Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/NA10.20250212
Published: June 17, 2025

Smarter Airports Start with Smarter Reviews: An LLM-Based Framework for Automating Airport Service Quality Analysis

Bismack Tokoli
Department of Data Science and Business Analytics
Florida Polytechnic University
Lakeland, FL, USA
btokoli9770@floridapoly.edu

Dr. Abdulaziz Alhamadani
Assistant Professor
Department of Data Science and Business Analytics
Florida Polytechnic University
Lakeland, FL, USA
aalhamadani@floridapoly.edu

Abstract

Airports play a vital role in global travel, serving millions of passengers daily. With more people traveling by air, airport service quality improvement is more important than ever to ensure continued passenger satisfaction and effective airport operation. Traditional methods for assessing Airport Service Quality (ASQ), such as structured surveys and sentiment analysis, often fail to capture the full spectrum of passenger experiences. Existing studies rely on sentiment scoring and topic modelling, which provide broad insights but lack the depth to identify specific service issues. This research proposes a novel approach leveraging Large Language Models (LLMs) to automate the analysis of airport customer reviews. Our methodology involves a four-phase process: (1) identification of extreme positive and extreme negative reviews using clustering-based methods on the sentiment scores and customer ratings, (2) keyword extraction using LLMs, (3) keywords categorization and (4) topics generation. The proposed method is applied to the Skytrax airport reviews dataset. With the combination of LLM-based keyword extraction and topic modelling, our system gives a richer understanding of passengers' experience in the form of overt concerns such as security delays, unclean facilities, or unfriendly staff interaction. This study aims to help airports make better decisions and improve customer experiences. Our findings demonstrate that LLMs can significantly enhance the granularity and relevance of ASQ insights compared to traditional methods. The proposed framework presents a scalable, automated solution for airport operators to systematically improve passenger experience and operational efficiency based on unstructured feedback.

Keywords
Airport Service Quality, Keyword Extraction, Large Language Models (LLMs), Sentiment Analysis, Airport Reviews

1. Introduction
Airports are critical hubs in the global transportation network, serving as the primary connection between air travel and ground transportation. As the aviation industry continues to grow, with global passenger traffic expected to double by 2040, the quality of services provided at airports has become a key factor in ensuring passenger satisfaction and operational efficiency. Airport service quality comprises of the various aspects of passenger experience within an

airport, including efficiency of check-in processes, cleanliness of facilities, availability of amenities, and the professionalism of staff. According to the Airports Council International (ACI), passenger satisfaction is directly linked to increased non-aeronautical revenues, such as retail and dining sales, which are crucial for the financial sustainability of airports. In 2023, the ACI reported that airports with higher ASQ scores experienced a 10-15% increase in passenger spending, underscoring the economic importance of service quality. Furthermore, passenger feedback has become increasingly influential, with studies indicating that a 1% increase in overall passenger satisfaction can lead to a 1.5% increase in non-aeronautical revenue, emphasizing the direct correlation between service quality and airport profitability. (Pholsook et al. 2019). Given the significant role airports play in the global economy and the increasing expectations of passengers, understanding and improving ASQ is more important than ever.

Despite the importance of passenger feedback, current research methodologies often fall short in fully capturing the nuanced insights embedded within customer reviews. Existing research on ASQ has primarily focused on sentiment analysis and topic modelling, which, while useful, do not fully capture the depth of information contained in customer reviews. For example, studies by Martin-Domingo et al. (2019) have used sentiment analysis and topic modelling to assess ASQ based on Google reviews and Twitter data, respectively. Also, deep learning models, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have been applied to analyze passenger feedback (Barakat et al. 2021). However, these models may struggle to fully understand the context and complexity of natural language, particularly in short, informal texts like tweets or online reviews. This gap in the literature highlights the need for more sophisticated methods that can perform a comprehensive analysis of customer reviews, identifying not just sentiment but also specific issues and actionable insights.

To address these limitations, this paper proposes a novel approach to automating the text analysis of airport service quality customer reviews using Large Language Models (LLMs). By leveraging the advanced natural language processing capabilities of LLMs, we can go beyond simple sentiment scores and topic summaries to identify the key issues passengers are raising and generate actionable recommendations for improving airport services. This approach is particularly relevant in the context of the growing volume of online reviews, which provide a rich but underutilized source of data for understanding passenger experiences. Our proposed methodology involves a four-phase approach to automate the analysis of airport customer reviews. We will perform sentiment analysis to classify reviews based on their sentiment scores and corresponding customer ratings of the airport. This will allow us to cluster reviews into categories reflecting varying levels of passenger satisfaction. We will then employ Large Language Models to conduct an in-depth analysis of the text within each cluster. The goal is to extract key issues pertinent to airport services. Furthermore, we aim to develop an automated system that, upon receiving a new review, can promptly identify salient issues for efficient airport management. Ultimately, this research aims to offer a scalable solution for airports to enhance service quality, leading to improved passenger satisfaction and operational efficiency.

1.1 Objectives

This study aims to enhance the analysis of airport service quality (ASQ) by leveraging advanced natural language processing techniques. Specifically, it seeks to analyze airport customer reviews using sentiment analysis to classify feedback based on passenger satisfaction levels and cluster reviews according to sentiment scores and customer ratings. By employing Large Language Models (LLMs), the study aims to extract key and specific issues beyond conventional sentiment scores and topic summaries, providing a more in-depth understanding of passenger concerns. The research also aims to develop an automated system that can analyze new reviews in real-time and identify critical service issues. Ultimately, this study strives to offer a scalable solution for improving airport service quality, enhancing passenger satisfaction, and optimizing operational efficiency.

2. Literature Review

2.1 Airport Service Quality

Airport service quality has been a critical area of research in the aviation industry, given its direct impact on passenger satisfaction and willingness to recommend. Several studies have explored the dimensions of airport service quality, often using structured surveys and service quality models such as SERVQUAL (Parasuraman et al. 1988). SERVQUAL, which measures service quality based on five dimensions-tangibles, reliability, responsiveness, assurance, and empathy has been widely applied in the airport context (Fodness and Murray 2007; Bogicevic et al. 2013). The Airports Council International (ACI) conducts the widely recognized Airport Service Quality (ASQ) program, which assesses airports based on passenger surveys (ACI 2018). However, these methods are often costly, time-consuming, and provide limited real-time insights. These studies have highlighted the importance of factors such as cleanliness, staff behavior, and efficiency in security and immigration processes as key determinants of passenger satisfaction. For instance, Fodness and Murray (2007) conducted a study on passengers' expectations of airport service quality, identifying that passengers prioritize efficiency, comfort, and convenience. Similarly, Bogicevic et al. (2013) found that airport service quality significantly influences passenger satisfaction and loyalty, with cleanliness and staff behavior being the most critical factors.

However, these studies primarily rely on structured surveys, which, while useful, may not capture the full spectrum of passenger experiences due to limitations such as survey fatigue and low response rates (Luo 2009; Viswanathan and Kayande 2012). Recent studies have explored automated text analysis to evaluate ASQ. Robertson et al. (2023) proposed leveraging online reviews for service quality assessment but primarily used linguistic analysis tools such as LIWC without deep contextual understanding. Also, research by Martín-Domingo et al. (2019) examined service quality through Twitter data but relied on lexicon-based sentiment analysis, which lacks nuance. In contrast to these traditional approaches, our work leverages Large Language Models (LLMs) to automate the analysis of unsolicited online reviews, providing a more dynamic and scalable method for understanding passenger experiences. Our work extends beyond the traditional SERVQUAL dimensions by incorporating sentiment analysis and clustering techniques to identify key issues. Moreover, while much of the existing literature focuses on airports, some studies have examined service quality in the context of airlines (Prentice et al. 2023; Brochado et al. 2019). These studies often highlight the importance of in-flight services, punctuality, and customer service. However, the focus on airlines rather than airports represents a significant gap in the literature, as airports and airlines offer distinct service experiences. Our work addresses this gap by focusing specifically on airport service quality, providing insights that are directly applicable to airport management.

2.2 Sentiment Analysis

Sentiment analysis has emerged as a powerful tool for understanding customer feedback in various industries, including aviation. In the context of airports, sentiment analysis of online reviews allows for the extraction of valuable insights into passenger experiences, which can be used to improve service quality. Several studies have applied sentiment analysis to airport reviews, often using machine learning techniques to classify reviews as positive, negative, or neutral (Korfiatis et al. 2019; Shahbaznezhad and Rashidirad 2021). For example, Korfiatis et al. (2019) used topic modelling and sentiment analysis to analyze airline passengers' online reviews, identifying key themes such as in-flight services, punctuality, and customer service. Similarly, Shahbaznezhad and Rashidirad (2021) explored the relationship between online reviews and customer satisfaction in the airline industry, finding that positive reviews were strongly correlated with higher satisfaction levels.

However, these studies primarily focus on airlines rather than airports, leaving a gap in the literature regarding the application of sentiment analysis to airport service quality. Our work builds on these studies by applying sentiment analysis to airport reviews, providing a more focused understanding of passenger experiences in the airport context. our approach uses BERT to perform advanced sentiment analysis. This allows for a more nuanced analysis of passenger feedback, capturing not only the sentiment but also the underlying emotional tone and language patterns that influence passenger satisfaction. By addressing these gaps, our work contributes to the growing body of literature on sentiment analysis in the aviation industry, offering new methods for analyzing and interpreting customer feedback.

2.3 Automating Airport Customer Reviews

Deep learning models and Large Language Models (LLMs) have gained significant attention in recent years for their ability to analyze large volumes of unstructured data, making them particularly well-suited for automating the analysis of customer service reviews. In the context of airports, deep learning models have been applied to various tasks, including sentiment analysis, topic modelling, and customer feedback classification (Araújo et al. 2020; Lam et al. 2019; Robertson et al. 2021). For instance, Araújo et al. (2020) used deep learning models to perform multilingual sentiment analysis on airline reviews, demonstrating the effectiveness of these models in capturing sentiment across different languages. Similarly, Lam et al. (2019) applied deep learning techniques to analyze wine reviews, highlighting the potential of these models to extract nuanced insights from customer feedback. In the airport context, Robertson et al. (2021) used deep learning models to analyze online reviews of accommodation providers, identifying key patterns that influenced customer satisfaction. While deep learning models and LLMs offer significant advantages in terms of accuracy and scalability, they often require large amounts of labelled data and computational resources, which can be a barrier to their widespread adoption. A major limitation of existing deep learning approaches is their reliance on feature extraction and predefined sentiment categories. Our work addresses these gaps by exploring the potential of LLMs for automating the analysis of airport customer service reviews. This allows for a more comprehensive analysis of passenger experiences, providing airport management with actionable insights for improving service quality.

3. Methods

This research proposes an LLM based automated issue identification from airport reviews. The proposed approach comprises four phases. The first phase involves conducting sentiment analysis on the reviews. The reviews are then clustered based on the results from the sentiment analysis and the overall rating using KMeans, Hierarchical, DBSCAN and Spectral clustering to identify the extremely positive and negative clusters of reviews. The second phase focuses on extracting all the keywords from these clusters using LLMs to capture salient aspects of passenger experiences. In the third phase, the extracted keywords are categorized into meaningful groups, from which topics are generated in the final phase. This structured approach enables the automated identification of key issues airport reviews, facilitating data-driven decision-making for airport management and service improvement. Figure 1 shows the overall framework.

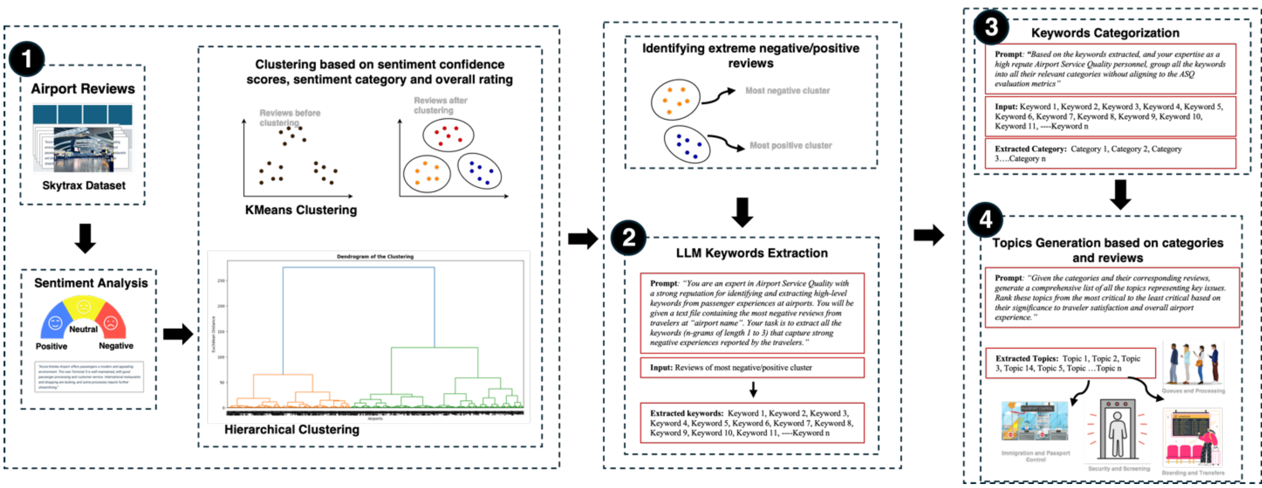

Figure 1. Illustrative Architecture of the framework.

3.1 Sentiment analysis with BERT

Sentiment analysis is pivotal in understanding customer feedback within the aviation industry. This study employs Bidirectional Encoder Representations from Transformers (BERT), a state-of-the-art model in natural language processing (Devlin et al. 2019). BERT generates contextualized word embeddings, capturing semantic and syntactic features of text, and classifies each review into positive, neutral, or negative sentiment categories, assigning confidence scores ranging from -1 to 1. Traditional sentiment analysis models such as VADER, rely on lexicons and rule-based approaches however BERT uses deep learning to understand the contextual meaning of words within a sentence. This enables it to handle complex subtle variations and intricacies in tone, such as sarcasm, negations, and word dependencies, which statistical models often struggle with. Also, BERT's ability to fine-tune on domain-specific data further enhances its accuracy, making it more effective in capturing sentiment variations in diverse customer reviews. (Sun et al. 2019).

3.2 Clustering

The main goal of clustering is to extract the extreme negative and positive reviews of passengers. The "overall rating" grants the passenger the discretion to rate their experience on a scale of 1-10 (lowest to highest). Relying solely on this rating to justify if a passenger had a negative or positive experience may be misleading. The ratings can be biased, subjective or sometimes influenced by the person's mood at a certain time and occasion (Althubiti et al. 2025). For example, a passenger gives a rating of 10 with review *"This is best airport I have used"* and another gives a 5 for a similar review, *"Everything was smooth. Best airport in the world"*. Clearly, the ratings could have been influenced by the mood of the individual. Therefore, to be more objective and unbiased, providing a more reliable way of clustering the reviews, the overall ratings were supplemented with the sentiment scores from the sentiment analysis.

The traditional KMeans was chosen for its simplicity and efficiency in dealing with large datasets. KMeans is an iterative, centroid-based clustering algorithm that partitions a dataset into similar groups based on the distance between their centroids. Agglomerative Hierarchical Clustering (AHC) was used for its ability to reveal nested structures within the data. AHC builds nested clusters by merging them. Spectral Clustering uses eigenvalues of a similarity matrix to reduce the dimensions before applying a clustering algorithm such as the k-means. Spectral clustering was selected for its effectiveness in identifying clusters based on their data-induced graph structure (Petukhova et al. 2025). Finally, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a base algorithm for density-based clustering. It can discover clusters of different shapes and sizes from a large amount of data, which containing noise and outliers. The selected algorithms and their respective parameters are listed in Table 1.

Table 1. Algorithms and respective parameters.

Algorithm	Parameters
K-Means	n_clusters=k, random_state=42, n_init=10
AHC	n_clusters=k, metric='euclidean', linkage='ward'
Spectral	n_clusters=k, affinity='nearest_neighbors', random_state=42, assign_labels='discretize'
DBSCAN	eps=0.014, min_samples=6

3.3 Keywords extraction using LLMs

In this subsection, we describe how the LLM is used in extracting the keywords from the most positive and most negative airport reviews derived from the cluster analysis. Keyword extraction is a fundamental task in natural language processing (NLP) that aims to generate a concise set of terms that best summarize and characterize a given text (Firoozeh et al. 2020). Given the industry-oriented nature of this research, we focus on extractive keyword generation to ensure factual integrity and mitigate the risks associated with LLM hallucination, which can introduce inaccuracies in critical domains such as airport service evaluation (Over and Hurst 2001). To achieve this, we employ ChatGPT, a highly sophisticated transformer-based model, utilizing a structured prompting approach. Specifically, we apply zero-shot prompting, wherein the model is instructed to identify key terms without prior examples, thereby leveraging its extensive pre-trained knowledge base. Zero-shot prompting is computationally efficient as it eliminates the need for fine-tuning or labeled training data, making the process faster and more cost-effective. We incorporate a system message that explicitly defines the LLM's role and operational constraints, ensuring consistency and relevance in keyword extraction. The extracted keywords are constrained to n-grams of length one to three, enabling the model to capture both individual terms and meaningful multi-word expressions indicative of passenger sentiments. A representative example of our prompt for extracting keywords from the most negative reviews is as follows:

Actual Prompt: "You are an expert in Airport Service Quality with a strong reputation for identifying and extracting high-level keywords from passenger experiences at airports. You will be given a text file containing the most negative reviews from travelers at "airport name". Your task is to extract all the keywords (n-grams of length 1 to 3) that capture strong negative experiences reported by the travelers."
Input: Reviews file - The review file is the text file containing all the reviews from the extremely negative cluster for a specified airport. For example, "Heathrow Airport" in London.
Sample Results:
"Rudeness", "Unprofessionalism", "Harassment Indifference", "Unhelpful staff", "Rude staff", "Poor customer service", "Staff indifference", "Overpriced duty-free shops", "Poor airport signage", "Unhelpful customer service"

3.4 Keywords Categorization

Following keyword extraction, the LLM categorizes the extracted keywords into semantically coherent groups.
Categorizing the extracted keywords helps uncover recurring patterns in passenger feedback, making it easier to analyze and address specific concerns. Without categorization, raw keywords may appear fragmented and difficult to interpret, whereas the structured groupings provide a clearer summary of major issues. This process also allows airport management to focus on specific problem areas, such as customer service or baggage handling, rather than analyzing scattered data points. To achieve this, the LLM clusters keywords based on semantic similarity, ensuring that related terms (e.g., "delayed flights" and "long wait times") are grouped together. By leveraging transformer-based word embeddings, the LLM is able recognize synonyms and related phrases. Instead of aligning with the ASQ (Airport Service Quality) evaluation metrics, we adopt an approach that directly reflects passenger concerns as expressed in

their reviews, allowing for a more unbiased categorization of the key issues. The following structured prompt is used to facilitate this step:

Actual Prompt: "Based on the keywords extracted, and your expertise as a high repute Airport Service Quality personnel, group all the keywords into all their relevant categories without aligning to the ASQ evaluation metrics"
Input: Keywords - "Rudeness", "Unprofessionalism", "Harassment Indifference", "Unhelpful staff", "Rude staff", "Poor customer service", "Staff indifference", "Overpriced duty-free shops", "Poor airport signage", "Unhelpful customer service"

Sample Results:
"Customer Service & Staff Conduct" - "Rudeness", "Unprofessionalism", "Harassment Indifference", "Unhelpful staff", "Rude staff", "Poor customer service", "Staff indifference", "Unhelpful customer service"
"Shopping & Dining" - "Overpriced duty-free shops"
"Infrastructure & Facilities" - "Poor airport signage"

3.5 Topics Generation

Following the keywords categorization, the LLM generates all the topics for each category using the categories and the reviews, ranking the topics from the most critical to the least critical. To generate topics, the LLM first analyzes the categorized keywords and their corresponding reviews. It applies topic modeling techniques such as Latent Dirichlet Allocation (LDA) and BERTopic, which uses word embeddings to detect common discussion patterns across multiple reviews. The LLM clusters similar keywords and reviews into distinct topics, ensuring that each topic reflects a meaningful concern. The generated topics are then ranked based on the volume of mentions, sentiment polarity, and contextual severity. Ranking the topics from the most critical to the least critical is essential for prioritizing passenger concerns. Not all issues in airport reviews have the same level of urgency; some, such as security delays or lost baggage, have a greater impact on passenger experience than minor inconveniences like limited seating. By ranking topics, the most pressing concerns can receive immediate attention. To achieve this, the LLM assigns importance to topics based on keyword frequency, sentiment intensity, and the severity of associated complaints, ensuring that reviews mentioning safety risks or operational failures are ranked higher than service-related issues. The LLM evaluates topics by considering the number of reviews mentioning the issue, the average sentiment score, and contextual urgency, such as distinguishing between security concerns and complaints about amenities. By applying NLP-based ranking metrics like TF-IDF and sentiment weighting, the LLM structures topics in a way that ensures major operational challenges are addressed first, followed by service-related improvements and minor complaints. An example of the prompt for generating the topics is given below.

Actual Prompt: "Given the categories and their corresponding reviews, generate a comprehensive list of all the topics representing key issues. Rank these topics from the most critical to the least critical based on their significance to traveler satisfaction and overall airport experience."
Sample Results:
Category: Security & Immigration Inefficiency
Topics: "Long security lines causing delays and missed flights", "Inefficient screening leading to unnecessary secondary checks", "Excessive security checks", "Unfriendly immigration officers", Incompetent security checks, "Chaotic passport control with unclear or misleading directions", "Delayed processing at immigration, especially for non-EU travelers"

4. Data Collection

4.1 Dataset

This study was conducted using the Skytrax airport reviews dataset. Skytrax is the one of the largest and the well-known website where air travelers voice their opinions via text reviews about their experiences from various airports and airlines across the globe. The dataset contains 17,721 reviews from 11,834 passengers across 741 airports over a period of 2002 to 2015. The highest number of reviews authored by one passenger is 64, representing 0.36% of the total number of reviews in the dataset. 82.41% of passengers have only written one review and 99.13% have written less than ten reviews. Each review includes the author's name, the author's country, the airport name, date, the content of the review, overall rating, amongst other variables.

4.2 Data preprocessing

For the dataset, a series of preprocessing steps were taken to ensure the quality of the input data. Initially, missing values from the overall rating column, which is a critical feature for the analysis were removed. Because all the 741 airports in the dataset were not equally represented by the number of reviews, the distribution of the number of reviews per airport was right skewed which ultimately affected the overall rating distribution. Given the skewed distribution of reviews per airport, only airports with more than 100 reviews were selected to ensure a balanced representation. The "overall rating" distribution for the top 5 airports with the highest number of reviews in the dataset did confirm some statistical similarity which shows a good case for the selected subset of the data. Figure 2 shows the overall rating distributions for the top 5 airports with the most reviews. The review contents were pre-processed, and a word cloud was generated to visualize the most frequent words in the reviews.

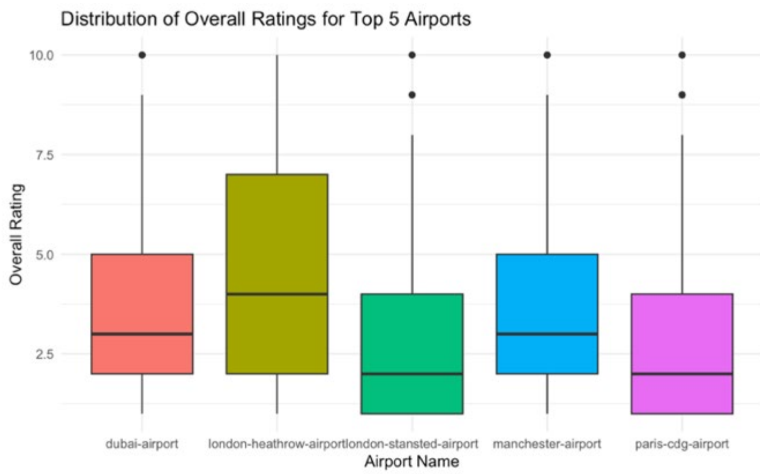

Figure 2. Distribution of Overall Ratings for top 5 airports

5. Results and Discussion

5.1 Clusters Optimality and Quality

The optimal number of clusters and the cluster qualities was determined using the elbow method and the silhouette analysis. The initial approach for the KMeans algorithm leverages the elbow method, a graphical technique that plots the sum of squared distances between data points and their assigned centroids against a range of 2-15 k values. To evaluate the cluster quality, the silhouette method was implemented. This technique evaluated how well each data point fits within its assigned cluster compared to neighboring clusters. Silhouette coefficients range from – 1 to 1, with values closer to 1 indicating superior clustering. The optimal clusters and silhouette scores for the selected algorithms are listed in Table 2.

Table 2. Silhouette Scores for the clustering algorithms.

Algorithm	Optimal k	Silhouette score
K-Means	5	**0.62**
AHC	5	0.59
Spectral	5	0.37
DBSCAN	7	0.54

The KMeans algorithm with K=5 emerges as both optimal and efficient. It yielded the highest silhouette scores of 0.62. The KMeans generated are selected for analysis of the extreme negative and positive clusters. Figure 3 displays the elbow and silhouette plots for the KMeans algorithm and the average ratings per cluster.

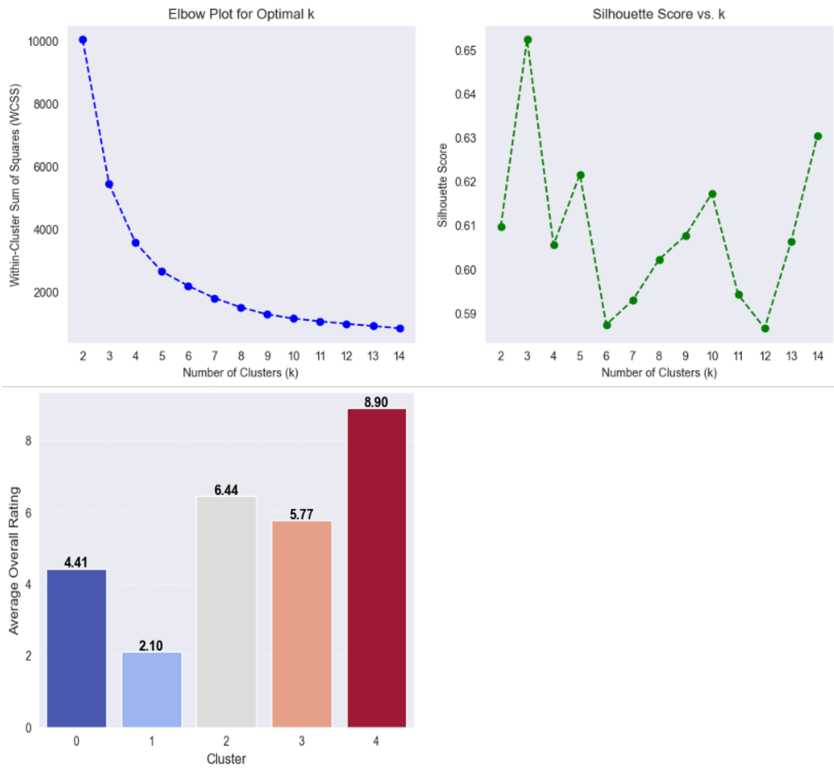

Figure 3. Choosing optimal K for K-Means and average ratings per cluster.

5.2 Cluster Analysis

The optimal five clusters generated by the KMeans algorithm show a significantly clear distinction between the reviews in all the clusters. The average overall ratings per cluster is displayed in Figure 3. Notably, cluster 1 with an average rating of 2.10 is the extremely negative cluster and cluster 4 being the extremely positive cluster. Figure 4 shows the distribution of ratings per cluster.Cluster 1 consists predominantly of low ratings, with most reviews scoring 1.0 (1,265 reviews), followed by 2.0 (980 reviews), 3.0 (771 reviews), and 4.0 (417 reviews). This cluster highlights dissatisfaction among passengers, with a strong concentration of the lowest possible rating. The high frequency of 1-star ratings (over 1,200 instances) suggests severe dissatisfaction, potentially due to issues such as poor customer service (e.g., rude staff interactions, lack of assistance), long wait times (delays in baggage handling, security checks, or immigration), unhygienic or poorly maintained facilities (dirty restrooms, inadequate seating, overcrowding), inefficient airport processes (check-in, security, or boarding inefficiencies).

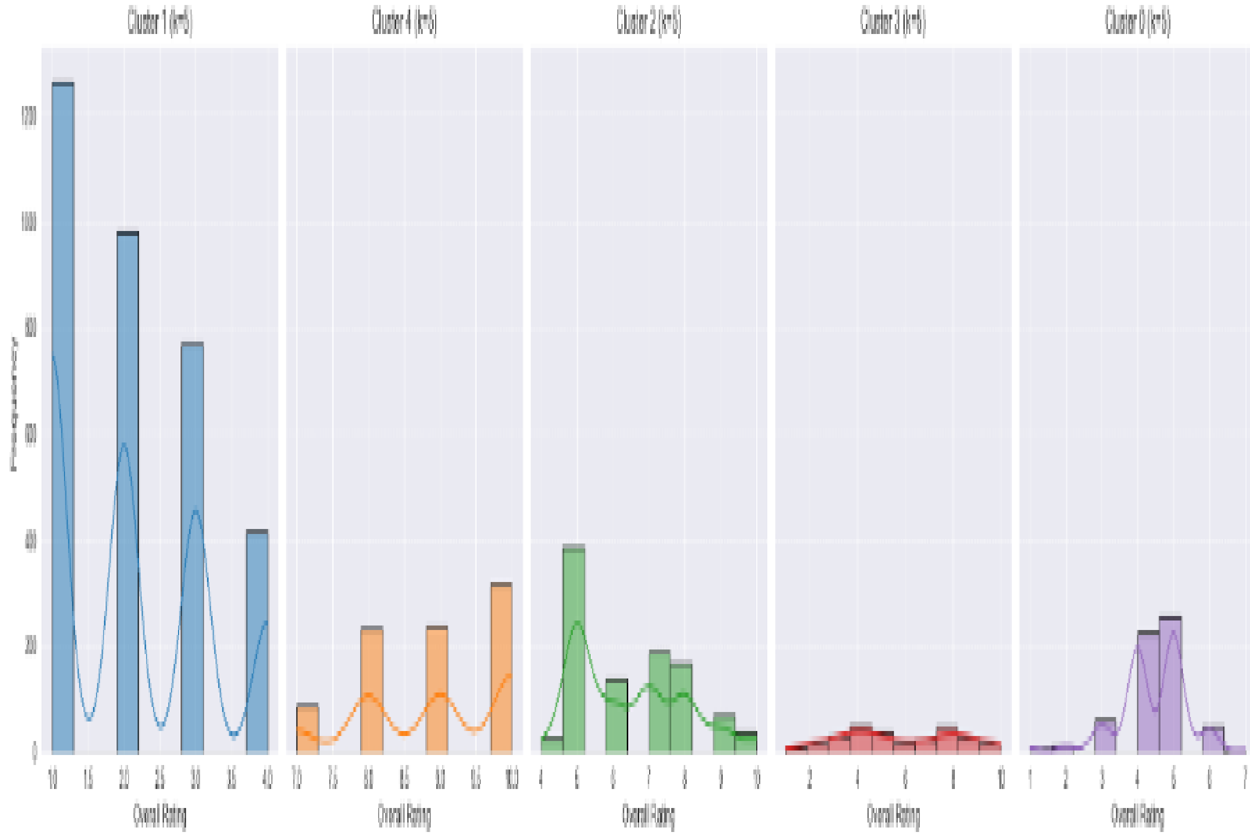

Figure 4. Distribution of overall ratings per cluster.

The presence of 2-star and 3-star ratings also indicates that some passengers had mixed experiences, where certain aspects of their journey were unsatisfactory but not entirely negative.

In contrast, Cluster 4 contains the most positive passenger experiences, with reviews heavily concentrated at the upper end of the rating scale. The distribution is as follows: 10.0 (318 reviews) – Highest Rating, 9.0 (237 reviews), 8.0 (232 reviews), 7.0 (88 reviews). This cluster represents passengers who were highly satisfied with their airport experience. common positive factors contributing to these ratings may include efficient airport operations (smooth check-in, security, and boarding processes), excellent customer service (friendly and helpful staff interactions), comfortable facilities (clean lounges, ample seating, modern infrastructure), seamless connectivity (efficient baggage handling, good transportation links). The presence of ratings between 7 and 10 suggests a consistently positive experience, with some minor areas for improvement (e.g., occasional delays or minor inconveniences). The pseudocode for the overall framework is shown below:

Proposed Method
Input: AirportReview Data REVattr_N **Output:** Topic-Based Review Data TOPICREV_O Initialize: K-clusters Optimal value k = 0, i = 0 1: **while** i < N **do** 2: ApplySentimentAnalysis(REVtext_i) 3: AppendSentimentScores(REVpos, neg, neu, com_i) 4: **end while** 5: **procedure** K-MEANS(REVpos, neg, neu, com, overall_rating_i) 6: Find and assign optimal value k for clusters 7: Assign and append cluster labels to each row REVclu_N 8: Retrieve Most Positive and Negative Clusters from REVclu_N 9: Extract Keywords from Positive and Negative Clusters using LLM 10: Group Extracted Keywords into Categories using LLM 11: Extract Topics from Categories and Associated Reviews 12: Retrieve All Reviews Corresponding to Each Topic 13: Store Extracted Topics and Corresponding Reviews to TOPICREV_O 14: **end procedure** **Return** TOPICREV_O

5.4 Case Study

In this section we perform some case studies to demonstrate the capabilities of the Large Language model in extracting the keywords, categories and topics from the extremely positive and negative clusters. The first case study demonstrates how LLM can identify keywords and from the most positive and negative reviews of Heathrow Airport in London and the second case study focuses on the Dubai Airport. According to the Airport Councils International, several fundamental issues such as long check-in process, security, and environmental challenges play a crucial role in how travelers feel at a particular airport. Tables 3 and 4 summarizes the keywords, categories and topics generated for both the negative and positive reviews for both airports.

Passenger reviews of London Heathrow Airport highlight both positive and negative experiences, emphasizing key areas of strength and concern. Many travelers appreciate the airport’s modern and well-maintained facilities, spacious terminals, and extensive duty-free shopping options, which enhance the overall experience. Positive interactions with efficient security staff and quick passport control processes contribute to a smooth travel experience for some passengers. Also, the convenience of well-organized transit connections and comfortable waiting areas underscores the airport’s efforts to accommodate travelers. Clean restrooms and a diverse selection of dining options further add to passenger satisfaction, particularly for those with long layovers. These aspects, when functioning efficiently, significantly contribute to Heathrow’s reputation as a major international hub. However, frequent complaints about long security queues, inefficient baggage handling, and unhelpful staff indicate areas requiring urgent improvement. Many passengers report delays at immigration, unclear terminal signage, and a lack of assistance, particularly for elderly and disabled travelers. Overcrowding in waiting areas, slow baggage retrieval, and expensive airport food further contribute to dissatisfaction, especially during peak travel hours. Negative feedback on inconsistent customer service and logistical inefficiencies suggests that operational improvements are necessary. While Heathrow’s modern infrastructure and amenities receive praise, addressing these persistent concerns would greatly enhance the overall passenger experience and solidify its standing as a leading global airport.

Table 3. Keywords and Topics for London Heathrow Airport

London Heathrow Airport		
	Negative	Positive
Sample Keywords	“security”, “immigration”, “passport”, “customs”, “screening”, “security check”, “passport control”, “slow immigration”, “inefficient staff”, “long security queues”, “inefficient passport control”, “rude immigration officers”, “staff”, “assistance”, “rude”, “unhelpful”, “rude staff”, “inefficient staff”, “slow	“Clean toilets”, “Well-organized terminal”, “Easy transit”, “Duty-free variety”, “Spacious lounge”, “Comfortable seating”, “Security efficiency”, “Helpful service”, “Excellent airport experience”
Sample Categories	1. Security & Immigration inefficiency 2. Poor Customer Service & Staff Conduct 3. Terminal Navigation & Wayfinding Issues	1. Efficiency and Queues 2. Staff Service and Friendliness 3. Convenience
Sample Topics	1. Security & Immigration Inefficiency (Most Critical) ● Long security lines causing delays and missed flights ● Inefficient screening leading to unnecessary secondary checks ● Incompetent security checks with random selection and re-scanning 2. Poor Customer Service & Staff Conduct ● Rude and unprofessional staff, particularly at security and check-in ● Harassment and power abuse by certain employees, particularly during security checks	1. Efficiency and Queue (Most Critical) ● Fast immigration process ● Quick passport control ● Smooth security checks 2. Staff Service and Friendliness ● Efficient ground staff ● Helpful customer service ● Polite security personnel

Table 4. Keywords and Topics for Dubai Airport

Dubai Airport		
	Negative	Positive
Sample Keywords	“Immigration queue”, “Passport control”, “Security check”, “Long immigration queues”, “Security check chaos”, “Transit security nightmare”, “Extremely slow immigration”	“Clean”, “Spacious”, “Efficient”, “Helpful”, “Modern”, “Comfortable”, “Fast”, “Organized”, “Smooth”, “Convenient”
Sample Categories	1. Immigration & Security Delays 2. Overcrowding & Congestion 3. Poor Facilities & Cleanliness	1. Efficient Immigration & Security 2. Friendly & Helpful Staff 3. Comfort & Convenience
Sample Topics	1. Immigration & Security Delays (Most Critical) ● Long immigration queues ● Chaotic and understaffed security checks 2. Overcrowding & Congestion ● Severe overcrowding in terminals and transit areas ● Overcrowded and disorganized boarding gates	1. Efficient Immigration & Security (Most Critical) ● Fast immigration process ● Smooth security checks ● Hassle-free transit 2. Friendly & Helpful Staff ● Professional and courteous service ● Polite security personnel

Positive experiences at Dubai Airport highlighted its modern and efficient infrastructure, with many praising the spacious terminals and smooth security processes. Passengers appreciated the fast immigration and quick passport control, which contributed to a hassle-free transit experience. The availability of extensive duty-free shopping and a wide range of dining options also received positive feedback, enhancing the overall travel satisfaction. The consistent mention of clean and well-maintained facilities, along with helpful and friendly staff, further underscored the airport's commitment to providing a comfortable and convenient environment for travelers. However, Dubai Airport also faces significant challenges, particularly regarding immigration and security delays, and overcrowding. Frequent complaints about long immigration queues, slow passport control, and chaotic security checks indicate critical areas needing immediate improvement. Overcrowding in terminals and insufficient seating were also common concerns, leading to a frustrating experience for many passengers. Additionally, negative feedback regarding poor facilities, such as filthy toilets and unreliable WiFi, suggests a need for enhanced maintenance and digital services. The presence of unhelpful or rude staff, coupled with flight and transit issues, further exacerbates passenger dissatisfaction, highlighting the necessity for better staff training and improved operational efficiency.

6. Conclusion

As air travel continues to grow and evolve, ensuring high Airport Service Quality (ASQ) is essential for maintaining passenger satisfaction and operational effectiveness. This paper introduces a novel, scalable framework for automating the analysis of airport customer reviews using Large Language Models (LLMs). By leveraging a four-phase methodology combining sentiment analysis, clustering, keyword extraction, and topic generation, we bridge the gap between high-level sentiment scoring and the granular insights required for actionable improvements. Our approach not only identifies broad sentiment trends but also uncovers the specific, context-rich issues that traditional methods often overlook. We demonstrate the effectiveness of LLMs in enhancing the depth, clarity, and relevance of ASQ analysis using the Skytrax dataset. Case studies of Heathrow and Dubai airports showed the method's effectiveness in extracting meaningful insights from reviews, helping airport management identify areas for improvement and strengths. In conclusion, this research highlights the transformative potential of LLMs in the aviation domain and as airports strive for excellence in an increasingly competitive environment, our proposed method offers a powerful tool for continuous feedback analysis and quality improvement, turning unstructured passenger voices into structured, impactful insights. Due to space limitations, we are unable to show the full output of the keywords, categories and topics.

References

Althubiti, K., Alhamadani, A., Khan, M. and Shah, M.G.H., Unveiling negative memorable experiences of hotel guests: An innovative algorithmic analysis, *International Journal of Hospitality Management*, vol. 126, p. 104087, 2025.

Araújo, M., Pereira, A. and Benevenuto, F., A comparative study of machine translation for multilingual sentence-level sentiment analysis, *Information Sciences*, vol. 512, pp. 1078–1102, 2020.

Barakat, H., Yeniterzi, R. and Martín-Domingo, L., Applying deep learning models to Twitter data to detect airport service quality, *Journal of Air Transport Management*, vol. 91, p. 102003, 2021.

Bogicevic, V., Yang, W., Bilgihan, A. and Bujisic, M., Airport service quality drivers of passenger satisfaction, *Tourism Review*, vol. 68, no. 4, pp. 3–18, 2013.

Brochado, A., Rita, P., Oliveira, C. and Oliveira, F., Airline passengers' perceptions of service quality: Themes in online reviews, *International Journal of Contemporary Hospitality Management*, vol. 31, no. 2, pp. 855–873, 2019.

Devlin, J., Chang, M. W., Lee, K. and Toutanova, K., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:1810.04805, 2019.

Dhini, A. and Kusumaningrum, D. A., Sentiment analysis of airport customer reviews, *Proceedings of the 2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pp. 502-506, Bangkok, Thailand, 2018, https://doi.org/10.1109/IEEM.2018.8607335.Reimer, D., Entrepreneurship and Innovation, Available: http://www.ieomsociet.org/ieom/newsletters/, July 2020.

Firoozeh, N., Nazarenko, A., Alizon, F. and Daille, B., Keyword extraction: Issues and methods, *Natural Language Engineering*, vol. 26, no. 3, pp. 259-291, 2020.

Fodness, D., & Murray, B. (2007). Passengers' expectations of airport service quality. The Journal of Services Marketing, 21(7), 492–506.

Korfiatis, N., Stamolampros, P., Kourouthanasis, P. and Sagladinos, V., Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews, *Expert Systems with Applications*, vol. 116, pp. 472–486, 2019.

Lam, J., Lambrechts, M., Pitt, C. and Afsharipour, A., When writing about wine: How ratings impact reviews, *Journal of Wine Research*, vol. 30, no. 4, pp. 335–345, 2019.

Luo, Y., Using internet data collection in marketing research, *International Business Research*, vol. 2, no. 1, pp. 196–202, 2009.

Martin-Domingo, L., Martín, J. C. and Mandsberg, G., Social media as a resource for sentiment analysis of airport service quality (ASQ), *Journal of Air Transport Management*, vol. 78, pp. 106-115, 2019.

Over, P. and Hurst, M., Introduction to duc-2001: an intrinsic evaluation of generic news text summarization systems, *Proceedings of DUC2001*, 2001.

Parasuraman, A., Zeithaml, A. V. and Berry, L., SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality, *Journal of Retailing*, vol. 64, no. 1, pp. 12–40, 1988.

Petukhova, A., Matos-Carvalho, J.P. and Fachada, N., Text clustering with large language model embeddings, *International Journal of Cognitive Computing in Engineering*, vol. 6, pp. 100-108, 2025.

Pholsook, T., Wipulanusat, W., Thamsatitdej, P., Ramjan, S., Sunkpho, J. and Ratanavaraha, V., A three-stage hybrid SEM-BN-ANN approach for analyzing airport service quality, *Sustainability*, vol. 15, no. 11, p. 8885, 2023.

Prentice, C., Hsiao, A., Wang, X. and Loureiro, S. M. C., Mind, service quality, relationship with airlines, *Journal of Strategic Marketing*, vol. 31, no. 1, pp. 212–234, 2023.

Robertson, J., Ferreira, C. and Paschen, J., Reading between the lines: Understanding customer experience with disruptive technology through online reviews, *Australasian Marketing Journal*, vol. 29, no. 3, pp. 215–224, 2021.

Robertson, J., Vella, J., Duncan, S., Pitt, C. and Caruana, A., Beyond surveys: Leveraging automated text analysis of travellers' online reviews to enhance service quality and willingness to recommend, *Journal of Strategic Marketing*, 2023.

Shahbaznezhad, H. and Rashidirad, M., Exploring firms' fan page behavior and users' participation: Evidence from airline industry on Twitter, *Journal of Strategic Marketing*, vol. 29, no. 6, pp. 492–513, 2021.

Skytrax, Skytrax Airport Reviews Dataset, Available: https://www.skytraxratings.com, Accessed on January 29, 2025.

Sun, C., Qiu, X., Xu, Y. and Huang, X., How to Fine-Tune BERT for Text Classification, China National Conference on Chinese Computational Linguistics, pp. 194-206, 2019.

Viswanathan, M. and Kayande, U., Commentary on "common method bias in Marketing: Causes, mechanisms, and procedural remedies", *Journal of Retailing*, vol. 88, no. 4, pp. 556–562, 2012.

Biographies

Bismack Tokoli is an MS. Data Science student at Florida Polytechnic University with strong research interests in machine learning, deep learning and the applications of Artificial Intelligence. He holds a BSc in Statistics and hopes to expand his knowledge base through impactful research.

Dr. Abdulaziz Alhamadani is an Assistant Professor of Data Science and Business Analytics at Florida Polytechnic University. His major research interests span machine learning, data mining, and deep learning, with a focus on applications such as text classification, event detection, pandemic forecasting, and ethical AI. Dr. Alhamadani graduated from Virginia Tech in 2024 with a Ph.D. in Computer Science and its applications. His research has been published in leading journals and conferences such as *EMNLP*, *IEEE Big Data*, and *ASONAM*. He is dedicated to student-centered teaching and actively mentors' students in research and applied data science. Dr. Alhamadani also contributes to the academic community as a peer reviewer and collaborator in interdisciplinary research initiatives.