

Confusion Matrix Vader - Sentence Level

Actual \ Predicted	negative	neutral	positive
negative	4896	655	1700
neutral	772	386	1071
positive	666	1401	9979

```
Accuracy: 0.47175508687169004
Precision: 0.6017546329239795
Recall: 0.47175508687169004
F1 Score: 0.5181767946757196
```

Confusion Matrix NRCLex - Sentence Level

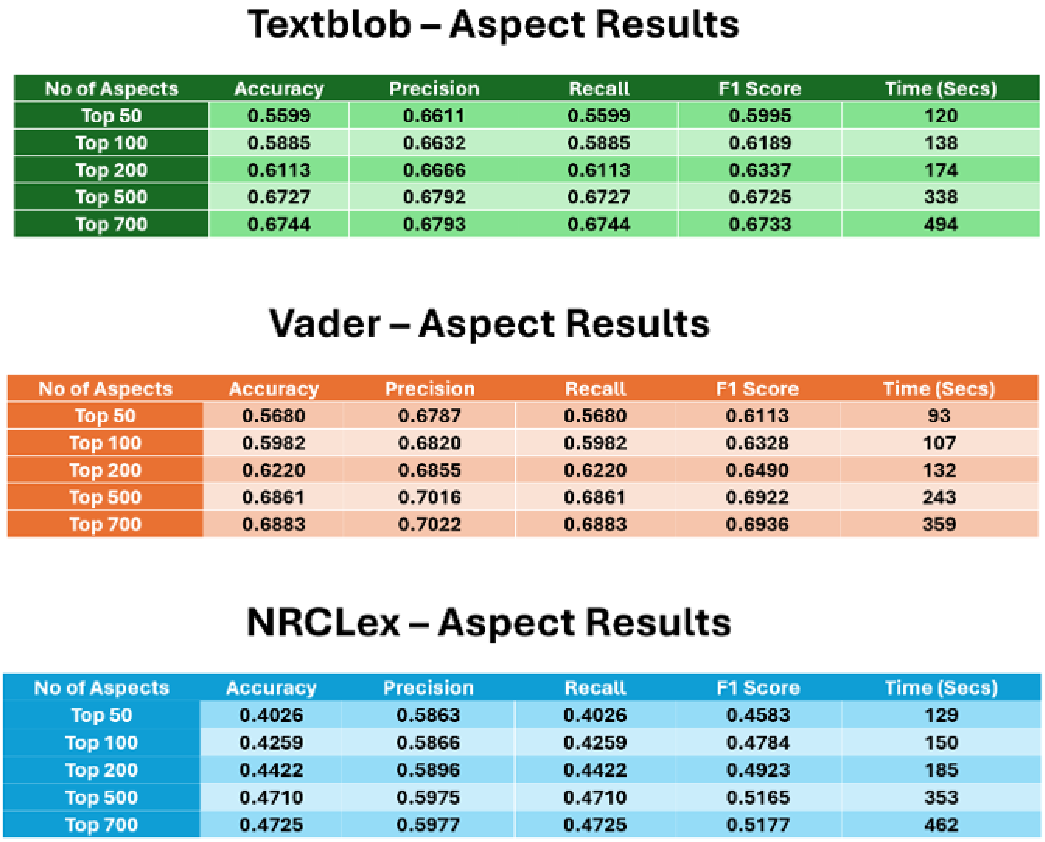
Actual \ Predicted	negative	neutral	positive
negative	3192	1762	2297
neutral	639	756	834
positive	1078	4761	6207

Figure 18: Evaluation Scores for Sentence Review

VADER outperforms TextBlob and NRCLex in all metrics, with TextBlob slightly behind VADER. NRCLex has the lowest performance overall. VADER performs slightly better at the sentence level than with full reviews.

7.3 How do NLP Methods Perform in Sentiment Analysis at the Aspect Level of Google Reviews?

Using the same parameters, aspect-level results are shown for the top 50, 100, 200, 500, and 700 aspects. Results beyond 500 aspects show minimal gains but significantly increased computation times.



Textblob – Aspect Results

No of Aspects	Accuracy	Precision	Recall	F1 Score	Time (Secs)
Top 50	0.5599	0.6611	0.5599	0.5995	120
Top 100	0.5885	0.6632	0.5885	0.6189	138
Top 200	0.6113	0.6666	0.6113	0.6337	174
Top 500	0.6727	0.6792	0.6727	0.6725	338
Top 700	0.6744	0.6793	0.6744	0.6733	494

Vader – Aspect Results

No of Aspects	Accuracy	Precision	Recall	F1 Score	Time (Secs)
Top 50	0.5680	0.6787	0.5680	0.6113	93
Top 100	0.5982	0.6820	0.5982	0.6328	107
Top 220	0.6220	0.6855	0.6220	0.6490	132
Top 500	0.6861	0.7016	0.6861	0.6922	243
Top 700	0.6883	0.7022	0.6883	0.6936	359

NRCLex – Aspect Results

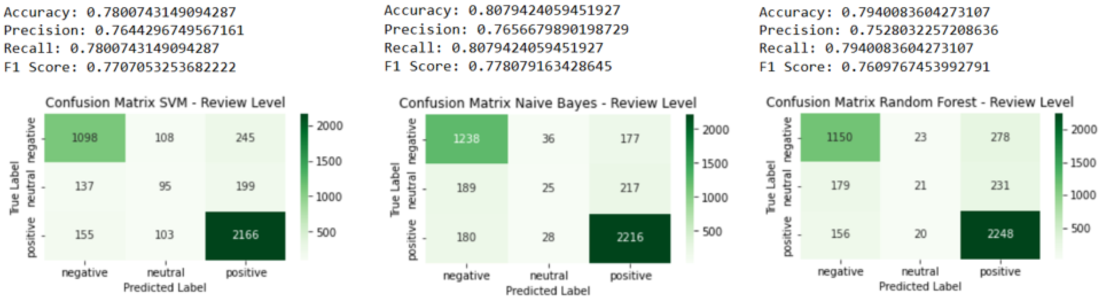
No of Aspects	Accuracy	Precision	Recall	F1 Score	Time (Secs)
Top 50	0.4026	0.5863	0.4026	0.4583	129
Top 100	0.4259	0.5866	0.4259	0.4784	150
Top 200	0.4422	0.5896	0.4422	0.4923	185
Top 500	0.4710	0.5975	0.4710	0.5165	353
Top 700	0.4725	0.5977	0.4725	0.5177	462

Figure 19: Evaluation Scores for Aspect Review

The sentiment analysis results show varying performance metrics across the three tools. TextBlob improves with more aspects, reaching an accuracy of 0.6744 and an F1 score of 0.6733 for 700 aspects, taking 494 seconds. VADER slightly outperforms TextBlob, achieving 0.6883 accuracy and 0.6936 F1 score in 359 seconds. NRCLex has the lowest metrics, with 0.4725 accuracy and 0.5177 F1 score, taking 462 seconds. Overall, VADER and TextBlob offer superior performance compared to NRCLex, but none surpass the sentence-level results, with VADER remaining the best model.

7.4 How do Different ML Models Perform in Predicting Review Ratings from Google Reviews?

Switching to machine learning techniques and using SVM, Naive Bayes, and Random Forest to see if there is an improvement in evaluation scores from previous experiments.



```
Accuracy: 0.7800743149094287
Precision: 0.7644296749567161
Recall: 0.7800743149094287
F1 Score: 0.7707053253682222
```

```
Accuracy: 0.8079424059451927
Precision: 0.7656679890198729
Recall: 0.8079424059451927
F1 Score: 0.778079163428645
```

```
Accuracy: 0.7940083604273107
Precision: 0.7528032257208636
Recall: 0.7940083604273107
F1 Score: 0.7609767453992791
```

Figure 20: Evaluation Scores for Machine Learning Methods

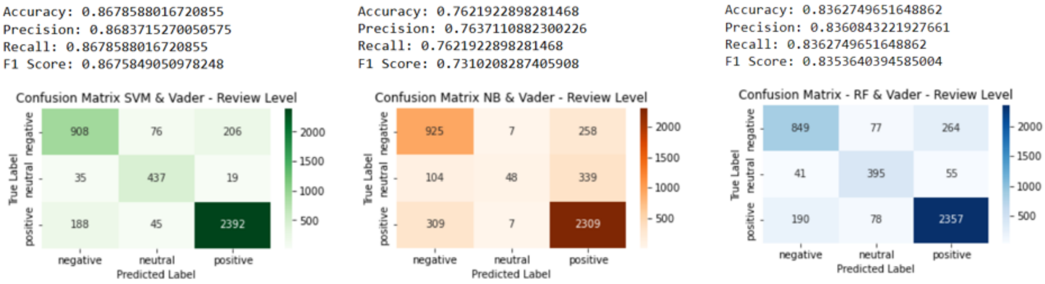
Among the classifiers, Naive Bayes has the highest accuracy (0.8079), precision (0.7657), recall (0.8079), and F1 score (0.7781). Random Forest slightly outperforms SVM in accuracy (0.7941 vs. 0.7801) and recall (0.7941 vs. 0.7801), while SVM has a higher F1 score (0.7707 vs. 0.7609). The confusion matrices show all models surpass previous evaluations in predicting positives, with Naive Bayes best at reducing false positives and negatives. The neutral category is the most challenging, with fewer correct predictions compared to negative and positive categories.

7.5 Can Combining the Best NLP Method with ML Methods Improve the Sentiment Evaluation of Google Reviews?

Vader produced the best results from the NLP method so in this experiment we will blend it with, SVM, Naive Bayes, and Random Forest to see if the evaluation results will increase.

The performance results of the three sentiment analysis models are as follows:

- **SVM & Vader**: Highest accuracy (86.76%), precision (86.84%), recall (86.76%),



```
Accuracy: 0.8678588016720855
Precision: 0.8683715270050575
Recall: 0.8678588016720855
F1 Score: 0.8675849050978248
```

```
Accuracy: 0.7621922898281468
Precision: 0.7637110882300226
Recall: 0.7621922898281468
F1 Score: 0.7310208287405908
```

```
Accuracy: 0.8362749651648862
Precision: 0.8360843221927661
Recall: 0.8362749651648862
F1 Score: 0.8353640394585004
```

Figure 21: Evaluation Scores for Blended Approach

and F1 score (86.76%), with strong positive review classification and minimal misclassifications.

- **RF & Vader**: Next best, with accuracy (83.63%), precision (83.61%), recall (83.63%), and F1 score (83.54%), showing effective classification but with more misclassifications than SVM & VADER.
- **NB & VADER**: Lowest metrics, with accuracy (76.22%), precision (76.37%), recall (76.22%), and F1 score (73.10%), indicating more misclassifications, especially in neutral reviews.

Overall, SVM & Vader is the most effective, followed by RF & Vader, while NB & Vader under performed.

7.6 Discussion

7.6.1 Natural Language Processing

Evaluating the Google reviews at different levels (Whole, Sentence, and Aspect) using NLP methods (Vader, NRCLex, and Texblob) against the Google Reviews star rating (5 and 4 = Positive, 3 = Neutral, and 2 and 1 = Negative), Vader outperformed TextBlob and NRCLex across all evaluation metrics. TextBlob performed moderately well but was slightly less effective than VADER. NRCLex showed the lowest performance across all metrics, suggesting it is the least reliable model for this task. VADER at the sentence level achieved the highest evaluation scores, but the computational demands and speed issues at the aspect level hindered the ability to determine if this approach could perform better.

Suggested Improvements - Larger, more diverse datasets and models like BERT or GPT with contextual embeddings could improve scores. At the aspect level, using dependency parsing, neural networks, and cloud-based solutions for better speed and computation would enhance practicality.

7.6.2 Machine Learning

Using machine learning techniques at the complete review level, Naive Bayes performed best in accuracy, precision, recall, and F1 score, followed by Random Forest and SVM. All models excelled in classifying positive reviews, but struggled with neutral ones. Machine learning models outperformed traditional NLP algorithms.

Suggested Improvements - Using advanced feature engineering like TF-IDF, word embeddings, or domain-specific features could enhance model performance. Applying SMOTE or adjusting class weights may also address class imbalance.

7.6.3 NLP and ML Combined

The blended approach combining Vader with machine learning showed SVM & Vader as the top performer, followed by Random Forest & Vader, with Naive Bayes & Vader being the least effective. This hybrid method achieved the best results, though increased complexity may impact on computational efficiency.

Suggested Improvements - Improve the blending process to reduce complexity and improve computational efficiency without impacting performance.

Overall, the experiments demonstrate that while traditional NLP models like Vader, TextBlob, and NRCLex provide a solid baseline for sentiment analysis, integrating these with machine learning methods significantly enhances performance.

The blended approach, particularly with Vader and SVM, yielded the highest accuracy, precision, recall, and F1 scores, showcased the potential of hybrid models. These findings underscore the value of combining natural language processing systems with machine learning to effectively capture the nuances in sentiment analysis tasks.

8 Conclusion and Future Work

The purpose of this research was to develop and evaluate classification models that categorise Google review data for airports by using both supervised machine learning techniques and traditional natural language processing methods. The findings uncover the strengths and limitations of each method in analysing everyday review data, offering insights into the most effective approach for sentiment analysis.

The results have shown that a combination of machine learning and NLP technique delivers the best results. The accuracy, precision, recall and F1 scores in the low to mid-eighties were achieved when blending SVM and Random Forest machine learning methods with Vader.

The NLP only approach is an easier deployment technique however the best evaluation scores achieved was by Vader in the early seventies, which indicates that the model is reasonably effective. The machine learning only methodology improved on the NLP approach and increased the evaluation scores into the late seventies and early eighties with Naïve Bayes producing the best results.

The project also examined the impact of reviews at both the sentence and aspect levels. Using the NLP only approach, we observed a modest improvement in evaluation scores at the sentence level with minimal impact on code execution time and processing speed. However, at the aspect level, there was a significant increase in code execution time and processing speed. As the number of aspects increased, processing times increased significantly.

The methods deployed in this project can be used with other textual data available within the airport environment, bringing multiple data together in a coherent way. This approach can also be enhanced to focus on aspect-based sentiment analysis, where specific parts of a product or service (e.g. Security, Way-finding, Cleanliness) are evaluated separately, providing more detailed insights into customer feedback. The aspect approach can also be used to benchmark products and services across different airports, providing insight that is currently not available within the aviation industry.

References

Airport, A. I. (2024a). Facts and figures, https://www.aia.gr/company-and-business/the-company/facts-and-figures. Accessed: 15th July 2024.

Airport, C. (2024b). Strong growth at copenhagen airport: Close to 27 million travellers, https://www.cph.dk/en/about-cph/investor/traffic-statistics/2024/01/strong%20growth%20at%20copenhagen%20airport%20close%20to%2027%20million%20travellers. Accessed: 15th July 2024.

Airport, D. (2024c). Almost 32 million through dublin airport's terminals in 2023, https://www.dublinairport.com/latest-news/2024/01/24/

almost-32-million-through-dublin-airport-s-terminals-in-2023. Accessed: 15th July 2024.

Airport, Z. (2024d). Passenger numbers closer to precovid levels, https://newsroom.flughafen-zuerich.ch/en/zurich-airport-passenger-numbers-closer-to-pre-covid-levels/. Accessed: 15th July 2024.

Behdenna, S., Barigou, F. and Belalem, G. (2018). Document level sentiment analysis: a survey, *EAI endorsed transactions on context-aware systems and applications* **4**(13): e2–e2.

Cambria, E., Schuller, B., Xia, Y. and White, B. (2016). New avenues in knowledge bases for natural language processing, *Knowledge-Based Systems* **108**(C): 1–4.

Dhini, A. and Kusumaningrum, D. (2018). Sentiment analysis of airport customer reviews, *2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pp. 502–506.

Europe, A. (2018). The passenger at the heart of the airport business, https://aci-europe.org/. Accessed: 09th June 2024.

Ghiassi, M., Skinner, J. and Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network, *Expert Systems with applications* **40**(16): 6266–6282.

Google (2024a). Add, edit, or delete google maps reviews ratings, https://support.google.com/maps/answer/6230175?hl=en&co=GENIE.Platform%3DAndroid/. Accessed: 12th June 2024.

Google (2024b). Prohibited restricted content, https://support.google.com/contributionpolicy/answer/7400114?sjid=10159888675302090209-EU/. Accessed: 12th June 2024.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. and Pedreschi, D. (2018). A survey of methods for explaining black box models, *ACM computing surveys (CSUR)* **51**(5): 1–42.

HappyOrNot (2024). We help turn feedback into revenue, https://www.happy-or-not.com/en/?_gl=1*gu127e*_up*MQ..&gclid=Cj0KCQjwvb-zBhCmARIsAAfUI2vIrsRenryrig7NPPsjBMki3z5_b5LEqhTA6670yNchEAbgIXWPWPYaAujYEALw_wcB. Accessed: 17th June 2024.

Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text, *Proceedings of the international AAAI conference on web and social media*, Vol. 8, pp. 216–225.

IAA (2024). Service quality and passenger focus, https://www.iaa.ie/docs/default-source/car-documents/1c-economic-regulation/dublin-airport-appendix-4-service-quality-report.pdf?sfvrsn=a38310f3_1. Accessed: 17th June 2024.

International, A. C. (2024). About us, https://www.aci-europe.org/about/about-us.html. Accessed: 01st July 2024.

Lee, K. and Yu, C. (2018). Assessment of airport service quality: A complementary approach to measure perceived service quality based on google reviews, *Journal of Air Transport Management* **71**: 28–44.

Loria, S. et al. (2018). textblob documentation, *Release 0.15* **2**(8): 269.

MAG (2024). Strong december performance for uk's largest airports group sees mag serve 60m passengers in 2023, `https://mediacentre.magairports.com/strong-december-performance-for-uks-largest-airports-group-sees-mag-serve-60m-passe` Accessed: 15th July 2024.

Mäntylä, M. V., Graziotin, D. and Kuutila, M. (2018). The evolution of sentiment analysis—a review of research topics, venues, and top cited papers, *Computer Science Review* **27**: 16–32.

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

Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon, *Computational intelligence* **29**(3): 436–465.

Nghiêm-Phú, B. and Suter, J. R. (2018). Airport image: an exploratory study of mccarran international airport, *Journal of Air Transport Management* **67**: 72–84.

Prentice, C., Wang, X. and Manhas, P. S. (2021). The spillover effect of airport service experience on destination revisit intention, *Journal of Hospitality and Tourism Management* **48**: 119–127.

Rao, G., Huang, W., Feng, Z. and Cong, Q. (2018). Lstm with sentence representations for document-level sentiment classification, *Neurocomputing* **308**: 49–57.

Sann, R. and Lai, P.-C. (2020). Understanding homophily of service failure within the hotel guest cycle: Applying nlp-aspect-based sentiment analysis to the hospitality industry, *International Journal of Hospitality Management* **91**: 102678.

Sun, S., Luo, C. and Chen, J. (2017). A review of natural language processing techniques for opinion mining systems, *Information fusion* **36**: 10–25.

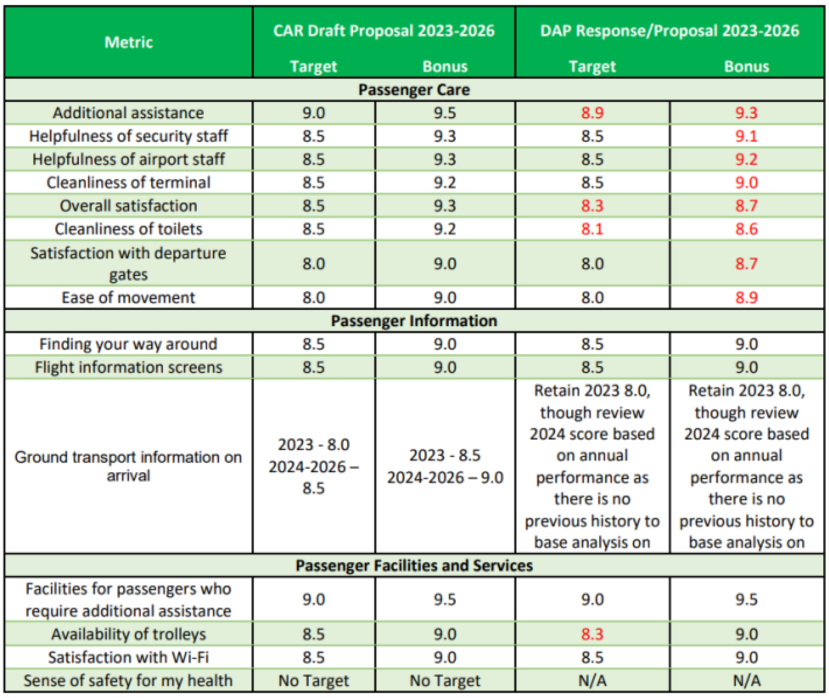
Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M. (2011). Lexicon-based methods for sentiment analysis, *Computational Linguistics* **37**(2): 267–307. **URL:** *https://aclanthology.org/J11-2001*

Thet, T. T., Na, J.-C. and Khoo, C. S. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards, *Journal of information science* **36**(6): 823–848.

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

Zhang, L., Wang, S. and Liu, B. (2018). Deep learning for sentiment analysis: A survey, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **8**(4): e1253.

9 Appendix



Metric	CAR Draft Proposal 2023-2026		DAP Response/Proposal 2023-2026	
	Target	Bonus	Target	Bonus
Passenger Care				
Additional assistance	9.0	9.5	8.9	9.3
Helpfulness of security staff	8.5	9.3	8.5	9.1
Helpfulness of airport staff	8.5	9.3	8.5	9.2
Cleanliness of terminal	8.5	9.2	8.5	9.0
Overall satisfaction	8.5	9.3	8.3	8.7
Cleanliness of toilets	8.5	9.2	8.1	8.6
Satisfaction with departure gates	8.0	9.0	8.0	8.7
Ease of movement	8.0	9.0	8.0	8.9
Passenger Information				
Finding your way around	8.5	9.0	8.5	9.0
Flight information screens	8.5	9.0	8.5	9.0
Ground transport information on arrival	2023 - 8.0 2024-2026 – 8.5	2023 - 8.5 2024-2026 – 9.0	Retain 2023 8.0, though review 2024 score based on annual performance as there is no previous history to base analysis on	Retain 2023 8.0, though review 2024 score based on annual performance as there is no previous history to base analysis on
Passenger Facilities and Services				
Facilities for passengers who require additional assistance	9.0	9.5	9.0	9.5
Availability of trolleys	8.5	9.0	8.3	9.0
Satisfaction with Wi-Fi	8.5	9.0	8.5	9.0
Sense of safety for my health	No Target	No Target	N/A	N/A

Figure 22: CAR draft proposal SQMs for 2023 to 2026