Analysis of Airline Service Quality Through Online Reviews: A Multi-Agent LLM Approach Based on SERVPERF

Author: Jaleh Ashrafi

Supervisor: Hamid Bekamiri

Master Thesis
Spring 2025

Acknowledgment

I would like to present my gratitude to my supervisor, Mr. Hamid Bekamiri, for his insightful guidance, support, and friendly attitude during the process of writing this thesis. His knowledge and valuable feedback significantly influenced the direction and quality of my research. His encouragement gave me the strength and motivation to overcome challenges and successfully complete this academic journey.

I am also deeply thankful to my husband, Sajjad, for his unconditional love, patience, and constant support. His belief in me gave me the strength and confidence to keep going. This achievement would not have been possible without him.

I would also like to express my heartfelt gratitude to my family and friends, whether near or far. Their emotional support and understanding meant a lot to me throughout this process.

Table of Contents

List of Figures

List of Tables

Abstract

This research aims to analyze airline service quality by integrating service evaluation models with artificial intelligence (AI) techniques. Specifically, it combines the SERVPERF model, which measures perceived service performance across five dimensions (tangibles, reliability, assurance, responsiveness, and empathy), with modern tools from the Smart Tourism Ecosystem, including big data analytics and Large Language Models (LLMs). The study analyzes 10,115 verified online reviews from 11 European airlines sourced from the Skytrax platform. It implements a three-stage methodology, which begins with a qualitative content analysis of 50 reviews to identify key service themes and sub-themes based on customer feedback, and then links them to the SERVPERF model. These extracted themes were then used to guide aspect-based sentiment classification with a multi-agent setup involving Gemini 1.5 Flash and GPT-3.5 Turbo. Sentiment polarity (positive, neutral, or negative) was assigned to each SERVPERF dimension, and the results were validated using manual review and micro F1 score. Finally, Ordinary Least Squares (OLS) regression was performed on the sentiment-labeled aspects to assess the significance of each service dimension in shaping customer satisfaction. The study offers several significant contributions. It demonstrates how deep learning techniques and LLMs can be applied to extract structured insights from unstructured, large-scale user-generated data, providing an efficient complementary approach to conventional survey-based research. This marks a significant advancement in service quality research by enabling high-volume, real-time evaluation across multiple service dimensions. Moreover, the findings offer actionable implications for airline managers and service designers. By identifying the most influential service quality dimensions from the passenger's perspective, airlines can prioritize the most important features for their customers. This prioritization, when aligned with a structured framework, enables real-time monitoring of service feedback and supports data-driven decision-making, ultimately facilitating strategic adjustments and enhanced passenger experiences.

In summary, this research highlights how traditional models like SERVPERF can be effectively applied and extended through AI-powered tools, offering practical insights into service evaluation within the Smart Tourism Ecosystem. This integration points out the potential of intelligent technologies to develop service innovation, adapt to dynamic customer expectations, and contribute to the development of a smarter tourism system.

Keywords: Airline Service Quality, SERVPERF, Large Language Models, Sentiment Analysis, Online Reviews.

1. Introduction

In today's highly competitive tourism industry, service quality can significantly shape customer satisfaction and overall business performance. Eraqi (2006) points out that relying only on price and promotional strategies is no longer sufficient in this sector. Though these actions remain significant, they cannot take the place of quality. Thus, to remain competitive, companies must instead prioritize new strategic goals, emphasizing quality-focused policies (Eraqi, 2006, as cited in Silvestri et al., 2017). This change reflects a broader understanding that customer satisfaction depends not only on cost-effectiveness but also on meeting a certain standard of service quality. In this regard, service quality has been shown to directly and significantly impact customer satisfaction (Costabile, 2001, as cited in Silvestri et al., 2017). Building on this connection, it becomes essential for tourism stakeholders to understand how service quality is perceived and measured, particularly if they aim to optimize operations and remain competitive. As tourism businesses became more aware of the importance of quality, the need to measure service quality also grew. This growing need led to the development of various tools and methods designed to evaluate service performance (Rodrigues et al., 2013). In the context of tourism, air transport is a critical facilitator of destination accessibility, contributing significantly to the expansion and integration of the tourism industry. Beyond its role in tourism, the airline industry is recognized as one of the world's major service sectors, essential in boosting national GDP, creating employment opportunities, fostering economic development, and generating tax revenue (Libent & Magasi, 2024As the demand for air travel continues to rise, particularly in recent years, the industry has experienced a shift in strategic priorities, placing greater emphasis on customer satisfaction as a key competitive advantage (Dike et al., 2023, as cited in Murugesan et al., 2024). To achieve customer satisfaction, service quality has been identified as a vital factor influencing passenger satisfaction and loyalty (Siqueira et al., 2023, as cited in Murugesan et al., 2024). Therefore, a deeper understanding of the elements that cause both airline passengers' positive and negative experiences can enable airlines to enhance their service offerings and strengthen long-term customer relationships (Suryani et al., 2023, as cited in Murugesan et al., 2024). As one of the main stakeholders in shaping a tourist's journey, airlines are recommended to ensure high service quality to meet customer expectations and enhance satisfaction. To fulfill this objective, airlines need to develop an accurate insight into customer needs and preferences, which is essential for aligning service delivery with expectations and improving overall market performance (Tahanisaz, 2020, as cited in Kim et al., 2024; Tsafarakis et al., 2018, as cited in Kim et al., 2024).

Another point that needs to be considered is that while all service quality dimensions contribute to overall service quality, some have a more immediate and substantial impact on customer satisfaction. Airlines are encouraged to strategically prioritize their improvements based on customer expectations and limited budgets. In this regard, assessing the comparative weight of each service dimension enables more targeted and efficient investments, ultimately leading to better service outcomes and enhanced customer satisfaction (Liu & Chen, 2022). For airlines operating under resource constraints, such analysis provides a strategic foundation for prioritizing improvements that align most closely with passenger expectations. However, identifying these priorities accurately depends on how customer perceptions are measured and understood. Much of the existing research in tourism and airline service quality has relied on survey-based methods or expert evaluations, which typically involve small, predefined sample groups and may not reflect the diversity of real-world travelers (Song & Liu, 2017, p. 16). While these approaches have

offered valuable insights, they often depend on subjective judgments and structured questionnaires, which possibly affect the objectivity of the research and narrow its generalizability (Liu & Chen, 2022). Since existing studies have struggled to assess details of service quality based on extracted service attributes accurately (James et al., 2017, as cited in Liu & Chen, 2022), researchers suggest integrating traditional service quality models with text mining approaches, which can enhance dimension extraction and provide a more comprehensive evaluation (Liu & Chen, 2022). In line with this recommendation, this study employs deep learning techniques to analyze large-scale, unstructured user-generated reviews, offering a more comprehensive and data-driven perspective on service quality. Doing so helps to provide a more objective and data-driven understanding of service quality, identifying key factors that impact airline passenger satisfaction. Therefore, with Machine Learning (ML), a considerable amount of customer-generated data can be analyzed and even discover hidden patterns that manual research methods might not consider. Based on the limitations discussed and the opportunities presented by data-driven approaches, this study proposes the following research questions:

- *How can airline service quality be measured on a large-scale dataset?*
- *Which service quality dimensions most significantly influence customer satisfaction across different airlines based on online reviews?*

This research contributes to the growing application of AI in tourism and service management by introducing a scalable and data-driven approach to measuring airline service quality. Through the use of deep learning, it offers a more objective and comprehensive analysis of passenger experiences, enabling airlines to understand customer needs better. These insights help airlines allocate resources more effectively, address critical service quality dimensions, and maintain a competitive position in the global aviation market.

2. Literature review

This literature review explains how new technologies such as big data, smart tourism systems, and AI are helping the tourism and aviation industries to assess their service quality. It shows how ML and NLP facilitate understanding customer experiences more accurately and in real-time. The review starts by looking at theoretical service quality models such as SERVQUAL and SERVPERF, which are often used to evaluate how well services meet customer expectations (Parasuraman et al., 1988; Cronin & Taylor, 1992). It also explains how online reviews have become an important source of feedback, offering detailed and honest insights into customer satisfaction. As the amount of customer data increases, especially through reviews and social media, the use of big data analytics is becoming more important. These tools help businesses better understand trends, predict customer needs, and improve decision-making. Next, the review highlights how AI and ML are being used in the tourism industry to personalize services. It explains how binary and multi-label classification support tasks like aspect-based sentiment analysis (ABSA) and are effective in analyzing customer feedback. Finally, advanced tools such as NLP and LLMs are discussed for their ability to interpret language, while LLM agents provide a dynamic, intelligent service analysis.

2.1 Service Quality and Tourism Context

2.1.1. Service quality

Service quality is a fundamental concept in the service industry that leads to customer satisfaction and business success, especially in aviation. Mustafa et al. (2005) highlight that improving the quality of service offered to passengers is a key concern for airlines. Consequently, providing exceptional service is currently recognized as a strategic necessity for maintaining competitiveness in the airline industry (Mustafa et al., 2005). In this context, service quality is widely acknowledged as a primary driver of customer satisfaction. Since the airline industry relies heavily on customer satisfaction, maintaining high service quality is essential for retaining customers and sustaining long-term profitability. As competition among airlines grows, service quality becomes a key differentiating factor. To fully explore the concept of service quality, we need to understand what a service truly means. The concept of services has been widely discussed in the literature, but there is no single definition that is universally accepted. However, various scholars provide helpful perspectives that shape this study's understanding of services, particularly in the tourism and airline sectors. The literature describes services as processes or actions performed for the benefit of others, often involving specialized skills and knowledge. Despite their different forms, all services rely on interaction between the provider and the customer to create value (Hartwig & Billert, 2018). From this perspective, a service is not a product but an experience that is co-produced by both parties. Lau, Wang, and Chuang (2011) define service as "*a process which has four fundamental elements: provider, client, mission, and value*" and further state that "*a service is a process by which the provider fulfills a mission for a client so that value is created for each of the two stakeholders*" (p. 50). This definition emphasizes the dynamic and relational nature of services, particularly in sectors such as tourism and aviation, where the interaction between service providers and customers is central to the value-creation process.

Now that the concept of service has been defined, the focus shifts to how service quality can be measured. Since the relationship between the provider and the client is essential in services, it also

complicates the measurement of how well the service is delivered. This complexity arises mainly from the typical characteristics of services: their intangibility, variability, and inseparability (Hartwig & Billert, 2018). First, service quality cannot be physically examined or quantified prior to consumption, making it difficult to evaluate in advance (Pollack, 2009). Second, services are often produced and consumed simultaneously, meaning customers evaluate them as they are delivered (Pollack, 2009). Third, service outcomes may vary depending on the people involved, providers and customers, highlighting the heterogeneity of services and making consistency a challenge. In addition, the customer often plays an active role in the delivery process, so their perception and participation directly shape how the service is experienced and judged (Hartwig & Billert, 2018).
Despite these service features, businesses require structured models and frameworks to assess service quality effectively. Without the support of structured evaluation models, it becomes increasingly complex for organizations to implement consistent, customer-focused improvements. In response to these challenges, researchers have introduced several theoretical models that offer structured approaches for evaluating and interpreting service quality.

Among these models, SERVQUAL has been widely applied by both academics and practicing managers across industries, particularly tourism. This model was introduced by Parasuraman et al. (1985) and initially consisted of ten service quality dimensions: tangibles, reliability, responsiveness, understanding the customers, access, communication, credibility, security, competence, and courtesy. The same authors later refined this model in 1988, reducing the original ten dimensions to five (tangibles, reliability, responsiveness, assurance, and empathy), leading to the development of the widely recognized SERVQUAL model (Parasuraman et al., 1988). This model considers the difference between expected and perceived service as the quality of the service itself (Parasuraman et al., 1988). Despite its broad application, the SERVQUAL model has several limitations. One major concern is that it focuses primarily on the functional aspects of service delivery, how the service is provided, while neglecting the technical quality or the actual outcome of the service (Liu & Chen, 2022). This can be particularly problematic in industries such as aviation, where the result (e.g., flight safety) is just as important as the service process itself (Wu & Cheng, 2013). In response to the limitations of the SERVQUAL model, Cronin and Taylor (1992) proposed a revised model known as SERVPERF, which focuses on performance-based evaluation of service quality (Wu & Cheng, 2013). While SERVQUAL assesses service quality by measuring the gap between what customers expect and what they perceive they have received, SERVPERF takes a different approach (Wu & Cheng, 2013). SERVPERF is also among the most widely recognized models, as it focuses particularly on customers' perceptions of service performance. Cronin and Taylor (1992) argued that it is unnecessary to include expectation-related items in service quality assessments, as customers naturally and implicitly compare their perceptions with expectations during the evaluation process. It focuses exclusively on the customer's evaluation of actual service performance, excluding the expectation component entirely (Cronin & Taylor, 1994). This simplified, performance-only model has gained considerable attention and is applied in various industries, including the airline sector, as a more direct and reliable method for assessing service quality (Wu & Cheng, 2013). The SERVPERF model retains the exact five dimensions as the SERVQUAL model: tangibles, reliability, responsiveness, assurance, and empathy. However, it differs by assessing service quality exclusively based on performance or customer perceptions without referencing expectations. (Wu & Cheng, 2013).

While models like SERVPERF offer structured and validated frameworks for evaluating service quality, their effectiveness depends on the accuracy and relevance of the input data. The next section will discuss online reviews as a valuable source for understanding honest customer opinions and evaluating their service experiences.

2.1.2. Online Reviews as a Source of Customer Experience Insights

In a time of wide accessibility to technology, online reviews can be considered a valuable resource for analyzing to uncover customer experience. Customers voluntarily share their experiences and opinions on various digital platforms and social media, offering a more authentic and high-volume source of service evaluations (Palese & Usai, 2018). Alongside this trend, the volume of online reviews for products and services has significantly increased. These reviews often include numerical ratings and detailed text that reflects the level and the reason for customer satisfaction or dissatisfaction (Liu & Chen, 2022). Such rich narratives often represent customers' real-life experiences, including their opinions, expectations, and emotional responses to the services they receive (Samir et al., 2023). So, these detailed reviews uncover the complexity of customer experiences in real time and provide qualitative data. Additionally, as online textual reviews have an open structure, a large-scale data sample, and the anonymous nature of contributors (Xu et al., 2017), these reviews tend to be more unbiased, aligned with customer experiences, and reliable (Sánchez-Franco et al., 2019). Berezina et al. (2016) further explain that customer online textual reviews show customer experiences in a more detailed way because of their open structure and can therefore reflect customer perceptions more accurately (Berezina et al., 2016, as cited in Xu et al., 2017). Since the investment cost in the hospitality industry is high and the online reviews are a key factor in consumer attitude and purchase intentions, it is valuable to examine service characteristics that tourists choose to highlight, relive, and narrate in their online reviews (Sánchez-Franco et al., 2019). However, despite these advantages, there are some shortcomings to online reviews. Customers vary in their individual needs and priorities when evaluating a service. While some customers may give greater importance to price, others may prioritize service quality or specific features. As a result, two customers may assign different ratings, but have received similar service (Samir et al., 2023). Another challenge is the multifaceted feedback from customers. Customers often give detailed feedback, highlighting strengths in some areas while pointing out weaknesses in others (Mudambi & Schuff, 2010). Consequently, this variation makes it difficult to accurately interpret customer preferences and expectations through online reviews and poses a challenge for delivering personalized and meaningful feedback (Samir et al., 2023).

To address the challenges of analyzing large volumes of textual online content, big data analytics has emerged as a suitable solution, as will discussed in the next section. By combining insights from online reviews with big data technologies within the smart tourism ecosystem, tourism businesses can more effectively monitor service quality and enhance customer satisfaction.

2.1.3. Big Data in tourism

Online reviews have greatly increased the amount of user-generated content in the tourism industry. This large volume of online reviews is a component of big data, as it is vast in scale and diverse in content. According to Song and Liu (2017), big data refers to datasets that are so extensive or complex that traditional data processing methods and software are insufficient for capturing, managing, and analyzing them efficiently (Song & Liu, 2017, p. 13). Consequently, this has led to the increasing use of big data analytics in tourism, offering new opportunities. As noted by Song and Liu (2017), in today's data-driven environment, organizations increasingly seek to extract actionable insights from big data to identify trends, enhance decision-making, and create new business opportunities (Song & Liu, 2017, p. 13). This shift towards big data analytics in tourism research allows for more comprehensive and objective insights, reducing biases often present in traditional survey-based studies. Moreover, using tourism big data through innovative analytical methods presents several advantages over traditional research approaches. Unlike conventional methodologies that rely on survey responses or self-reported intentions, big data draws directly from users' actual behaviors and interactions. In other words, it enables the analysis of real actions rather than relying on what individuals claim they would do or how they respond to predefined questions (Song & Liu, 2017, p. 16). This feature is crucial, as it relies on observed behavior rather than self-reported answers, enhancing research findings' accuracy.
An additional significant benefit is that when considering all available information sources collectively, it becomes evident that big data expands the sample size far beyond what conventional research typically utilizes, often by several orders of magnitude (Meeker & Hong, 2014, as cited in Song & Liu, 2017). The strength of big data lies in its ability to incorporate vast and diverse datasets, reducing the risk of bias that can result from limited or incomplete samples. As noted by Song and Liu (2017), this enhanced reliability enables a more holistic and accurate analysis, leading to conclusions that better reflect the full scope of the data rather than being constrained by traditional sampling limitations. (Song & Liu, 2017, p. 17). Another key benefit is that it is generated by tourists themselves, making it a direct and valuable source of insight into consumer behavior. This type of data significantly enriches tourism businesses' understanding of their target markets and proves particularly useful in analyzing consumer demand for a wide range of tourism products and services (Hendrik & Perdana, 2014; as cited in Song & Liu, 2017, p. 17). In addition to its origin, tourism big data is often structured and adaptable, allowing it to be linked with other information sources such as social media content and open public datasets. This capacity for cross-referencing enhances the depth and flexibility of analysis, whether using currently available data or integrating new data sources as they emerge (Song & Liu, 2017, p. 17).

As a result, decision-makers in tourism can develop more precise marketing strategies and service improvements based on a larger and more reliable data set. In addition, big data makes sure that tourism services align with actual visitor expectations. Such capabilities are essential when applying the SERVPERF dimensions (e.g., reliability, responsiveness, empathy) in a modern, data-driven context. Therefore, the potential for big data in tourism is huge, and it is essential that tourism organizations recognize its strategic significance rather than underestimate its value (Song & Liu, 2017, p. 19). By applying big data techniques, the tourism industry can evaluate service quality dimensions and optimize service delivery. In this context, big data analytics supports a smart tourism ecosystem, which enables the assessment of service quality through the integration of technology and real-time customer feedback.

2.1.4 Smart Tourism Ecosystem

The concept of Smart tourism is based on the smooth integration of data, technology, and digital innovation, allowing destinations, businesses, and tourists to connect in an interactive and interconnected system. Destinations become more responsive, businesses optimize their operations, and tourists benefit from more tailored services. Xiang and Fesenmaier (2017) discussed the key components of the smart tourism ecosystem across multiple levels, which consist of consumers, businesses, and destinations. At the consumer level, smart tourism focuses on delivering intelligent, data-driven support grounded in a timely and comprehensive understanding of the tourist experience. Within this framework, data becomes the foundation for this process, offering context-rich, dynamic, and real-time insights that enable a more authentic understanding of traveler behavior (Xiang & Fesenmaier, 2017, p. 303). This consumer-driven perspective highlights the significance of real-time data collection and adaptation, ensuring that tourism services evolve in response to user behavior and preferences. The next level is considered business. At the business level, smart destinations make use of widely available open data to develop practical strategies that support both their business objectives and day-to-day operations. Finally, at the destination level, the smart tourism concept involves the transformation of physical places, such as smart cities, into technology-enabled environments where innovation drives economic development and enhances social wellbeing through tourism (Xiang & Fesenmaier, 2017, p. 304). One significant feature of smart tourism across all levels is the active role of tourists in both consuming and generating data. Tourists, as contributors of data, share experiences through social media, online reviews, and location-based services, which help businesses and destinations to adapt their services accordingly. This shift from passive consumers to active participants underlines the role of user-generated content in shaping the tourism industry. As Gretzel et al. (2015) explain, the smart tourism experience is defined by a combination of efficiency and meaningful engagement. Tourists are no longer passive recipients; rather, they actively contribute to shaping the experience by generating, tagging, and enriching data that forms its foundation (Gretzel et al., 2015).

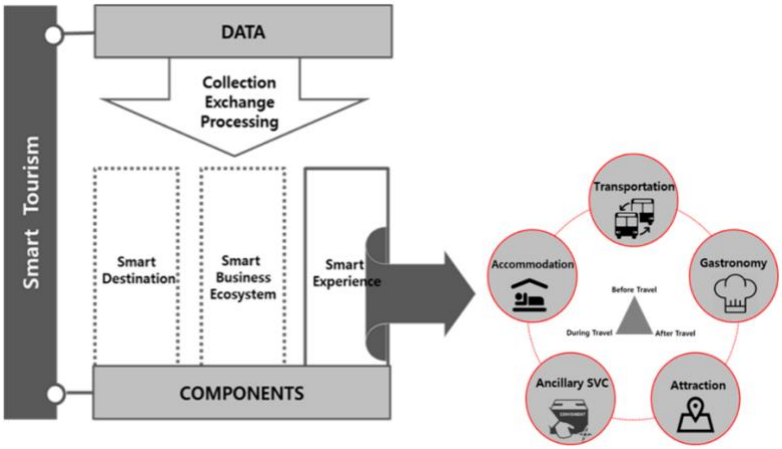

Figure 1: Smart tourism ecosystem framework. (Lee et al., 2020).

Figure 1 illustrates the smart tourism ecosystem. This framework is organized into three data-related layers, each supporting a distinct functional component: a smart information layer that focuses on data collection; a smart exchange layer that facilitates interconnectivity; and, a smart processing layer that handles data analysis, visualization, integration and effective use of data (Tu & Liu 2014, as cited in Lee et al., 2020). By segmenting the smart tourism framework into these layers, researchers and practitioners can better understand how data flows within tourism ecosystems, ultimately contributing to improved services.
Finally, smart tourism aims to enhance key aspects of the travel experience, including mobility, creativity, sustainability, resource efficiency, and overall quality of life, by relying on strategic technological investments and large-scale, coordinated initiatives (Xiang & Fesenmaier, 2017, p. 305). Reaching this objective depends on strong collaboration among governments, tourism businesses, and travelers to ensure the effective integration of technological innovations into the overall tourism experience. To put these goals into practice, Smart Tourism Destinations use several key features, including access to real-time user data, quick feedback systems to understand customer opinions, and platforms that allow different stakeholders to share data and improve services. In addition, they use historical data and patterns to predict what tourists expect, which helps create more tailored services and smart recommendation systems (Buhalis & Amaranggana, 2015).

Extracting meaningful insights from large and complex datasets requires advanced analytical techniques. Among these techniques, ML is recognized as a powerful tool within the tourism sector, particularly for user-generated content. The next section explores how ML techniques are increasingly being applied to enhance tourism services, personalize customer experiences, and support data-driven innovation.

2.2 Technological Foundations: Machine Learning

2.2.1. Machine Learning in Tourism

ML has changed the way businesses analyze and optimize customer experiences across industries, particularly in tourism and aviation. ML focuses on the collection of learning from large datasets and extracting meaningful insights without requiring direct human input (Navamani & Kannammal, 2015; as cited in Sancho Núñez et al., 2024). ML has become a significant disruptive and transformative factor in recent years. In the tourism industry. It challenges old systems by introducing new, faster, or smarter solutions, alongside transforming traditional systems by adding long-term strategic value (Sancho Núñez et al., 2024). By these advancements, tourism providers are now able to process large volumes of customer data to uncover valuable insights into traveler behavior, preferences, and trends (Sancho Núñez et al., 2024). These insights enable the creation of personalized recommendations, ranging from destinations and accommodations to activities and dining, tailored to individual interests and prior experiences. In addition, predictive models powered by ML assist in forecasting demand, adjusting pricing strategies, and improving logistical operations, which contribute to greater efficiency and increased profitability within the tourism sector (Sancho Núñez et al., 2024).

Various studies have explored how ML techniques can enhance service quality, improve customer satisfaction, and optimize operational efficiency. To better understand the application of ML in tourism, the following table summarizes key studies in this field. Table 1 provides an overview of the paper author(s), dataset used, and model, which shows how ML has been used to address different challenges and metrics for evaluating the model. This review forms a critical step in identifying current trends and potential directions for future research on the application of ML in tourism. For instance, Murugesan et al. (2024) analyzed 64,440 Skytrax reviews across 81 airlines using a wide range of models, including VADER sentiment analysis, LightGBM, Random Forest, and Neural Networks, achieving an impressive F1-score of 0.96. Similarly, Pales and Usai (2018) employed weakly supervised topic modeling and regression analysis on 74,775 online reviews, reporting a validation accuracy of 93.3% with high inter-rater reliability (Fleiss' Kappa = 0.858). Kumar and Zymbler (2019) applied CNN and other NLP methods on Twitter data from major airlines, obtaining 92.3% accuracy. In a large-scale analysis, Wang (2023) used logistic regression and factor analysis on over 129,000 customer records from 23 airlines, achieving 87.5% accuracy. For travel behavior analysis, Mendieta-Aragón and Garín-Muñoz (2023) applied logistic regression, MLP, and random forest on Spanish travel survey data, with 84.5% accuracy for the random forest model. Dimitriadou et al. (2024) used gradient boosting trees and other models on tourism data from 24 EU countries between 2010 and 2020, achieving an MAPE of 1.36% and an R^2 of 0.90. Lastly, Zhu et al. (2019) conducted semantic network analysis on over 42,000 Airbnb reviews using Leximancer, showing conceptual patterns through word frequency and semantic mapping.

These studies highlight the flexibility and efficiency of ML in using diverse datasets to provide high-accuracy insights in tourism and service quality research.

Table 1: Overview of studies, datasets, models, and evaluation metrics used in ML-based tourism research.

Author(s)	Data	Model	Evaluation
Murugesan et al. (2024)	64,440 Skytrax reviews from 81 airlines	VADER Sentiment Analysis, Logistic Regression, Naive Bayes, KNN, Decision Tree, Random Forest, Gradient Boost, Neural Network, XGBoost, LightGBM	LightGBM Accuracy: 97%, Precision: 0.97, Recall: 0.96, F1 Score: 0.96
Palese & Usai (2018)	74,775 online reviews from an Italian price comparison website	Weakly Supervised Topic Modeling (LDA), Linear Regression	Validation accuracy: 93.3% (Fleiss' Kappa = 0.858), Regression Analysis
Kumar & Zymbler (2019)	Twitter data from 146,731 tweets of major airlines globally (March 2019)	SVM, ANN, CNN, Word Embeddings (GloVe), N-gram Models, Association Rule Mining	Accuracy (CNN: 92.3%), Support, Confidence, Lift
Yunxia Zhu, Mingming Cheng, Jie Wang, Laikun Ma, Ruochen Jiang (2019)	42,085 Airbnb reviews from three U.S. cities (2016–2017)	Semantic network analysis using Leximancer software, conceptual aggregation	Word frequency, semantic dimension mapping
Yifei Wang (2023)	Airline customer data with 129,880 observations and 23 features (demographics, delays, etc.)	Logistic Regression, Factor Analysis, Comparative Analysis	Accuracy (87.5%), Statistical Significance
Mendieta- Aragón, A. & Garín-Muñoz, T.(2023)	Residents Travel Survey (RTS) of the National Statistics Institute of Spain (2016–2021), 69,752 observations for accommodation and 23,779 for transport	Logistic Regression, Multilayer Perceptron (MLP), Random Forest	Accuracy (84.5% for RF), AUC, Sensitivity, Specificity
Athanasia Dimitriadou, Periklis Gogas, & Theophilos Papadimitriou (2024)	Annual tourism data (2010–2020) for 24 EU countries with 17 key variables, including economic and political indices	Gradient Boosting Trees, Random Forest, Decision Trees, KNN, Support Vector Regression	MAPE (1.36%), RMSE, R2 (0.90 for GBT)

2.2.2. Machine Learning in Classification: Binary and Multi-Label Approaches

Binary and multi-label classification, as a ML technique, are fundamental to interpreting structured and unstructured datasets by assigning data points to specific, meaningful classes. These tasks range from simple binary decisions, yes or no answers to more classes, when multiple labels are assigned to a single data instance. Binary classification is a method used to categorize data into two opposing classes. It is useful when the goal is to separate data into two groups, like deciding if a customer will leave or stay. It is simple to use and easy to understand, which makes it popular in areas like finance, healthcare, and marketing (Zadeh et al., 2024). The classifier distinguishes between two groups by analyzing the features of each data point. Although both classification and regression are types of supervised learning, they differ in their outputs; classification deals with assigning discrete labels, whereas regression focuses on predicting continuous values (Zadeh et al., 2024). However, Zadeh et al. (2024) explain that this reliance on labeled data presents a challenge, especially in imbalanced or limited datasets (Zadeh et al., 2024).

As explained in a paper by Tidake & Sane in 2018, the rapid usage of the internet has led to generate high-volume data, which needs proper organization, such as text categorization. Over time, it became evident that many texts can simultaneously relate to multiple topics. This shift made an increased adoption of multi-label classification, a supervised learning method that assigns multiple relevant labels to a document by analyzing its features and content, more commonly implemented (Tidake & Sane, 2018). Multi-label classification goes beyond binary classification, allowing each instance to belong to multiple classes simultaneously. This is especially beneficial in tasks like ABSA, where reviews may discuss multiple aspects (e.g., "food," "service," "punctuality") and assign sentiments (positive, neutral, negative) to each. Multi-label classification is particularly useful in areas such as online review analysis, where a single post might contain multiple sentiments or topics.

These classification methods help us understand how ML can sort and predict data, specifically online reviews, which may contain different aspects. The next section looks at how similar ideas are used in NLP to work with language and text.

2.2.3. NLP and Deep Learning

NLP relies on the interaction between computers and human language. It is also known as a beneficial tool for interpreting human language through computer systems (Baral, S. 2024). Moreover, NLP is capable of implementing models, systems, and algorithms to address problems in understanding human language (Lauriola et al., 2022). Some tasks of NLP are described by Lauriola et al. (2022) as machine translation, question answering, and summarization.

In the context of tourism and service industries, NLP has a wide range of practical applications. One of them is related to referencing public opinion. Individuals often rely on the views of family and friends before making purchasing decisions, and organizations commonly use tools such as surveys and polls to gather information (Mowlaei et al., 2020). So, NLP serves as a powerful tool for businesses seeking to evaluate their performance regarding quality. It enables the analysis of customer feedback such as online reviews and survey responses to identify levels of satisfaction and prioritized areas of concern. These insights support the continuous improvement of services and products. Additionally, NLP is implemented in customer service systems to enable automation, technologies like chatbots, which provide fast, personalized assistance. NLP systems

also help prioritize and categorize complaints, enabling businesses to respond more efficiently (Mowlaei et al., 2020).
The effectiveness of NLP has significantly improved in recent years, mainly due to advances in deep learning. Deep learning models have become some of the most powerful tools in the field of NLP (Tay et al., 2020). Models such as BERT (Boukkouri et al., 2020) and GPT-3 (Dale, 2021) have made major progress in a variety of NLP tasks. These include text classification for sentiment analysis, personalized recommendation systems, such as those used in transport or tourism, and the automatic generation of content for applications like chatbots. Their ability to process language with greater accuracy and contextual understanding has expanded the scope and impact of NLP applications (Álvarez-Carmona et al., 2022). Figure 2 provides a visual representation of the relationship between AI, ML, and DL, emphasizing how each technology extends and enhances the capabilities of the other.

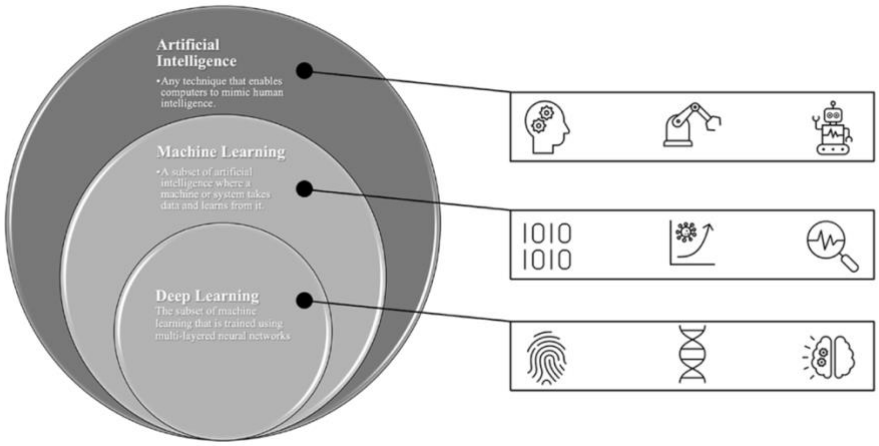

Figure 2: Positioning deep learning within the broader context of AI and ML (Essien & Chukwukelu, 2022).

2.2.4. LLM and Multi-Agent Systems

The development of LLMs is a significant step in NLP. These models show capabilities in understanding, generating, and interacting with human language. Notable examples include GPT (Generative Pre-trained Transformer) and BERT (Bidirectional Encoder Representations from Transformers), which are built with hundreds of billions, or more, parameters, and trained on vast text data (Yao et al., 2024). As Yang et al. (2023) clarify, LLMs should possess four essential characteristics. First, they must be capable of deeply understanding and interpreting natural language, allowing them to extract relevant information and carry out various tasks such as translation (Yang et al., 2023, as cited in Yao et al., 2024). Second, LLMs should be able to

generate coherent and human-like text in response to prompts, whether by completing sentences, drafting paragraphs, or producing entire articles. Third, these models need to demonstrate contextual sensitivity and domain-specific knowledge, referred to as "knowledge-intensive" capability (Yao et al., 2024), which enables them to tailor outputs based on specialized subject matter. Finally, effective LLMs should support problem-solving and decision-making by extracting and integrating information from textual data. This makes them particularly valuable for applications such as question-answering and information retrieval systems (Yao et al., 2024). As a result of these advanced features, many widely used LLMs are designed to be easily fine-tuned for domain-specific applications, including tourism (Gu, 2024).

These developments have led to the release of several well-known LLMs that are now widely used in both research and industry. OpenAI's ChatGPT, Meta AI's LLaMA, and Databricks' Dolly 2.0 are notable models that were developed and released in 2023. These models are not only technically advanced but also widely adopted in real-world settings. For instance, ChatGPT has gained over 180 million users, showing how deeply LLMs have been integrated into various domains (Yao et al., 2024). To support such wide usage, LLMs are typically designed with specific capabilities that ensure their effectiveness across tasks and industries. Another LLM which has reached increasing attention due to its optimized performance in both speed and contextual accuracy, is the Gemini series, developed by Google DeepMind. According to recent findings, Gemini models showcase impressive multimodal capabilities, enabling them to effectively process and understand input across text, image, audio, and video formats (Gemini Team, 2024). The Gemini family includes Ultra, Pro, and Nano variants, each tailored to different needs, from advanced reasoning tasks to deployment in memory-constrained environments. Among them, Gemini 1.5 has been specifically refined for enhanced speed and accuracy. What distinguishes the model is its capacity to handle complex queries while producing accurate and context-sensitive responses applicable across diverse domains (Mondillo et al., 2025). These features allow Gemini to extract meaningful aspect–sentiment relations even from unstructured data, which improves the reliability of statistical analyses.

At the same time, the GPT series, including models such as GPT-3, Codex, InstructGPT, and ChatGPT, has attracted significant attention for its advanced NLP capabilities (Ye et al., 2024). Like earlier models, GPT is trained in an unsupervised manner on large volumes of natural language text, resulting in a general-purpose language model that can be fine-tuned for specific NLP tasks. The GPT model adopts the transformer architecture introduced by Vaswani et al. (2017), distinguished by its deep attention-based layers that enhance context-aware processing (Vaswani et al., 2017, as cited in Gu, 2024). This structure enables GPT to capture long-range dependencies and contextual relationships in text, which is critical for understanding user intent and generating coherent responses. Moreover, the model has been developed in six successive versions: starting with GPT-1 (Radford et al., 2018), followed by GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), and continuing with GPT-3.5, and GPT-4 (Ouaddi et al., 2025). With each new release, improvements were achieved by increasing both the volume of training data and the model's complexity, scaling from 117 million parameters in GPT-1 to 1.5 billion in GPT-2, and reaching 175 billion in GPT-3 and later versions (Ouaddi et al., 2025). These improvements can address better language understanding, fewer hallucinations, and greater adaptability across application domains.

LLMs are also at the center of a new generation of intelligent systems known as LLM agents or multi-agent systems, supporting sophisticated user interaction and task execution across diverse domains (Yan et al., 2025). These systems use one or more LLMs as the core brain to interact with users, solve complex problems, and collaborate with other agents. This shows how AI is moving from doing simple tasks to working in more complex and smart ways. As shown in Figure 3, a typical LLM agent is made up of five key components: the core language model (LLM), a planning module, memory (both short-term and long-term), tool access, and an action module. Each component plays a distinct role in enabling the agent to understand, reason, and act within complex environments. With the core element being the brain, LLM itself is responsible for processing input, making decisions, and performing tasks involving reasoning and planning. Trained on large-scale human behavior data, LLMs enable the agent to break down complex tasks and communicate naturally through language (Yan et al., 2025). While the brain handles thinking and understanding, the agent also needs a way to break down tasks; that's where the next component comes in. The plan module is designed to break down complex tasks into a series of simpler, independently solvable steps, helping the agent address the user's request more effectively (Yan et al., 2025). By structuring problems into smaller components, this module enhances the agent's reasoning abilities, improves its grasp of the task, and increases the likelihood of producing accurate and dependable outcomes (Yan et al., 2025). However, completing tasks also depends on remembering important information, which leads to the next key component. Memory in LLM agents is generally divided into short-term and long-term categories (Yan et al., 2025). Short-term memory allows the agent to temporarily retain important information related to the current task, which leads to efficient performance (Yan et al., 2025). In contrast, long-term memory employs external storage and fast retrieval systems, enabling the agent to store and recall large volumes of information when required (Yan et al., 2025). This supports handling more complex tasks that need longer timeframes or previously acquired knowledge (Yan et al., 2025). Finally, to act on the plans and knowledge it gathers, the agent needs tools to interact with the world. LLM agents are capable of learning how to operate various tools and interfaces, enabling them to access real-time information, execute code, and retrieve proprietary data. This functionality supports more accurate and efficient task completion (Yan et al., 2025). By integrating these components, LLM agents can perform a wide range of tasks within complex environments, gradually progressing toward the capabilities associated with Artificial General Intelligence (AGI). Within this framework, the LLM serves as the core component, responsible for processing information and guiding decision-making (Yan et al., 2025).

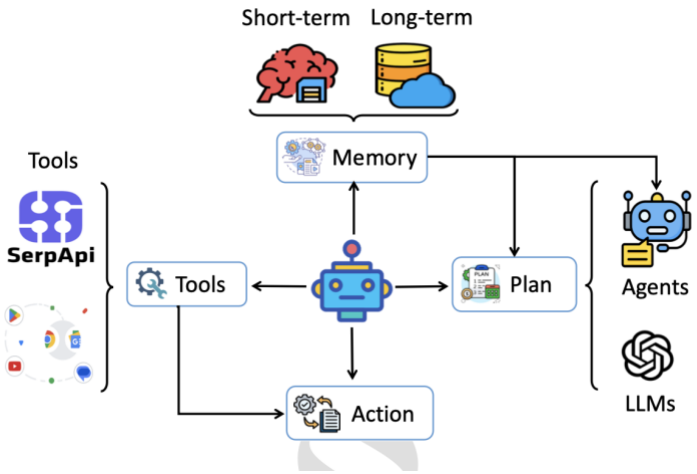

Figure 3: The structure of LLM agents (Yan et al., 2025).

In conclusion, the literature reviewed covers several key concepts relevant to this study, including service quality assessment models, the role of online reviews in capturing customer satisfaction, and the integration of advanced technologies such as NLP, LLMs, and intelligent agents in service analysis. Although these previous studies offer valuable insights, a unified theoretical perspective is still required for this research. Therefore, the next section introduces the theoretical framework that aims to align technological advancements with established service quality theories in the tourism industry.

3. Theoretical Framework

Understanding the role of theory in research is crucial for designing a structured and credible study. To clarify the role of theory in this research, it is important to first define what a theoretical framework is. According to Adom et al. (2018), the theoretical framework is a constructed and interconnected set of concepts derived from one or more established theories. It is designed to provide support for a research study. Also, its primary purpose is to enhance the relevance, credibility, and generalizability of the research findings within theoretical structures (Adom et al., 2018). Developing a theoretical framework is essential because it ensures that the data is interpreted in a clear, logical, and consistent way based on scientific reasoning (Neuman, 1997). Based on this definition, this study adopts two key theoretical perspectives to form the basis of its analytical approach: the SERVPERF model of service quality, which provides the conceptual structure for measuring service quality, and the Smart Tourism Ecosystem, which frames the role of digital technologies and user-generated content in evaluating service experiences. Together, these frameworks support the combination of traditional service quality dimensions with modern data-driven techniques such as deep learning.

Therefore, these frameworks guide this study with a strong theoretical base while also allowing for the use of modern data analysis techniques. This section will explain the theoretical perspectives in this research.

3.1. SERVPERF Model as a Foundation for Measuring Airline Service Quality

The SERVPERF model is used in this study to understand and evaluate the quality of airline services. It is a well-known framework that focuses on five key dimensions: tangibles, reliability, responsiveness, assurance, and empathy. These dimensions help explain how passengers perceive service quality based on both what they see and how they are treated. This model provides the structure for analyzing the online review dataset in this research. A deeper understanding of service quality assessment requires an exploration of its key dimensions, which serve as the foundation for the SERVPERF model. Among the five SERVPERF dimensions, tangibility presents all physical features that influence customers' perceptions of service quality. As Yu and Hyun (2019) explain, tangible elements are one of the features of service quality that provide visible and physical cues about the service's reliability and professionalism (Yu & Hyun, 2019). These components include the appearance of physical facilities, equipment, staff, and communication materials, along with other elements that enable interaction with the service (Yu & Hyun, 2019). Tangible elements help to shape how customers see the service because they give visible signs of its quality and professionalism. In the airline industry, the tangibility of service is assessed through various physical in-flight conditions, including the safety and comfort of seating, the quality and maintenance of equipment, the cleanliness of the aircraft interior, the appearance of flight attendants, as well as the availability of food, beverages, and entertainment materials provided during the flight (Yu & Hyun, 2019). When these tangible elements are poorly maintained, perceived as low quality, or fail to meet customer expectations, it is likely to result in dissatisfaction with the overall service experience (Yu & Hyun, 2019). In addition to physical features, delivering the promised services can also play a critical role in shaping customer opinion about service quality, and the dimension of reliability captures this. A reliable service is defined as the capability to consistently deliver the promised service accurately and dependably within a specific context. Reliability focuses on intangible service and has two dimensions: promises and doing it right. In the airline industry, baggage security, passenger safety, and the proper execution of emergency procedures, and so on, can be considered as reliability features (Yu & Hyun, 2019). Closely linked to reliability is responsiveness, the ability to act quickly and effectively when customers need assistance. Responsiveness is a key dimension of service quality, which involves timely reactions to customer demands and a readiness to assist when required. Studies have shown that responsiveness significantly contributes to customer satisfaction (Andaleeb & Conway, 2006) and aligns with the goal of enhancing service quality (Andersson & Mossberg, 2004). To meet this standard, personnel must deliver services promptly (Hansen, 2014) while demonstrating competence, enthusiasm, availability, and a strong sense of responsibility (Namkung & Jang, 2008). Moreover, a genuine willingness to help customers is essential in fulfilling this dimension of service quality (Hansen, 2014). Therefore, providing prompt and efficient responses ensures that customer concerns are addressed without delay, leading to overall satisfaction. While responsiveness ensures quick reaction to needs, assurance focuses on building trust and confidence in the service provider. This dimension can be defined as: "*creating trust and certainty, personnel knowledge*" (Hansen, 2014). To maintain high assurance levels, it is essential to train staff well and ensure they communicate clearly and professionally. In the airline industry, the assurance dimension is reflected in crucial qualities such as the perceived trustworthiness and competence or knowledge of service personnel, their ability to maintain a consistent and reliable experience from the digital booking phase to the actual flight, and the consistent display of polite and

professional behavior by service personnel throughout the entire customer journey (Yu & Hyun, 2019). Finally, beyond professional behavior and efficiency, the emotional connection between staff and passengers is reflected in the empathy dimension. It is defined as a personalized attention to customers, which can address individual passenger needs and offer tailored support (Yu & Hyun, 2019). Therefore, paying attention and caring for each customer can significantly enhance the sense of being valued and understood among customers.

Besides the SERVPERF model, the Smart Tourism Ecosystem is also used in this study to support the theoretical foundation. It shows how digital technologies and data systems help improve service. The following section will explain it in more detail.

3.2. Smart Tourism Ecosystem as a Digital Framework for Data-Driven Analysis

To implement the SERVPERF model, this research adopts the Smart Tourism Ecosystem framework (Gretzel et al., 2015) as an analytical lens for understanding how digital technologies and user-generated data transform tourism experiences. This ecosystem contains three linked layers (Gretzel et al., 2015); the smart information layer, which focuses on collecting data from sources such as user-generated content (UGC) and online reviews; the smart exchange layer, that links various digital platforms together; and the smart processing layer, which involves advanced analytics techniques, including deep learning and NLP. These three layers align with the structure of this study, in which online reviews are gathered as raw data, organized through content analysis into themes and sub-themes, and then further processed using an LLM to extract deeper insights (Gretzel et al., 2015). This layered approach reflects the smart tourism system, where data flows through collection, exchange, and intelligent use. This study particularly relies on the smart processing layer to apply LLMs to the online review dataset. Through this approach, service quality themes can be extracted automatically, supporting a more scalable and detailed analysis of customer feedback that goes beyond the limitations of traditional survey methods. Furthermore, the Smart Tourism Ecosystem framework supports the use of big data and AI technologies as effective tools for enhancing operational decision-making and personalizing tourism services. As described by Gretzel et al. (2015), smart tourism fundamentally depends on the capability to gather high volumes of data and to effectively store, process, integrate, analyze, and apply this data to enhance business innovation, service delivery, and operational efficiency (Gretzel et al., 2015). This ecosystem enables the transformation of the tourism experience by integrating real-time data, user-generated content, and smart infrastructure into service design and delivery. Within this framework, tourism becomes more dynamic, personalized, and responsive to the changing needs of travelers.

In addition to this data structure, tourists now contribute to the system by sharing experiences through online platforms, writing reviews, and interacting with digital services. These user-generated inputs not only reflect personal experiences but also serve as valuable data for service providers to monitor, assess, and enhance service quality. According to Gretzel et al. (2015), this feature is considered a significant feature of the smart tourism ecosystem, which highlights the active role of tourists in both consuming and producing data. This process can enhance data that enriches the basis of the experience (Gretzel et al., 2015). Therefore, this study builds on this understanding by using qualitative feedback from online reviews as a data source for measuring perceived service performance. Moreover, the smart tourism ecosystem offers a theoretical foundation for integrating the SERVPERF model into a data-driven environment. Big data

analytics is a powerful tool that involves various types of data, analysis techniques, and business applications (Xiang et al., 2015). Compared to traditional research methods, it offers a deeper and broader understanding of consumer behavior on a much larger scale in comparison with traditional models (Boyd & Crawford, 2012). The dynamic nature of smart tourism supports continuous feedback loops between users and service providers, allowing tourism services to be adjusted in near real-time. By connecting individual review content to broader service quality dimensions through automated analysis, this study bridges the gap between theoretical models and practical, technology-enabled evaluation.

The merging of SERVPERF and the Smart Tourism Ecosystem forms the core of this study's theoretical framework. SERVPERF provides the what, the dimensions of service quality to be assessed, while the Smart Tourism Ecosystem offers the how, a digital infrastructure for collecting, processing, and interpreting unstructured customer feedback. By integrating these two models, the research connects traditional service theory with modern digital methods, facilitating a more scalable assessment of airline service performance within the tourism sector. The following section will elaborate on the philosophical and methodological foundations of the study.

4. Methodology

4.1. Research Layout

This research is structured into several key phases, each contributing to the overall goal of understanding and evaluating airline service quality using modern AI techniques. The study begins with a literature review, which introduces the key theoretical concepts, namely the SERVPERF model and the Smart Tourism Ecosystem, as well as recent technological developments such as big data analytics, ML, and LLMs. These form the foundation for the research framework. Following this, a content analysis was performed on a sample of 50 airline reviews. This step helped define meaningful sub-themes and themes and organize them under the SERVPERF dimensions. The insights from this analysis were used to craft more targeted prompts for LLMs. The core of the research involves a three-step analytical process using LLMs: Gemini 1.5 Flash was first applied to identify key aspects and sentiments in the reviews, followed by a second validation and refinement step using GPT-3.5 Turbo, a cloud-based LLM developed by OpenAI. This multi-agent, prompt-based approach allowed for a scalable and structured extraction of customer feedback from unstructured text. To evaluate the performance of the models, a manual content analysis of 20 random samples was conducted, and the results were compared using the micro F1 score. This metric was selected due to class imbalance in the dataset and the multi-label nature of the task. Additionally, the results were visualized using confusion matrices for both sentiment and aspect classification. These visual tools help assess the accuracy of the model and highlight areas for improvement, particularly in underrepresented classes such as neutral sentiment or less frequent service aspects. Finally, to identify which service aspects were most influential on customer satisfaction, an OLS regression model was applied to the output of the ABSA. The regression coefficients allowed for the visualization of each dimension's importance through a bar chart, providing quantitative insights into which dimensions contributed most to positive or negative customer evaluations. The whole process is designed and illustrated in Figure 4.

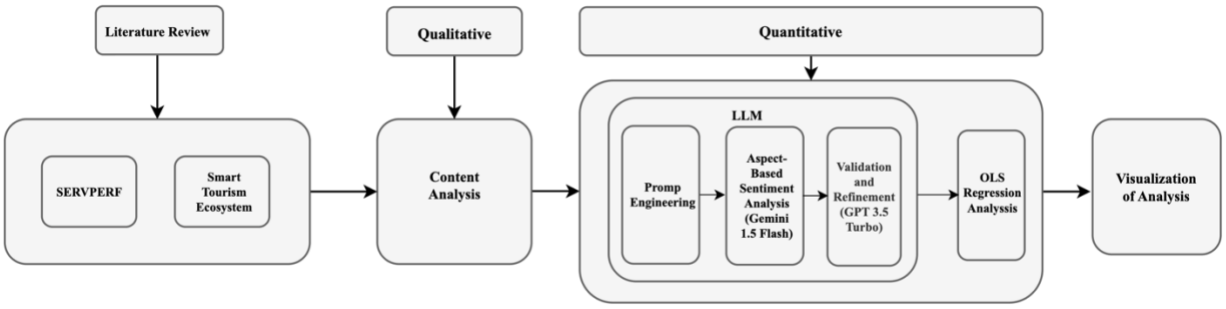

Figure 4: Research Layout designed by the researcher.

4.2. Philosophy of Science

This research aims to explore and classify service quality dimensions in airline customer experiences by analyzing online reviews through deep learning. For having a more focused and consistent research approach, defining a paradigm is essential. A research paradigm in social science reflects the researcher's core beliefs about knowledge and reality, which influence how the study is planned and conducted (Lincoln et al., 2011, as cited in Kaushik & Walsh, 2019). There is also another term for the definition of paradigm, worldview, which is described as a way of thinking about and making sense of the complexities of the real world (Patton 2002, p. 69 as cited in Kaushik & Walsh, 2019). For this research, a pragmatism paradigm was chosen, as it aligns with the study's emphasis on real-world experiences and context-dependent insights. Following Morgan's (2014) interpretation of Dewey's work, this paradigm recognizes that human actions are context-dependent and must be interpreted within the specific situations in which they are performed (Morgan 2014, as cited in Kaushik & Walsh, 2019). as the customer's evaluation of the same event changes based on how it is handled and the way it is managed. Another feature of this paradigm is explained by Morgan (2014); actions are connected to outcomes in ways that are subject to change, implying that even if the actions remain the same, shifts in context can lead to different consequences. This reflects the pragmatist view that our understanding of how to act is always short-term and shaped by current circumstances. Furthermore, actions are influenced by socially constructed worldviews, shared belief systems that guide behavior (Morgan 2014, as cited in Kaushik & Walsh, 2019). From a pragmatist perspective, no two individuals have identical life experiences; therefore, their worldviews and interpretations of actions inevitably differ (Kaushik & Walsh, 2019). Additionally, pragmatism aligns with the nature of customer experience data, where meaning depends on context, and perceptions vary from person to person. As customer reviews reflect individual experiences shaped by specific situations, changing outcomes, and diverse worldviews, pragmatism offers a suitable philosophical foundation that values practical, context-aware understanding.

From an ontological perspective, this study adopts a pragmatic realist position. Within this view, reality is assumed to exist, but our access to and understanding of it are always mediated by our actions, context, and available tools. Rather than viewing reality as entirely objective or purely constructed, pragmatism suggests that knowledge emerges from our practical engagement with the world. According to Morgan (2007), Dewey created a revised version of metaphysics that focused on the experience of actions in the world, emphasizing that reality is understood through

interaction, not independent observation (Morgan 2007, as cited in Kaushik & Walsh, 2019). In this study, service quality is seen as a real and meaningful phenomenon that can be explored through passenger reviews, but the patterns identified may vary depending on the analytic method, data type, or model structure. Therefore, deep learning models are used to detect common themes and dimensions across diverse user experiences, providing a structured but context-aware exploration of the service quality context. This ontological perspective forms the foundation for the study's approach to knowledge generation and informs the selection of appropriate methods.
Following this ontological perspective**,** the research paradigm defines the philosophical lens through which data are collected, interpreted, and used to answer research questions. This study is positioned within a pragmatist paradigm, which emphasizes the practical application of knowledge, the usefulness of results, and the integration of multiple methods to solve real-world problems. Creswell (2014) and Morgan (2007) explain that pragmatism is not limited to one specific philosophical system or view of reality. Instead, it allows researchers, especially in mixed methods studies, to freely combine both quantitative and qualitative approaches to better understand and solve research problems (Creswell, 2014; Morgan, 2007). Also, Cherryholmes (1992) explains that the role of the researcher is not to discover absolute truths, but to generate insights that are credible, useful, and applicable. Many research traditions from positivist/empiricist (quantitative) to phenomenological/interpretivist (qualitative) to versions of critical research aim at getting things right (Cherryholmes, 1992). Understanding present experiences requires attention to the events and conditions that preceded them. This aligns with the pragmatic view that meaning is shaped by prior context, which is especially relevant in interpreting customer feedback and service evaluations (Cherryholmes, 1992). With the exception of critical research, these traditions in different ways maintain those descriptions, theories, and explanations precede values, social policy, and educational practice. Within the pragmatic tradition, research focuses on understanding meaning through its outcomes. Pragmatists maintain that values and perspectives on human behavior and interaction come before the development of theoretical explanations, descriptions, or narratives (Cherryholmes, 1992).

The present study adopts a practical approach to investigating airline service quality. The primary objective of this study is to extract and categorize key dimensions of airline service quality based on large-scale user-generated content through deep learning. This reflects the pragmatist commitment to using the most effective tools and approaches to generate actionable knowledge.
The study combines deductive reasoning by applying existing theories like SERVPREF with inductive reasoning through content analysis, which helps refine the AI model based on emerging patterns. According to Morgan (2016), "*The pragmatic approach is to rely on a version of abductive reasoning that moves back and forth between induction and deduction—first converting observations into theories and then assessing those theories through action.*" (Morgan, 2016, p. 71). Also, Morgan (2007, p. 68) notes, *"The anomalies associated with the essential role that research questions rather than metaphysical assumptions play is little more than a restatement of the pragmatist approach itself."* (Morgan, 2016, p. 67). In this study, the use of AI and NLP is not driven by philosophical loyalty to any one paradigm but by their suitability to uncover meaningful patterns in large, unstructured datasets. Therefore, the overall reasoning in this study is abductive, as it combines deductive application of established theories with inductive insights from data, in line with the pragmatist paradigm that values practical problem-solving and methodological flexibility (Morgan, 2016).

This study follows a mixed methods approach in line with the pragmatist paradigm. In pragmatic research, the main focus is placed on the research problem itself, and researchers are encouraged to use any suitable methods, qualitative, quantitative, or mixed, to gain a comprehensive understanding of the issue (Cherryholmes, 1992). Using different methods together makes the results both reliable and meaningful. This research combines both quantitative and qualitative aspects in the research design. While the main method is based on data analysis using computational tools, content analysis was used to fine-tune the agent model and guide it in identifying meaningful service quality themes and sub-themes. Theoretical frameworks also support the interpretation of results. This combination helps provide a fuller understanding of the research problem. The following section outlines the data used in this study, including its sources and characteristics.

4.3. Data

In research, data collection methods are generally classified into two main types: primary and secondary data. Primary data refers to information collected directly by the researcher for the first time, whereas secondary data consists of existing information that has been previously gathered or generated by other sources (Ajayi, 2023). In this study, the "Skytrax Airline Reviews" dataset was used as a secondary data source. This dataset was sourced from the Kaggle website, a recognized platform for data science competitions and public datasets. However, this dataset was originally obtained from the Skytrax website. As explained on the Skytrax website, online reviews are verified by checking the reviewer's e-ticket, booking confirmation, or boarding pass, ensuring that the name and flight route match the information provided in the review (Skytrax website). This verification process increases the credibility of the dataset, making it a reliable source for academic research. Accordingly, this dataset has also been used in previous research for sentiment analysis and customer satisfaction prediction in the airline industry, notably in *"Sentiment analysis model for Airline customers' feedback using deep learning techniques"* (Samir, Abd-Elmegid & Marie, 2023) and *"Forecasting airline passengers' satisfaction based on sentiments and ratings: An application of VADER and machine learning techniques"* (Murugesan et al., 2024). The present study uses the same dataset, which includes 65,947 rows, with reviews from the years 2006 to 2019 and detailed customer reviews with ratings for 81 airlines. The dataset provides a rich set of attributes capturing various aspects of passenger experiences, such as airline ratings, cabin types, routes, and individual service ratings; however, for this study, only three columns were used.

- Airline: Name of the airline reviewed, with 81 unique airlines.
- Customer review: Textual review content.
- Overall: ranking from 1 to 10.

From the full Skytrax dataset of 81 airlines, a subset of 11 European airlines with a total of 9761 reviews was selected for in-depth analysis. This decision was made based on two main criteria: Data availability, as these airlines had a sufficient number of verified customer reviews to support reliable sentiment and content analysis. The other reason is regarding regional focus, allowing for a culturally and regulatorily consistent comparison of service quality across multiple carriers operating under similar market conditions. This focused approach improves the validity of the analysis while also creating a foundation for future studies to expand into other regions, such as Asia or North America, for cross-regional comparisons.

To prepare the dataset for analysis and modeling, a series of preprocessing steps was applied. First, selected columns were saved to a new file. Then, rows with missing values were handled carefully to ensure data consistency. The third step was deleting duplicate rows to avoid introducing bias or inaccuracies. Also, the reviews that were marked as not verified were deleted, and only verified reviews are used in this research. Next, the text in the "customer_review" column was passed through preprocessing to enhance consistency and remove noise for the NLP tasks. This process began by converting all characters to lowercase, ensuring uniformity across the dataset. URLs and HTML tags such as (r'http\S+', '', text) and (r'<.*?>', '', text) were then removed using regular expressions to eliminate irrelevant web-based content. Specific unwanted phrases such as "âœ… Trip Verified" were identified and excluded from the text, as they did not contribute meaningful information for sentiment or service quality analysis. Additionally, non-alphanumeric characters were excluded, and excessive whitespace such as [^a-zA-Z0-9\s] was deleted to create clean, readable strings. This refined version of the review text was then stored in a separate column for further tokenization and feature extraction processes in the analytical pipeline. The following section describes the methods employed to process, analyze, and interpret the dataset in line with the research objectives.

4.4. Method

4.4.1. Content Analysis

In this research, a content analysis method was employed on a sample of reviews to provide a more accurate prompt to the model. As noted by Stepchenkova et al. (2009), Content analysis is a research method commonly employed to examine diverse forms of textual data, such as media content, interview transcripts, online forum discussions, or travel journals without influencing the source of the data (Stepchenkova et al., 2009). In other words, content analysis is a technique that helps the researcher to describe what is communicated on a particular topic in a specific context with the highest possible level of objectivity, accuracy, and generalizability. So, the findings can be trusted. (Stepchenkova et al., 2009). To achieve this, researchers often organize data into themes and categories, which provide structure and clarity. In research, a theme is a common idea that links different categories and gives meaning to repeated topics or experiences. It is also defined as a red thread. Morse (2008) explains that a theme is like a central idea that appears throughout the data, sometimes clearly, sometimes more quietly, like a repeating tune in music. In qualitative content analysis, results are usually shown as categories and/or themes (Morse, 2008, as cited in Graneheim et al., 2017). According to Elo and Kyngäs (2008), the goal of qualitative content analysis is to produce a concise and comprehensive understanding of a phenomenon, where the results are expressed through concepts or categories that describe it (Elo & Kyngäs, 2008). In addition, effective content analysis depends on the researcher's ability to interpret and simplify the data, creating categories that accurately and reliably represent the topic being studied (Elo & Kyngäs, 2008).

In this research, for identifying sub-themes and themes of reviews in a systematic way, a qualitative content analysis was applied to a sample of 50 online airline reviews. The decision to extract sub-themes and themes was intentional, as the SERVPERF dimensions are relatively broad; adding sub-levels made the interpretation easier to perceive. The process began with an inductive

approach, where no predefined categories were used. First, general themes were identified directly from the reviews, and then these were further developed into more specific sub-themes, which were captured in detailed aspects of passengers' experiences. Once the sub-themes were established, they were organized under the broader dimensions of the SERVPERF model. This step added a theoretical lens to the analysis and helped to structure the findings. Although the initial coding was purely data-driven, the final categorization reflects an abductive reasoning process; the appearing patterns from the data were interpreted in an existing theoretical framework (SERVPERF), without modifying the theory itself. This abductive process made it possible to link the findings from the data with the SERVPERF framework, which helped improve both the structure and clarity of the analysis. After identifying the main themes and sub-themes, additional reviews were examined to check if any new insights would appear. Finally, an additional 30 further reviews were added; however, no new themes were added. As a result, the first 50 reviews were sufficient, and the data was considered saturated.

The analysis started by reading each review carefully and identifying specific points mentioned by passengers. These points were first grouped into sub-themes, such as baggage issues, unclear pricing, or staff attitude. Then, similar sub-themes were combined into broader themes, like service failure or communication problems. Finally, each theme was matched with one of the five SERVPERF dimensions: tangibles, reliability, responsiveness, assurance, or empathy. This step-by-step approach helped organize the data clearly and allowed for a better understanding of how different passenger experiences relate to service quality. The content analysis table shows the whole process in Table 2. Following the content analysis, the next step involved using an agentic system to conduct the main data analysis. This will be described in more detail in the following section.

Table 2: Content analysis of 50 online reviews

Review	Sub-theme	Theme	Dimensions
R4; R45; R44; R15; R27; R42; R43; R44	Inflight Entertainment options	Inflight entertainment	**Tangibles**
R4; R45; R16; R22; R27; R43	Easy-to-use interface Inflight entertainment		
R14; R15; R17; R19; R30; R33; R40; R41; R42; R43	Comfort of seats	Aircraft condition	
R1; R14; R17; R33; R4	Leg room		
R17; R15; R20	Lavatory		
R20; R31	Air conditioning		
R3; R6; R15; R23; R25; R27; R29; R40; R42; R43; R44	Food and beverage quality	Food and beverage service	
R38; R15; R19	Serving cold and warm		
R8; R46; R45; R42; R38; R23; R19; R13; R34; R35	Food and beverage options		
R3; R13; R27; R49; R1	Organization of airport environment	Airport experience	
R27; R42; R46	Airport Wi-Fi		
R2; R13; R26	Service delivery failure regarding paid or promised benefits	Service promise fulfillment, accountability, and recovery	**Reliability**
R9; R48; R11; R12; R34	Responsibility for service failure		
R2; R22; R12; R50	Booking management	Operation	
R3; R4; R6; R7; R8; R11; R16; R17; R18; R19; R20; R22; R28; R31; R36; R42; R45; R49; R50	Punctuality		
R6; R9; R12; R16; R20; R14; R24	Baggage handling		

R1; R11; R21; R22; R30; R35	Seat allocation		
R11; R14; R16; R17; R18; R25; R29; R36; R38; R40; R45; R46	Boarding process		
R19	Water leakage issue	Maintenance/Condition	
R47; R12	Overhead luggage storage		
R8; R5; R34	Compensation policy and consistent messaging	Policy and policy transparency	
R10; R47; R13; R25; R29	Cost structure		
R1; R49; R36;	Staff communication language barrier	Staff competence and professionalism	**Assurance**
R1; R6; R16; R4; R12; R14; R31; R46; R3; R34	Staff behavior and training		
R3; R 49; R7; R8; R18; R30; R24; R31; R39; R50	Clear and consistent information	Communication clarity and accuracy	
R11; R12; R50	Accuracy of provided information		
R18; R9; R10; R19; R34; R38; R39; R2; R21	Quick resolution of customer requests	Prompt service delivery and response	**Responsiveness**
R 24; R27; R11; R12; R31; R8; R10; R1	Speed of online customer support and updates		
R38; R39; R11	Ease of finding staff when needed	Staff availability and accessibility	
R39; R37; R27	Staff presence in key areas		
R48; R31; R4; R35; R2	Individualized service for special needs	Personalized attention	**Empathy**
R25; R11	Flexibility in service delivery		
R50; R22; R30; R8; R23; R28; R40; R43; R1	Attentiveness and warm and friendly behavior	Emotional support and sensitivity	
R45; R17; R23	Patience and tolerance toward customers		

In addition, a sample of review number 6 is presented in Table 3. Each part of the review is categorized into a sub-theme, an overall theme, and one of the SERVPERF dimensions.

" istanbul to budapest via dublin delays on each flight for both legs of the journey stuck in istanbul for almost 20 hours with ground staff who were rude slow or unhelpful and took an hour to get our bags the only only thing positive i can say is that the food on the plane is good everything else was disastrous"

Table 3: Content analysis of review number 6

Review	Sub-theme	Theme	Dimension
"istanbul to budapest via dublin delays on each flight…"	Punctuality	Operation	Reliability
"…ground staff who were rude slow or unhelpful…"	Staff behavior and training	Staff competence and professionalism	Assurance
"…and took an hour to get our bags…"	Baggage handling	Operation	Reliability
"…that the food on the plane is good…"	Food and beverage quality	Food and beverage service	Tangible

4.4.2. Large Language Models

After organizing samples in sub-themes and themes, the LLM technique is applied to automatically and efficiently analyze the total user reviews about airline service quality. This study used two advanced LLMs: Gemini 1.5 Flash and GPT-3.5 Turbo, each selected for their distinct strengths in processing user-generated content and facilitating automated content analysis. Gemini 1.5 Flash is Google's lightweight and fast AI model that can process text, images, and audio. Using optimization techniques, it delivers high speed while maintaining good accuracy across various tasks. Another LLM used in this research is GPT-3.5 Turbo**.** GPT-3 possesses linguistic competence and is capable of identifying semantic meaning across a wide range of continuous language contexts (Ye et al., 2023). In addition, Yang et al. (2023) and Hendy et al. (2023) explored the capabilities of ChatGPT, specifically the gpt-3.5-turbo model, in performing aspect-based text summarization and machine translation tasks (Yang et al., 2023; Hendy et al., 2023). The GPT-3 model series, with its 175 billion parameters, is recognized as a highly advanced tool for generating human-like text (Ye et al., 2023). GPT-3.5 Turbo is a refined version of GPT-3, designed to provide similar functionality with increased efficiency and lower cost (Campesato, 2024). It offers high-quality performance while being more practical for diverse use cases

(Campesato, 2024). In addition, LLMs can be customized and fine-tuned to meet the specific needs of different domains, including tourism and customer service. Therefore, for implementing these LLMs effectively, prompt engineering can be essential in guiding the models toward producing structured and contextually relevant outputs. Prompt engineering is the practice of designing effective textual inputs to guide AI systems toward producing outputs that better align with user goals. Similar to how a coach provides guidance to enhance performance, prompts help direct the behavior of language models (Campesato, 2024). The way a prompt is structured can greatly impact the accuracy, relevance, and overall quality of the response. This process ensures that the AI's output closely reflects the intended purpose of the user (Campesato, 2024). Such control over model behavior through prompts is especially valuable in scenarios where annotated training data is inadequate. Prompt-based classification has been widely used in zero-shot and few-shot learning tasks (Mao et al., 2023). Unlike traditional methods that require large labeled datasets and fine-tuning, prompt-based approaches reduce the need for manual data annotation. Instead of changing the model to fit a new task, prompting changes the input so that the task matches what the language model already knows. This method is especially helpful in situations with limited data, such as metaphor interpretation, text classification, and natural language inference (Mao et al., 2023).

First, the Gemini 1.5 flash was used to analyze airline reviews. This model found the main topics and service-related points in the text, and gave an initial summary of what customers said. Then, the results from Gemini 1.5 flash were sent to GPT-3.5 Turbo, which looked at the same reviews again to check, improve, and add more details. GPT-3.5 Turbo helped make the output more accurate by understanding context better and organizing the themes more clearly into the SERVPERF categories. Actually, this model is boosting the result, where the second model improves on what the first model analyzed. By using Gemini 1.5 flash for fast general analysis, and GPT-3.5 Turbo for deeper understanding, this study produced more reliable and complete results about how customers feel about airline service. Details of each prompt are available in Appendix A.
The structured sentiment data extracted from the two-stage LLM analysis provided a rich and context-aware representation of customer feedback. This processed output was then transformed into a format suitable for regression analysis, enabling the evaluation of which specific service aspects most strongly influence overall passenger satisfaction.

4.4.3. OLS Regression Analysis

To analyze which service dimensions have the strongest impact on overall passenger satisfaction, this study applied OLS regression as a quantitative statistical technique. OLS regression estimates the linear relationship between a dependent variable (in this case, overall satisfaction rating) and multiple independent variables, which correspond to the frequency or intensity of mentions of service aspects extracted from customer reviews. This method was chosen for its interpretability and widespread use in social science research when modeling linear relationships between variables. According to Wooldridge (2013), multiple regression analysis is particularly suitable for ceteris paribus analysis, as it enables researchers to control for several influencing variables at once when examining the effect of one independent variable on the dependent variable. It remains one of the most commonly used methods for empirical research in economics and other social sciences (Wooldridge, 2013). Additionally, the OLS method is widely applied for estimating the

parameters within multiple regression models (Wooldridge, 2013). In regression analysis, coefficients and P-values are essential outputs, as they indicate the statistical significance of the independent variables and explain their relationship with the dependent variable. A low P-value (typically below 0.05) suggests that an independent variable has a statistically significant effect (Frost, 2019). Since regression is a type of inferential statistical method, P-values help assess whether the observed relationships in the sample are likely to be present in the broader population (Frost, 2019).

In the context of this research, the goal was to evaluate the relative importance of each SERVPERF service dimension (Tangibles, Reliability, Responsiveness, Assurance, and Empathy) and their corresponding themes on customer satisfaction. After extracting and labeling review data using LLMs, each review was coded with sentiment (positive, neutral, negative) for each detected aspect. The aspect–sentiment data were organized into a table that could be used for regression analysis. In this table, each service aspect (like staff, food, or punctuality) was used as an independent variable. For each review, values were assigned as +1 if the aspect was mentioned positively, -1 if it was mentioned negatively, and 0 if it was neutral or not mentioned at all. In this way, each review became a row in the dataset, and each column represented one service aspect. The coded data were then aggregated and used as predictors in the regression model. The overall satisfaction rating from the dataset is used as the dependent variable. The resulting regression coefficients (β-values) indicate the strength and direction of influence each service aspect has on satisfaction, with positive coefficients representing a positive contribution to satisfaction and negative coefficients indicating dissatisfaction. Furthermore, P-values were calculated to assess statistical significance; predictors with P-values below 0.05 were considered. This approach enables a data-driven evaluation of which service quality features are most influential and should be prioritized by airline management.

4.4.4. Evaluation Metrics

Evaluation of the model is an essential key in the model, as it shows how accurate our model is for future usage, and it provides insight into the model's potential performance in real-world applications. When choosing an evaluation metric, paying attention to the dataset is important. This dataset is an imbalanced dataset, which means that some categories in the dataset have significantly more examples than others. This is a common issue in tasks like aspect detection and sentiment analysis, where certain aspects or sentiments may appear more frequently in the data. For example, in the airline reviews, aspects such as "Food and Beverage Service" and "Operation" are mentioned more often than aspects like "Inflight Entertainment" or "Maintenance/Condition". This imbalanced data can make it challenging for the model to learn effectively. If not addressed, the model might become biased toward predicting the majority classes, which leads to a poor performance on the less frequent categories. So, selecting an appropriate evaluation metric is crucial. In multi-label classification, the F1 score can be computed using macro or micro averaging methods (Baral, 2024). While the macro F1 score gives equal importance to each class and is typically suited for balanced datasets, the micro F1 score aggregates the performance across all classes, making it more appropriate for imbalanced data. (Baral, 2024).
To address the imbalanced dataset challenge, the micro F1 score was used as an evaluation metric. This metric collects the performance across all categories, ensuring that both frequent and rare classes contribute equally to the overall evaluation. The Micro F1 score is a commonly used metric

for evaluating model performance in classification tasks. It combines precision, which shows how many of the predicted results are correct, and recall, which shows how many of the actual correct results are found. The Micro F1 score gives a single number that balances these two measures, making it easier to understand overall performance. This score ensures that mistakes, such as missing an important category or falsely predicting one, are treated equally. That is why this metric was chosen for evaluating our model. For this goal, 20 random samples were chosen to define aspects and sentiments manually, and then the scikit-learn metrics library was used to evaluate the micro F1 score.

In conclusion, this chapter explained how the study was designed and conducted, using AI tools and statistical techniques to analyze customer reviews. In the next chapter, the results of these analyses will be presented and discussed, showing how service quality is measured and what service aspects most influence passenger satisfaction in different airlines.

5. Analysis

This section presents the findings of the study and addresses the two main research questions: *How can airline service quality be measured on a large-scale dataset?* And *which service quality dimensions most significantly influence customer satisfaction across different airlines based on online reviews?* To answer the first research question, this study applied a mixed-methods framework combining content analysis and ML, particularly a multi-agent system using LLMs. The measurement was based on extracting and categorizing aspects of service from user-generated content from 11 European airline reviews with a total of 9761, and linking them with established service quality dimensions, specifically extracted from the SERVPERF model. This approach allowed for a structured, scalable analysis of airline service quality that goes beyond traditional surveys. By mapping the extracted aspects into predefined categories and performing sentiment analysis, it was possible to quantify customer satisfaction and dissatisfaction in a way that reflects real customer experiences. Moreover, to address the second research question, the study conducted an OLS regression analysis using binary variables for extracted aspects against the overall satisfaction score. The results show that several aspects significantly affect overall passenger satisfaction, both positively and negatively. This method enabled a structured, scalable way to quantify and interpret service quality directly from unstructured customer reviews.
To provide a more focused understanding of how this approach works in practice, the following sections present a detailed case analysis of Lufthansa Airlines and British Airways. Then, a comprehensive analysis is provided. Additionally, the detailed results for the remaining airlines are presented in Appendices B to J, and the number of customer reviews per airline is summarized in Appendix K.

5.1. Service quality in Lufthansa Airlines

Service quality in Lufthansa Airlines was assessed by combining sentiment analysis and regression modeling based on the SERVPERF dimensions. A total of 1,354 customer review mentions were categorized into sub-themes of the service quality model and then into five dimensions: Tangibles, Reliability, Assurance, Responsiveness, and Empathy. Positive, neutral, and negative sentiments

were quantified for each dimension, and OLS regression was used to evaluate the statistical significance of each dimension's influence on overall customer satisfaction.

In the case of Lufthansa Airlines, the sentiment distribution across service aspects shows clear patterns in passenger perceptions, as presented in Figure 5 and Table 4. The highest number of positive comments was aligned with Staff Competence and Professionalism (n = 783), 'Food and Beverage Service' (n = 555), and 'Aircraft Condition' (n = 520), all of which fall under the Assurance and Tangibles dimensions. These numbers reflect Lufthansa's strong performance in delivering high-quality service through professional staff, well-maintained aircraft, and amenities. On the other side, 'Staff Availability and Accessibility' (n = 78 negative; 10 positive) shows significantly higher negative sentiment, suggesting considerable challenges in the Responsiveness dimension. Similarly, 'Policy and Policy Transparency' also performed poorly (n = 71) and only had minimal positive mentions (n = 2), indicating customer dissatisfaction with refund rules, rebooking procedures, and overall transparency. Moreover, 'Operation' (n = 435 negative; 421 positive) and 'Service Promise Fulfillment, Accountability, and Recovery' (n = 411 negative; 270 positive) received a mix of both positive and negative feedback. These findings reflect inconsistent operational standards and suggest that expectations related to Reliability may not have been adequately fulfilled. Although less frequently mentioned aspects like 'Emotional Support and Sensitivity' (n = 18 positive; 17 negative) and 'Personalized Attention' (n = 75 positive; 55 negative) show that some passengers noticed and appreciated these elements, however, the experience was not consistent for everyone. This suggests that Lufthansa is offering some level of personal care, but there is still room for improvement.

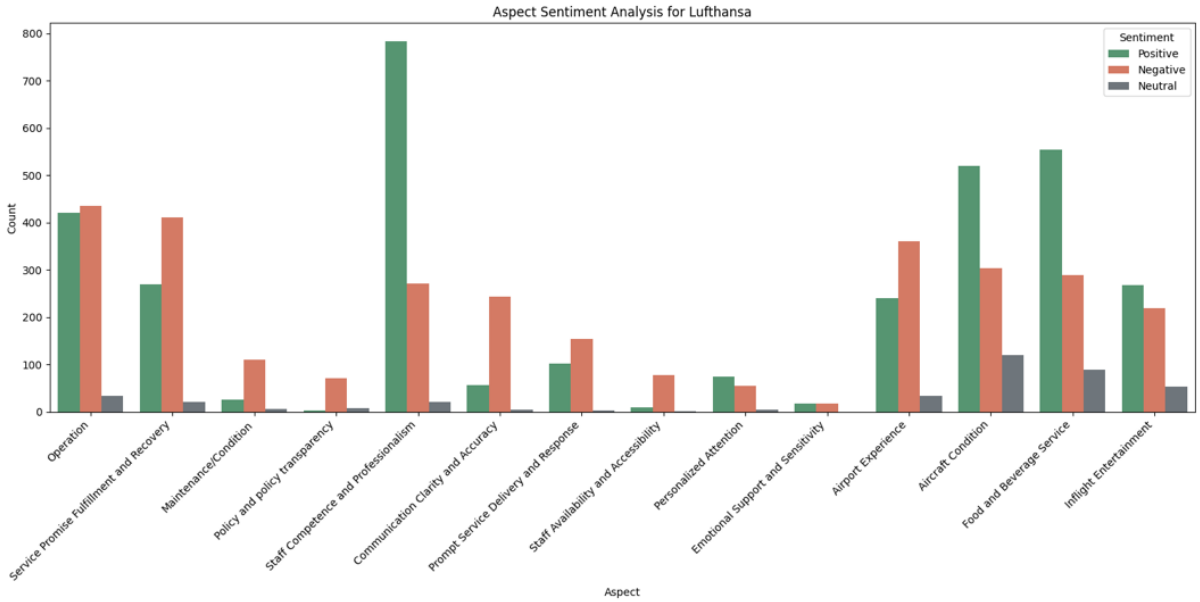

Figure 5: ABSA of service quality themes in Lufthansa Airlines.

Table 4: ABSA of service quality themes and dimensions in Lufthansa Airlines.

Dimension	Themes	Negative	Neutral	Positive
Tangibles	Aircraft Condition	303	120	520
	Food and Beverage Service	289	89	555
	Airport Experience	361	33	241
	Inflight Entertainment	219	54	268
Reliability	Maintenance/Condition	110	6	25
	Policy and policy transparency	71	7	2
	Operation	435	34	421
	Service Promise Fulfillment, Accountability and Recovery	411	20	270
Assurance	Staff Competence and Professionalism	272	21	783
	Communication Clarity and Accuracy	243	5	56
Responsiveness	Prompt Service Delivery and Response	154	3	102
	Staff Availability and Accessibility	78	1	10
Empathy	Personalized Attention	55	5	75
	Emotional Support and Sensitivity	17	0	18

While the previous analysis showed how customers feel about different service dimensions, it is also important to understand which of these aspects has the most significant effect on overall satisfaction. Therefore, OLS regression analysis was used to measure how strongly each theme influences the overall rating. Figure 6 and Table 5 summarize the influence of different service themes on overall passenger satisfaction, based on an OLS regression model. The column Coef. (β) shows how strongly each theme affects the overall rating; a positive coefficient indicates a positive impact, while a negative coefficient shows a negative effect. The column P-value shows the significance level, and values below 0.05 are statistically significant and should be interpreted with confidence. The results for Lufthansa Airline show that 'Food and Beverage Service' (β = 1.31, P = 0), 'Emotional Support and Sensitivity' (β = 1.28, P = 0), and 'Staff Competence and Professionalism' (β = 1.14, P = 0) have the strongest positive influence on satisfaction. These themes belong to the SERVPERF dimensions of Tangibles, Empathy, and Assurance, respectively, confirming that both physical comfort and human interaction are essential features of customer experience in Lufthansa Airlines.

On the negative side, 'Communication Clarity and Accuracy' (β = -1.71, P = 0) and 'Staff Availability and Accessibility' (β = -1.39, P = 0) had the most negative impact. These fall under

the Responsiveness and Assurance dimensions, indicating that customers are particularly dissatisfied when information is unclear or staff are unavailable, both critical failures in real-time service delivery. Additionally, 'Policy and Policy Transparency' (β = -0.90, P = 0) and 'Maintenance/Condition' (β = -0.59, P = 0.01) show significant negative influence, suggesting that policy transparency and maintenance standards are areas needing improvement. Aspects such as 'Inflight Entertainment', 'Airport Experience', and 'Prompt Service Delivery' had weaker coefficients or were not statistically significant ($p > 0.05$), which means their impact on satisfaction is either negligible or not clearly supported by the current data.

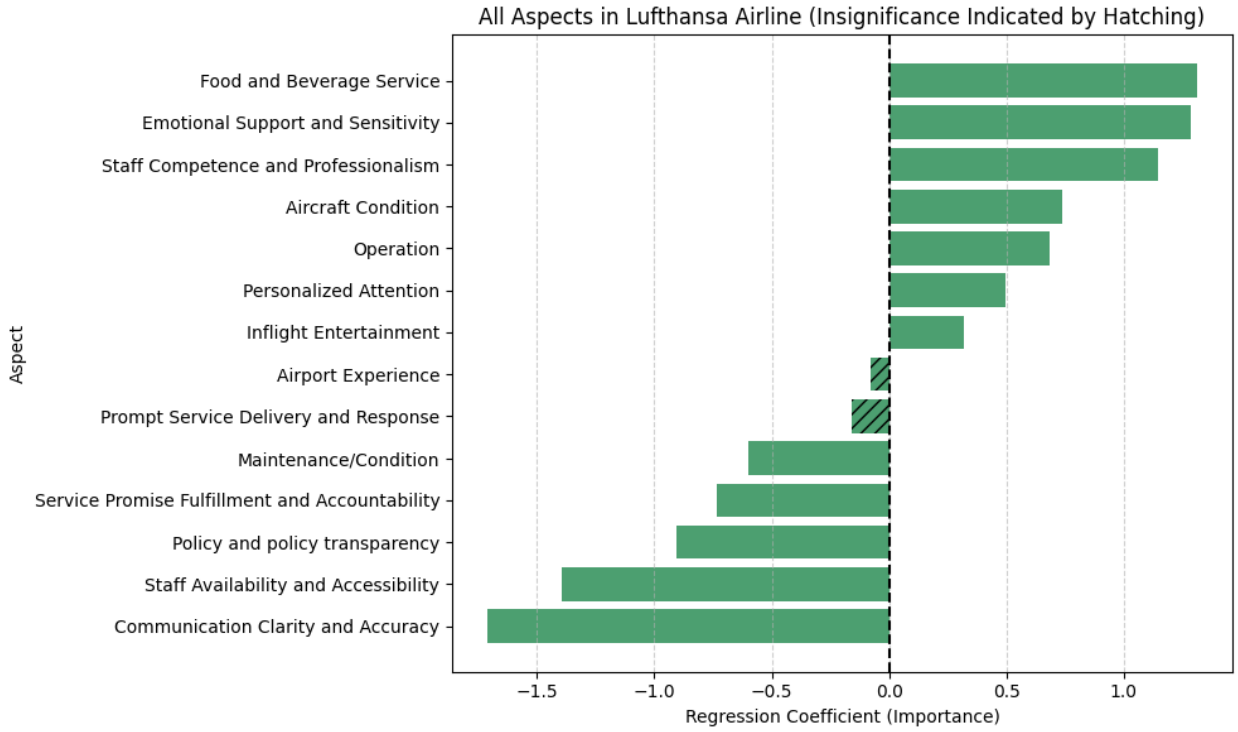

Figure 6: OLS regression analysis of all service themes in Lufthansa Airlines with indication of statistical insignificance.

Table 5: OLS regression analysis of service quality themes and dimensions in Lufthansa Airlines.

Dimension	Themes	Coefficient	P-value
Tangibles	Airport Experience	-0.07	0.58
	Aircraft Condition	0.73	0*
	Food and Beverage Service	1.31	0*
	Inflight Entertainment	0.31	0.03*
Reliability	Service Promise Fulfillment and Accountability	-0.73	0*
	Operation	0.68	0*
	Maintenance/Condition	-0.59	0.01*
	Policy and policy transparency	-0.90	0*
Assurance	Staff Competence and Professionalism	1.14	0*
	Communication Clarity and Accuracy	-1.71	0*
Responsiveness	Prompt Service Delivery and Response	-0.15	0.41
	Staff Availability and Accessibility	-1.39	0*
Empathy	Personalized Attention	0.49	0.03*
	Emotional Support and Sensitivity	1.28	0*

Note: * Indicates statistical significance at $p < 0.05$

Taken together, the combined findings from sentiment distribution and regression analysis provide a comprehensive understanding of how different service dimensions influence passenger satisfaction, as demonstrated in Table 6. In this section, only the statistically significant dimensions are discussed ($p > 0.05$). For instance, the theme Reliability received a high number of negative mentions (n = 1,027 out of 1,812 total), and also showed a statistically significant negative coefficient in the regression model ($\beta = –0.38$, P = 0). This combination of high negative sentiment and statistical significance indicates a critical service weakness that directly contributes to customer dissatisfaction. Conversely, Empathy was mentioned far less frequently (n = 170 in total), however, its regression coefficient was the highest among all dimensions ($\beta = +0.89$, P = 0.02), suggesting that although it is less discussed, it can play a highly influential role in shaping satisfaction. This reflects a hidden opportunity; the airline may considerably improve perceived service quality by enhancing empathetic interactions, such as emotional support and personalized attention. Furthermore, sentiment analysis can also help identify interpretive contradictions. For example, the Assurance dimension received mostly positive mentions (n = 783 positive out of 1,076); however, it had a significant negative regression coefficient ($\beta = –0.28$, P = 0). This discrepancy, where Assurance is frequently mentioned positively but has a significant negative impact on satisfaction, suggests that the negative instances of Assurance might carry more weight in shaping the overall satisfaction score. It is also possible that positive mentions of Assurance are

not strong enough to compensate for negative evaluations in other critical dimensions. Thus, integrating sentiment volume with regression results enables a more comprehensive understanding of which service dimensions matter most, which need urgent attention, and which may offer hidden potential for service innovation.

To visualize the overall customer perception of each service quality dimension in the SERVPERF model (Tangibles, Reliability, Assurance, Responsiveness, Empathy), two bar charts are provided. Figure 7 shows the average sentiment for each dimension, calculated by subtracting the number of negative reviews from positive ones and dividing by the total. This metric reflects the average emotional tone associated with each dimension. In Figure 8, dimensions with P-values below 0.05 were considered statistically significant, and those that were visualized without hatching are insignificant.

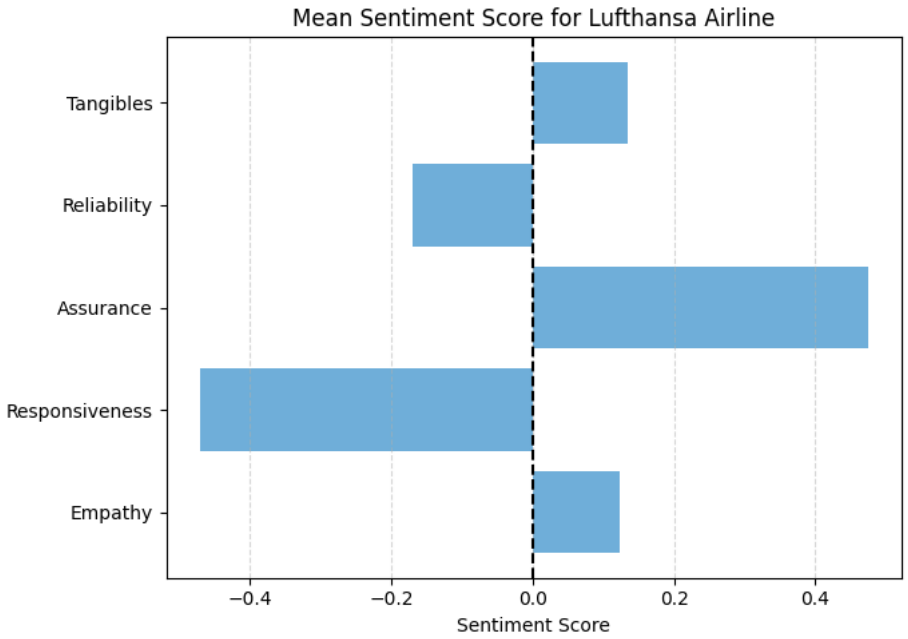

Figure 7: Mean sentiment scores of service quality dimensions in Lufthansa Airlines.

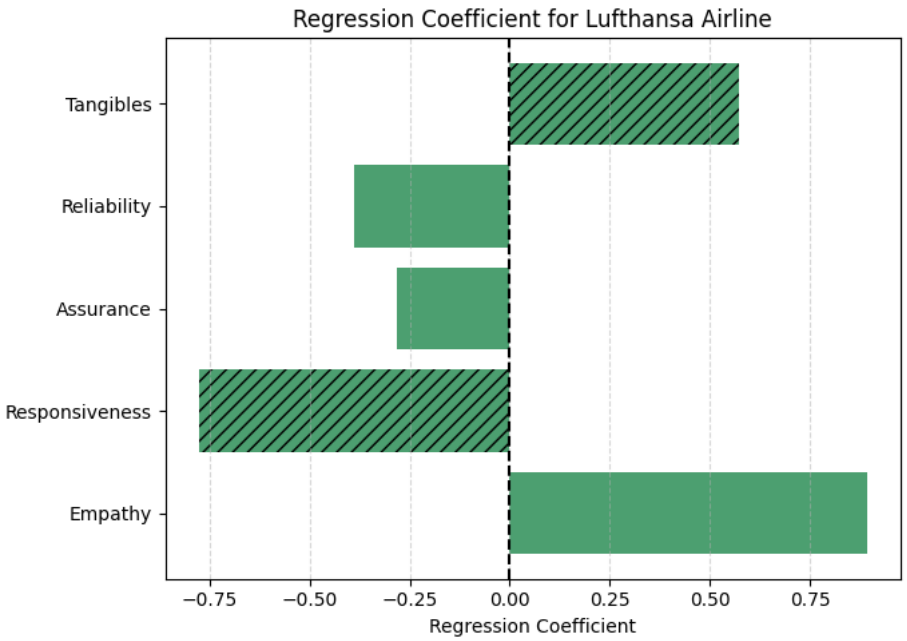

Figure 8: OLS regression analysis of service quality dimensions in Lufthansa Airlines.

Table 6: Summary of sentiment and OLS regression analysis for service quality dimensions in Lufthansa Airlines.

Dimension	Negative	Neutral	Positive	Total	Coefficient	P-value
Tangibles	1172	296	1584	3052	0.57	0.15
Reliability	1027	67	718	1812	-0.38	0*
Assurance	272	21	783	1076	-0.28	0*
Responsiveness	475	9	168	652	-0.77	0.20
Empathy	72	5	93	170	0.89	0.02*

Note: * Indicates statistical significance at $p < 0.05$

5.2. Service quality in British Airways

The second airline for a detailed analysis is British. A total of 1,620 customer reviews from British Airways were classified into five core service quality dimensions: Tangibles, Reliability, Assurance, Responsiveness, and Empathy. Sentiment distribution (positive, neutral, negative) was quantified for each aspect, and OLS regression was employed to assess the statistical impact of each dimension on overall customer satisfaction. The sentiment analysis highlights clear trends in how customers evaluate different dimensions of service quality (More information in Figure 9 and Table 7). The highest volume of positive comments was concentrated within the Assurance and Tangibles dimensions. Specifically, 'Staff Competence and Professionalism' (n = 639 positive), 'Food and Beverage Service' (n = 408 positive), and 'Aircraft Condition' (n = 316 positive) which received the most favorable feedback, highlighting British Airways' strengths in staff professionalism and the quality of onboard amenities. In contrast, several service aspects received substantial negative sentiment. 'Service Promise Fulfillment and Recovery' recorded 772 negative mentions, indicating a significant level of dissatisfaction related to unmet expectations and issue resolution. 'Communication Clarity and Accuracy' also received negative feedback (n = 339 negative; 53 positive), reflecting customer frustration with unclear or insufficient information. The Responsiveness dimension, overall, including 'Prompt Service Delivery' and 'Staff Availability and Accessibility', received limited positive sentiment, suggesting notable concerns about timely support and staff presence. Additionally, aspects under the Reliability dimension, such as 'Operation', 'Maintenance/Condition', and 'Policy and Policy Transparency', were also associated with higher negative than positive feedback. These patterns point to operational shortcomings and procedural issues that affect the consistency of service delivery. On the other hand, Empathy-related aspects such as 'Personalized Attention' and 'Emotional Support and Sensitivity' were mentioned less frequently but showed a more balanced tone. For instance, 'Personalized Attention' received 101 positive and 52 negative mentions, suggesting that while empathy is not a dominant theme in reviews, it is appreciated by customers when encountered.

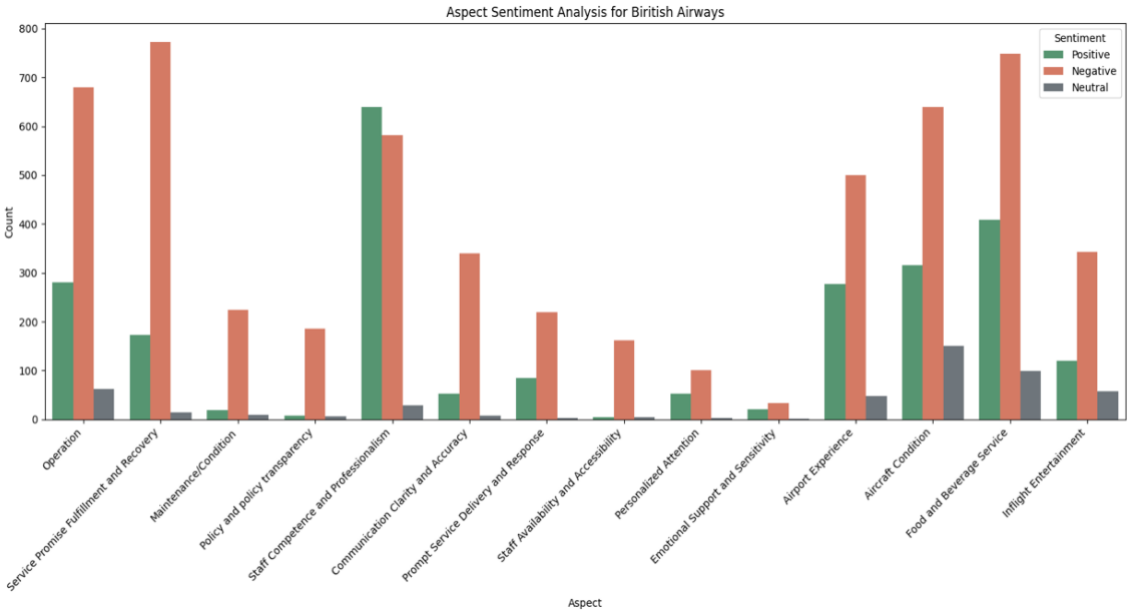

Figure 9: ABSA of service quality themes in British Airways.

Table 7: ABSA of service quality themes and dimensions in British Airways.

Dimension	Themes	Negative	Neutral	Positive
Tangibles	Food and Beverage Service	748	99	408
	Aircraft Condition	639	150	316
	Airport Experience	500	47	277
	Inflight Entertainment	343	57	120
Reliability	Operation	680	62	280
	Service Promise Fulfillment and Accountability	772	14	172
	Maintenance/Condition	224	9	19
	Policy and policy transparency	185	6	7
Assurance	Staff Competence and Professionalism	581	29	639
	Communication Clarity and Accuracy	339	8	53
Responsiveness	Prompt Service Delivery and Response	219	3	85
	Staff Availability and Accessibility	161	5	5
Empathy	Personalized Attention	101	3	52
	Emotional Support and Sensitivity	33	1	21

Figure 10 and Table 8 summarize the influence of different service aspects on overall passenger satisfaction for British Airways, based on an OLS regression model. The coefficient values indicate the direction and strength of each aspect's influence: positive coefficients reflect a positive impact on satisfaction, while negative coefficients reflect a damaging effect. Statistical significance is indicated by P-values, with values below 0.05 indicating reliable relationships. Among the most influential positive predictors of satisfaction were 'Staff Competence and Professionalism' (β = 0.95, P = 0) and 'Operation' (β = 0.89, P = 0). These themes fall under the dimensions of Assurance and Reliability, suggesting that skilled personnel and smooth flight procedures can play an important role in shaping customer experience at British Airways.

On the negative side, several service themes demonstrated significant negative effects on satisfaction. 'Staff Availability and Accessibility' (β = –1.89, P = 0), 'Policy and Policy Transparency' (β = –1.10, P = 0), and 'Service Promise Fulfillment and Accountability' (β = –0.82, P = 0) showed the strongest negative coefficients. These themes, primarily within the Responsiveness and Reliability dimensions, suggest that issues with personnel availability, unclear policies, and unmet service expectations are sources of dissatisfaction for British Airways customers. Similarly, 'Communication Clarity and Accuracy' (β = –1.79, P = 0) and 'Maintenance/Condition' (β = –0.22, P = 0) displayed negative coefficients, though their influence was statistically insignificant. Meanwhile, 'Airport Experience' (β = 0.70, P = 0) and 'Aircraft Condition' (β = 0.76, P = 0) contributed positively to satisfaction, indicating the value of the Tangibles dimension. The results highlight that professional service delivery and operational reliability are key satisfaction drivers, while breakdowns in responsiveness and unclear policies significantly impact the customer experience.

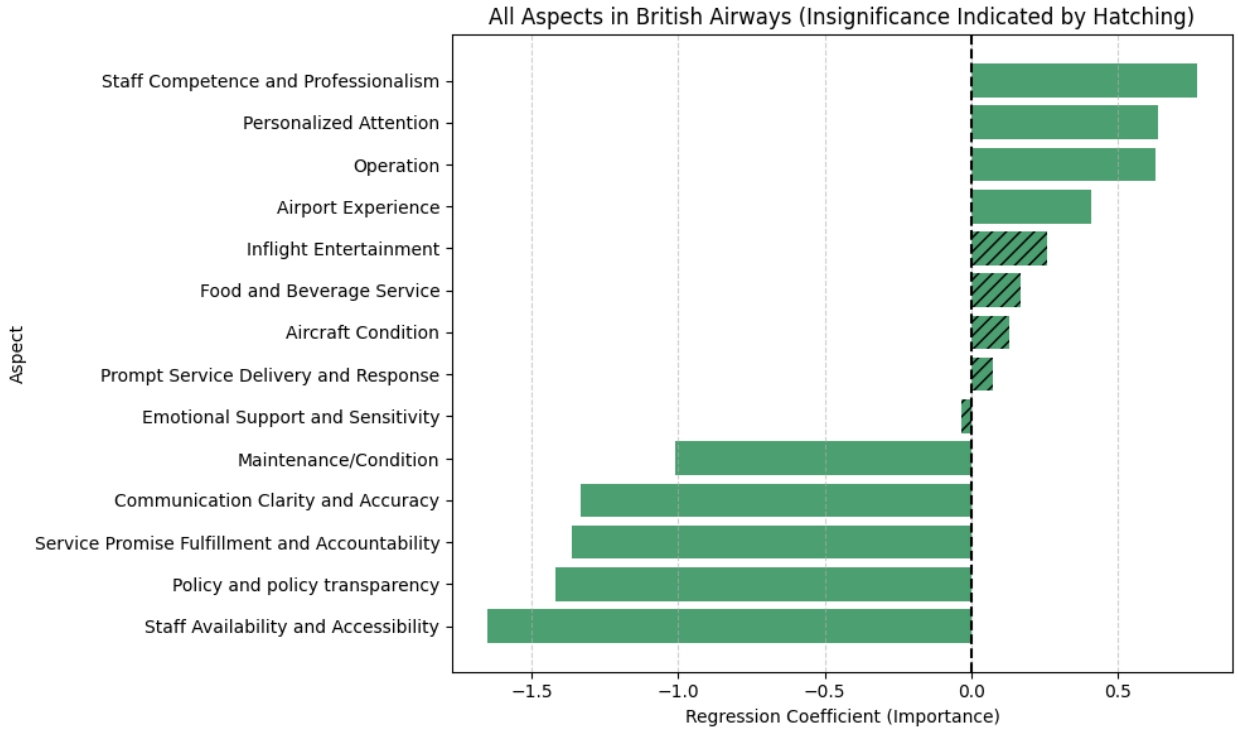

Figure 10: OLS regression analysis of all service themes in British Airways with indication of statistical insignificance.

Table 8: OLS regression analysis of service quality themes and dimensions in British Airways.

Dimension	Themes	Coefficient	P-value
Tangibles	Airport Experience	0.70	0*
	Aircraft Condition	0.76	0*
	Food and Beverage Service	0.29	0.30
	Inflight Entertainment	0.74	0*
Reliability	Service Promise Fulfillment and Accountability	-0.82	0*
	Operation	0.89	0*
	Maintenance/Condition	-0.22	0.57
	Policy and policy transparency	-1.10	0*
Assurance	Staff Competence and Professionalism	0.95	0*
	Communication Clarity and Accuracy	-1.79	0*
Responsiveness	Prompt Service Delivery and Response	0.79	0.01*
	Staff Availability and Accessibility	-1.89	0*
Empathy	Personalized Attention	0.59	0.21
	Emotional Support and Sensitivity	-1.47	0.01*

Note: * Indicates statistical significance at $p < 0.05$

To conclude, the sentiment analysis and regression findings offer a detailed view of how different service dimensions contribute to customer satisfaction at British Airways, as demonstrated in Table 9. In the following, only the statistically significant dimensions are discussed ($p > 0.05$). One of the considerable dimensions in customer satisfaction is Assurance, which covers 'Staff competence and Professionalism' and 'Communication Clarity and Accuracy' themes. This dimension showed a relatively balanced number of sentiments to total mentions ($p = 639$, $n = 581$, out of 1,249) and had a statistically significant negative coefficient ($\beta = –0.42$, $P = 0$), indicating its crucial role in shaping customer dissatisfaction. On the other hand, Responsiveness is also one of the problematic dimensions. It has a total review of 878, which has a high proportion of negative sentiment ($n = 719$ negative vs. 143 positive). In addition, it recorded a negative regression coefficient ($\beta = –0.55$, $P = 0$), suggesting persistent issues in staff availability and prompt responsiveness. These results highlight clear priorities for service improvement, particularly the need to enhance responsiveness and assurance consistency.

To illustrate customer perception across the five SERVPERF dimensions (Tangibles, Reliability, Assurance, Responsiveness, Empathy), two bar charts are presented. Figure 11 shows the average sentiment for each dimension, calculated as the difference between positive and negative mentions relative to the total. Figure 12 displays regression coefficients, where dimensions with $P < 0.05$ are considered statistically significant (solid bars), while others are shown with hatching.

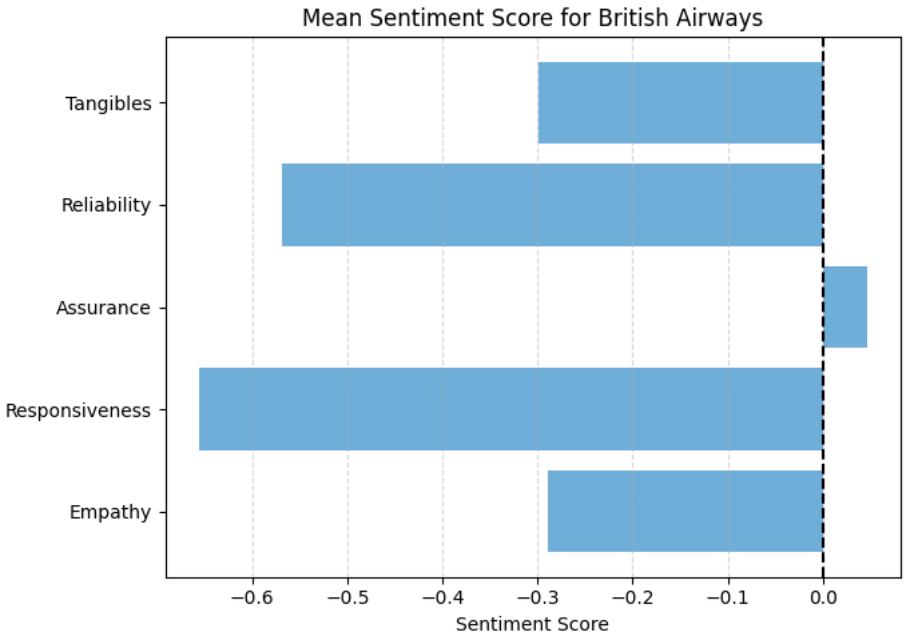

Figure 11: Mean sentiment scores of service quality dimensions in British Airways.

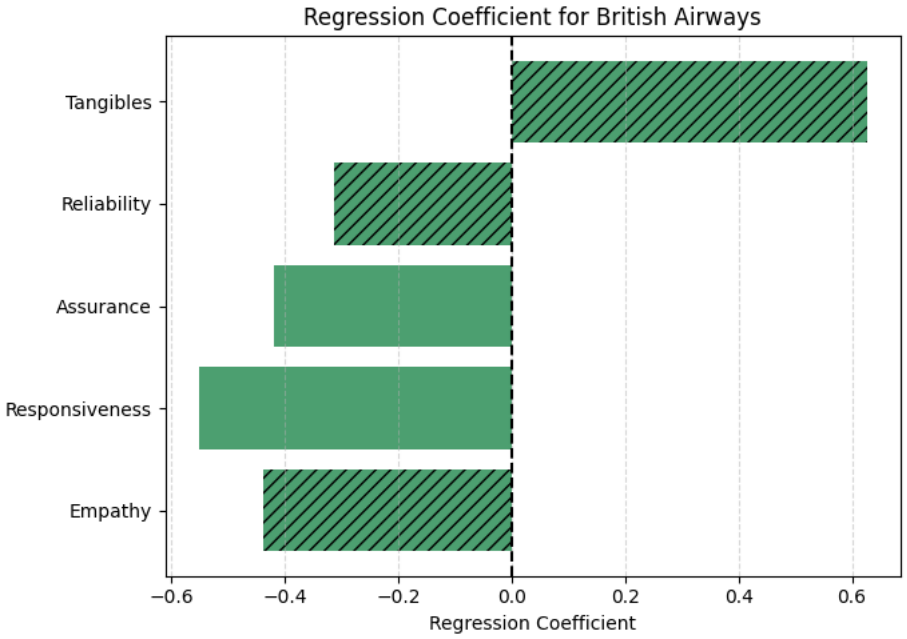

Figure 12: OLS regression analysis of service quality dimensions in British Airways.

Note: Insignificance indicated by hatching

Table 9: Summary of sentiment and OLS regression analysis for service quality dimensions in British Airways.

Dimension	Negative	Neutral	Positive	Total	Coefficient	P-value
Tangibles	2230	353	1121	3704	0.62	0.08
Reliability	1861	91	478	2430	-0.31	0.14
Assurance	581	29	639	1249	-0.42	0*
Responsiveness	719	16	143	878	-0.55	0*
Empathy	134	4	73	211	-0.43	0.11

Note: * Indicates statistical significance at $p < 0.05$

5.3. Comparative Summary of Airline Service Quality Results

This section presents a comparative overview of the main service quality extracted from 11 European airlines, which combines insights from ABSA and OLS regression results, as demonstrated in Table 10. By analyzing both the volume and tone of customer feedback and the statistical impact of service dimensions on satisfaction, a clearer picture of each airline's service performance emerges.

Figure 13 displays that Reliability was a particularly prominent dimension for Lufthansa, Norwegian, and Ryanair, where it received a significantly higher number of mentions compared to other service aspects. A high number of mentions, especially with negative sentiment, often points to customer dissatisfaction with delays, cancellations, baggage handling, or service inconsistencies, and generally, service delivery failure. Similarly, Tangibles stood out in Air France, indicating that passengers frequently commented on the physical aspects of service, such as aircraft condition and onboard amenities. In addition, the dimension Assurance was mentioned across all airlines, making it the only service aspect that appeared consistently in every carrier's feedback, and Reliability was also frequently noted across multiple airlines. This suggests that passengers value trust, professionalism, and dependable service the most. While Reliability and Assurance were mentioned most often, the two dimensions of Empathy and Responsiveness were not as commonly discussed. Empathy appeared mainly in airlines Lufthansa and KLM, showing that only a few airlines stood out for offering personal attention or care. Moreover, Responsiveness was not only mentioned less frequently but also carried mostly negative sentiment when it was discussed. This suggests that while passengers might not always comment on responsiveness, when they do, it is often due to dissatisfaction. It may highlight this dimension as a weak point in service delivery. In addition, it should be noted that only the service dimensions with statistically significant impact ($p < 0.05$) are included in this summary, and thus, not all dimensions for every airline are represented in the figure.

Figure 14 complements this view by presenting the statistically significant service dimensions ($p < 0.05$) derived from OLS regression. Only the statistically significant service dimensions were

included in the regression-based analysis and visualizations. Among all the carriers, Air France stood out with the highest positive coefficient for Tangibles (0.82), indicating that passengers placed high value on physical service elements such as aircraft condition, entertainment, and catering. However, Tangibles did not appear as a statistically significant driver for other airlines. Instead, Empathy was identified as the strongest positive dimension for airlines like Lufthansa (0.89) and KLM (0.29), which emphasizes the value of emotional support, personalized service, and attentive staff interactions. Meanwhile, a clear trend of negative coefficients across Responsiveness, Reliability, and Assurance was evident in nearly all airlines. Air France, for instance, received a strong negative coefficient in Reliability (-0.55), and British Airways showed significant negative impacts in both Responsiveness (-0.55) and Assurance (-0.42). While performing well in Empathy, Lufthansa still had significant shortcomings in Responsiveness (-0.77) and Reliability (-0.38). Interestingly, the Assurance dimension was the only aspect that was statistically significant for all airlines with a negative coefficient. However, Figure 13 reveals a more balanced or even positive tone in passenger feedback regarding the professionalism and communication of airline staff. This contrast may indicate that professionalism and clear communication are expectations in air travel, which are acknowledged when present but heavily criticized when absent. Notably, KLM was an exception in the regression analysis, where Assurance had a positive coefficient, indicating a direct and favorable impact on overall customer satisfaction.

These findings suggest that customer satisfaction depends not only on the range of services provided by airlines, but also on the reliability and level of care and professionalism with which those services are delivered.

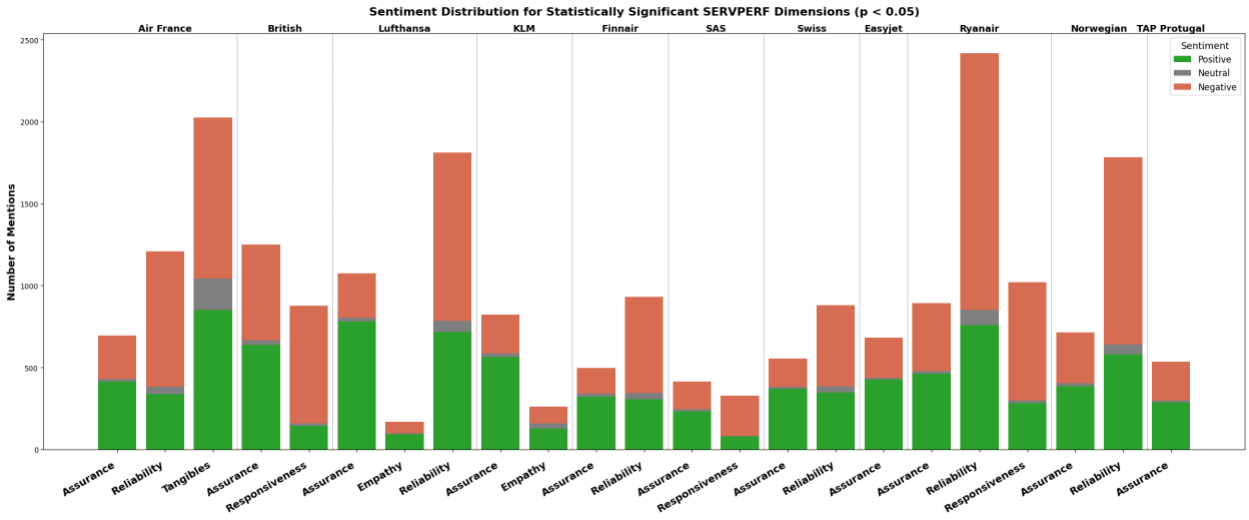

Figure 13: Comparative sentiment analysis of service quality dimensions across all selected Airlines.

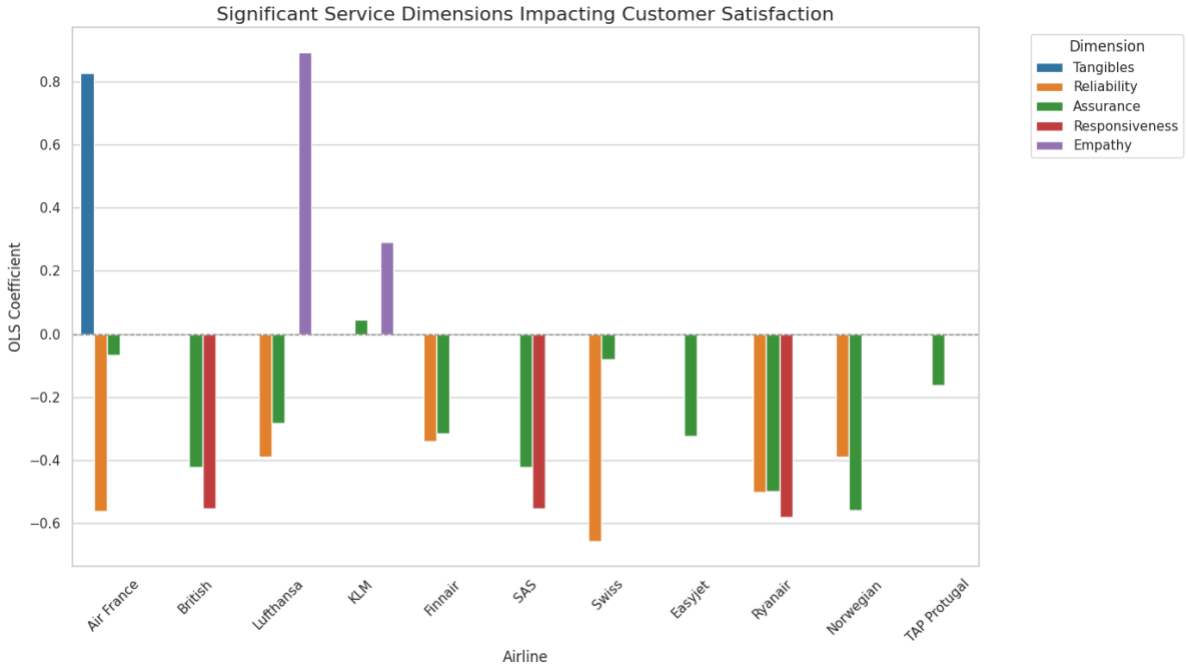

Figure 14: OLS regression analysis of service quality dimensions across all selected Airlines.

Table 10: Summary of sentiment and OLS regression analysis for service quality dimensions across all Airlines.

Airlines	Dimension	Negative	Neutral	Positive	Total	Coefficient	P-value
Air France	Tangibles	982	193	851	2026	0.82	0.01
	Reliability	826	45	339	1210	-0.55	0.02
	Assurance	264	18	414	696	-0.06	0
British	Assurance	581	29	639	1249	-0.42	0
	Responsiveness	719	16	143	878	-0.55	0
Lufthansa	Reliability	1027	67	718	1812	-0.38	0
	Assurance	272	21	783	1076	-0.28	0
	Empathy	72	5	93	170	0.89	0.02
KLM	Assurance	234	22	566	822	0.04	0
	Empathy	101	31	129	261	0.29	0
Finnair	Reliability	585	40	305	930	-0.34	0
	Assurance	156	18	322	496	-0.31	0
SAS	Assurance	169	13	232	414	-0.42	0
	Responsiveness	243	3	81	327	-0.55	0
Swiss	Reliability	495	38	348	881	-0.65	0
	Assurance	173	12	370	555	-0.08	0
EasyJet	Assurance	247	10	426	683	-0.32	0
Ryanair	Reliability	1566	94	759	2419	-0.50	0.02
	Assurance	415	16	462	893	-0.49	0
	Responsiveness	722	15	285	1022	-0.57	0
Norwegian	Reliability	1142	58	582	1782	-0.38	0
	Assurance	308	19	387	714	-0.55	0
TAP Portugal	Assurance	234	13	288	535	-0.16	0

5.4. Evaluation Metric

The accuracy and reliability of the model's detection of aspects were assessed using standard evaluation metrics. To measure its effectiveness, the model's predictions were compared with manually annotated data through the use of confusion matrices and the Micro F1 score, providing a comprehensive evaluation of its classification performance. In this study, the model achieved a Micro F1 score of 0.9211 for detecting service aspects. This shows that the model was very accurate in identifying the main aspects discussed in the reviews. Additionally, for sentiment analysis, the Micro F1 score was 0.9298, which means the model was also reliable at determining whether the sentiment was positive, negative, or neutral. These scores indicate the model's strong capability in accurately processing customer feedback and producing dependable results.

As shown in Figure 15, the sentiment analysis performed well in identifying both negative and positive reviews. It correctly identified 80 out of 80 negative sentiment labels and 25 out of 26 positive labels, which shows a high accuracy for these categories. However, it has challenges with neutral sentiment, correctly identifying only 1 out of 3 cases. Most errors happened when the neutral reviews were misclassified as either negative or positive. Since this analysis was done using a prompt-based multi-agent approach and not a traditional training model, the results suggest that the agents found it more difficult to recognize neutral or mixed sentiments. Improving the prompt instructions or adding clarification examples may help increase accuracy for the neutral class.

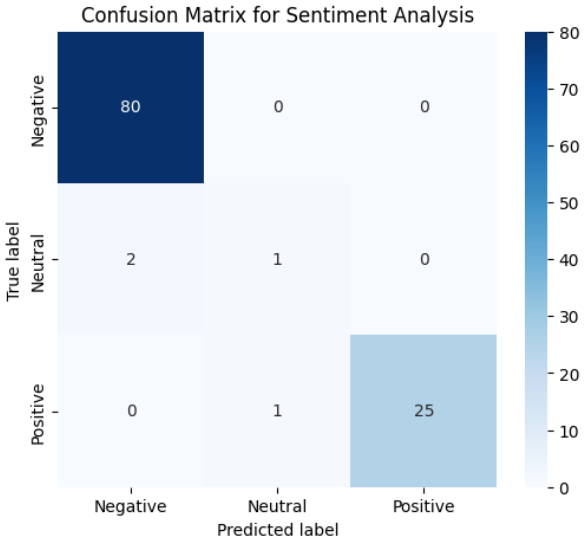

Figure 15: Performance evaluation of sentiment analysis using a confusion matrix.

Similarly, as shown in the aspect classification confusion matrix (Figure 16), the multi-agent system demonstrates strong performance for the most frequently mentioned service quality dimensions. For example, it correctly identifies 19 instances of "Service Promise Fulfillment and Accountability," 17 for "Staff Competence and Professionalism," and 14 for "Operation." Most predictions align with the truth, which suggests that the system is generally able to classify aspects accurately with minimal misclassification. However, some categories, such as "Policy and Policy Transparency," "Personalized Attention," and "Maintenance/Condition," appear less frequently and show minor confusion with other labels. This may be due to limited representation or overlapping language in customer reviews. Overall, the model handles dominant service quality aspects effectively, but performance on less common categories could be improved through prompt refinement or targeted evaluation.

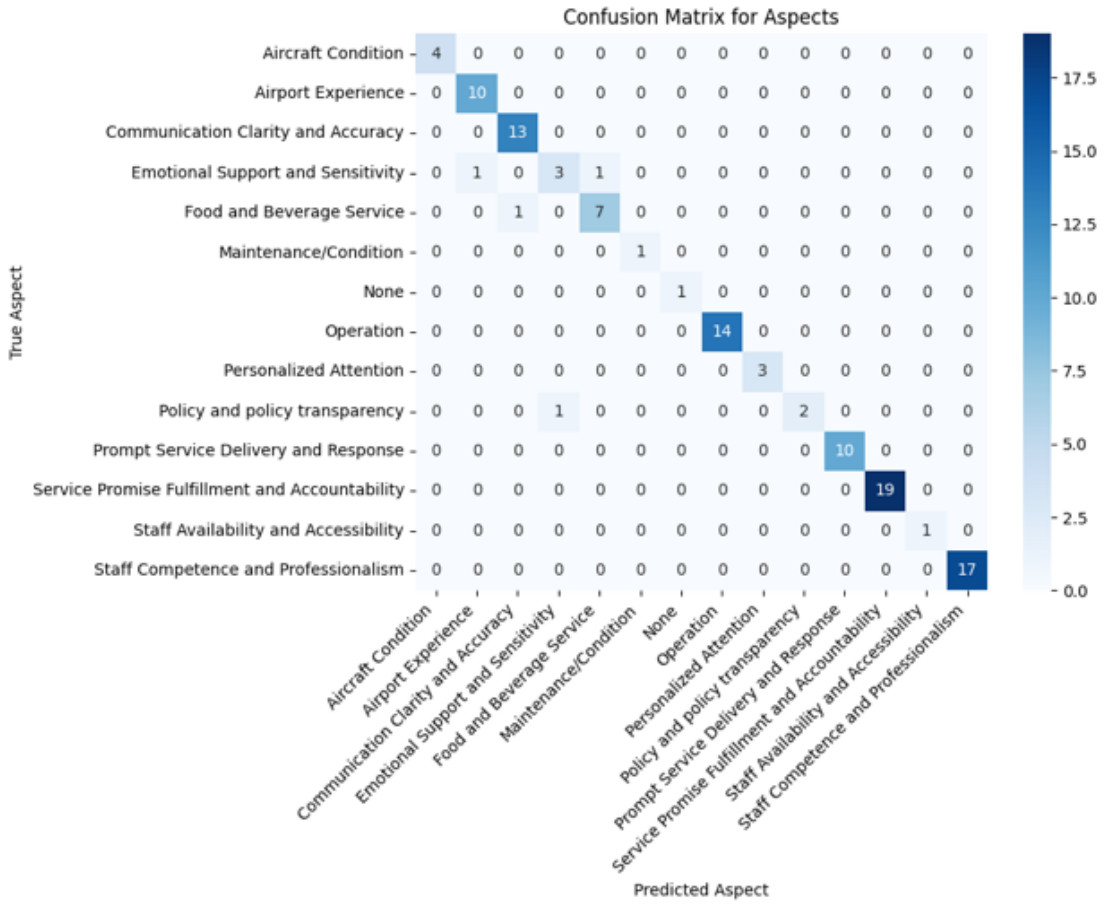

Figure 16: Performance evaluation of aspect analysis using a confusion matrix.

6. Conclusion

This research analyzed 10,115 verified online reviews of 11 European airlines with AI technologies. With the rise of high-volume user-generated content on social media and review platforms, commonly referred to as big data, LLMs can be a powerful tool for extracting meaningful insights. This approach aligns with the smart tourism ecosystem theory and provides a valuable lens for situating online reviews within a broader digital context. Additionally, the SERVPERF model was chosen as another theory for service quality, which has five core dimensions: tangibles, reliability, assurance, responsiveness, and empathy. This theory is used in a multi-agent system for extracting service dimension, but the concept was broad, and these broad categories are not always directly mentioned by customers in reviews. Therefore, a more detailed definition was necessary for the AI agent. To achieve this, a sample of 50 online reviews was analyzed to extract themes and sub-themes to expand the broad concept of the SERVPERF theoretical framework and design a suitable prompt for using in a multi-agent system. As a result, the manual validation step strengthened the credibility of the automated results, bridging the gap between human interpretation and ML. Following this, the refined prompt was used for ABSA on the full dataset, supported by visualization. The multi-agent system, which combined Gemini 1.5 Flash and GPT-3.5 Turbo, demonstrated strong performance in large-scale review analysis. In this method, evaluation is a key step. Furthermore, 20 samples of the results were checked manually, and the micro F1 score for the imbalanced dataset was used to show evaluation metrics. Regarding this, the model reached a performance score of 0.9211 for service dimension classification and 0.9298 for sentiment analysis. This whole process could answer the first research question, which was focusing on the method how airline service quality can be measured on a large-scale dataset. Then, OLS regression analysis provided a clear answer to the second research question, which aimed to identify which service quality dimensions were most influential on overall customer satisfaction in different airlines based on online review analysis. Dimensions with a P-value below 0.05 were considered statistically significant and interpreted as either a challenge or an opportunity, depending on the direction of the coefficient.

The findings of this study revealed that certain service dimensions, reliability, responsiveness, and particularly assurance, were more strongly associated with overall satisfaction than others. Interestingly, the assurance was significant across all airlines but mostly with a negative coefficient, indicating that poor performance in this area contributes to dissatisfaction. The assurance dimension was defined through four sub-themes: staff communication language barrier, staff behavior and training, clear and consistent information, and accuracy of provided information. All are showing the importance of staff communication and professionalism in customer satisfaction. Notably, only KLM could achieve a positive coefficient in assurance, which may offer other airlines a potential best-practice model in this dimension. Also, the tangibles dimension, while still relevant, showed comparatively lower impact on satisfaction. This may be due to the fact that physical amenities are often standardized across airlines, reducing their influence as differentiating factors and shifting customer focus toward service consistency and trust-based interactions. Only Air France showed tangibles as a statistically significant dimension, and it had a positive impact on customer satisfaction. In addition, Lufthansa, Air France, Finnair, Swiss Air, Ryanair, and Norwegian had challenges in the reliability dimension. This dimension consists of themes such as service promise fulfillment and accountability, operation, maintenance, and policy transparency. These findings reflect a systemic gap between service expectations and

actual performance. Therefore, reliability can also be a critical area for improvement. Another notable observation was that the empathy dimension, which only showed in two airlines, KLM and Lufthansa, had a strong positive influence on overall satisfaction with confidence of 0.29 and 0.89, respectively. This suggests that empathy is not common across all airlines, but when it is present, it makes a significant difference. It helps customers feel cared for and understood, which improves their overall experience.

This study also reflects some main contributions to the tourism industry. Methodologically, it demonstrates the effectiveness of combining qualitative content analysis with multi-agent LLMs for analyzing large-scale unstructured data. Theoretically, the integration of SERVPERF with the Smart Tourism Ecosystem provides a practical framework for evaluating service quality in the digital age. Empirically, by analyzing 10,115 verified reviews from 11 European airlines through sentiment classification and regression analysis, the study identifies which SERVPERF dimensions most strongly affect passenger satisfaction. These data-driven results offer targeted, airline-specific insights that highlight both challenges and improvement opportunities. Practically, it contributes a scalable, AI-supported approach that enables airlines to transform customer feedback into actionable improvements, promoting smarter and more responsive service strategies. In conclusion, this study demonstrates that combining theoretical frameworks with advanced AI tools can provide meaningful, context-aware evaluations of service quality, supporting smarter decision-making in the smart tourism paradigm.

7. Limitations and Future Studies

While this study provides valuable insights into airline service quality using user-generated content, it has some limitations that can be potential directions for future research. Firstly, the current analysis was limited to reviews sourced exclusively from the Skytrax website. Despite its credibility, this dataset represents only a subset of the broader range of customer insights. Future studies can enhance both the validity and generalizability of findings by extracting data from multiple platforms such as TripAdvisor, Google Reviews, or airline-specific feedback systems. A multi-platform approach would allow researchers to capture a more diverse range of customer experiences. Secondly, this study focused only on airlines operating within the European region. Although this regional scope allowed for in-depth and consistent analysis, it limits the cross-cultural applicability of the findings. Future research can expand the geographic scope to include airlines from other regions, such as Asia-Pacific, North America, the Middle East, or Latin America, enabling comparative analysis across different regulatory, cultural, and service environments. This would facilitate a deeper understanding of how regional and cultural contexts influence service expectations, satisfaction levels, and review behavior. In addition to geographic diversity, future studies can also explore comparative research between different airline business models, such as low-cost carriers (LCCs) versus full-service carriers (FSCs). Such comparisons would be useful for identifying whether and how expectations of service quality differ based on the fare structure, route length, or service inclusions. Moreover, expanding the types of data used in the analysis can significantly enrich insights. While this study focused particularly on textual reviews, integrating other data modalities such as images (e.g., of seating, meals), voice reviews, videos, flight metadata, or even social media activity can provide a more comprehensive understanding of service quality perceptions. Advances in multimodal AI now make it feasible to

analyze diverse data types within a unified framework, allowing for more comprehensive interpretations. In addition, if future studies have access to stronger technical resources, they can use sub-theme prompts instead of general themes. This would help identify more detailed and specific aspects of service quality, leading to a deeper understanding of customer experiences.

Finally, future research can explore longitudinal analysis by examining how perceptions of service quality have changed over time, especially in response to industry events such as global crises (e.g., COVID-19), policy changes, or the implementation of new technologies (e.g., biometric boarding, AI chatbots). Tracking such shifts could help airlines proactively adapt their service strategies to meet changing customer expectations.

By addressing these research gaps, future studies can build on this work to develop more comprehensive, globally relevant, and technologically integrated approaches to service quality evaluation within both the airline industry and the tourism sector.

References

Adom, D., Hussein, E. K., & Agyem, J. A. (2018). Theoretical and conceptual framework: Mandatory ingredients of a quality research. *International Journal of Scientific Research, 7*(1), 438–441. https://www.researchgate.net/publication/322204158

Ajayi, V. O. (2023). A review on primary sources of data and secondary sources of data. *European Journal of Education and Pedagogy, 2*(3), 1–3. https://www.researchgate.net/publication/370608670

Andaleeb, S. S., & Conway, C. (2006). Customer satisfaction in the restaurant industry: An examination of the transaction-specific model. *Journal of Services Marketing, 20*(1), 3–11. https://doi.org/10.1108/08876040610646536

Andersson, T. D., & Mossberg, L. L. (2004). The dining experience: Do restaurants satisfy customer needs? *Food Service Technology, 4*(4), 171–177. https://doi.org/10.1111/j.1471-5740.2004.00105.x

Baral, S. (2024). *Using large language models for aspect-based sentiment analysis* (Master's thesis). Aalborg University Business School, Aalborg, Denmark.

Boukkouri, H.E., Ferret, O., Lavergne, T., Noji, H., Zweigenbaum, P., & Tsujii, J. (2020). CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. arXiv preprint arXiv:2010.10392. https://doi.org/10.48550/arXiv.2010.10392.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Buhalis, D., & Amaranggana, A. (2015). Smart tourism destinations: Enhancing tourism experience through personalisation of services. In I. Tussyadiah & A. Inversini (Eds.), *Information and communication technologies in tourism 2015* (pp. 377–389). Springer. https://doi.org/10.1007/978-3-319-14343-9_28

Campesato, O. (2024). *Large language models for developers: A prompt-based exploration of LLMs* (1st ed.). Mercury Learning and Information. https://doi.org/10.1515/9781501520938

Cherryholmes, C. H. (1992). Notes on pragmatism and scientific realism. *Educational Researcher, 21*(6), 13–17. https://www.jstor.org/stable/1176502

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Sage Publications.

Cronin, J. J., & Taylor, S. A. (1992). *Measuring service quality: A reexamination and extension*. Journal of Marketing, **56**(3), 55–68. https://doi.org/10.1177/002224299205600304

Dale, R. (2021). GPT-3: What's it good for? *Natural Language Engineering*, *27*(1), 113–118. https://doi.org/10.1017/S1351324920000601

Danisman, E. (2019). *Skytrax airline reviews* [Data set]. Kaggle. https://www.kaggle.com/datasets/efehandanisman/skytrax-airline-reviews

Dardas, A. Z., Williams, A., & Scott, D. (2020). Carer-employees' travel behaviour: Assisted-transport in time and space. *Journal of Transport Geography*, *82*, 102558. https://doi.org/10.1016/j.jtrangeo.2019.102558

Dimitriadou, A., Papadopoulos, T., & Vlachos, I**.** (2024). Tourism demand forecasting using Gradient Boosting Trees and other regressors: Insights from EU data (2010–2020). *European Journal of Tourism Research, 35*, 123–137.

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing, 62*(1), 107–115. https://doi.org/10.1111/j.1365-2648.2007.04569.x

Essien, A. E., & Chukwukelu, G. (2022). Deep learning in hospitality and tourism: A research framework agenda for future research. *International Journal of Contemporary Hospitality Management, 34*(12), 4480–4515. https://doi.org/10.1108/IJCHM-09-2021-1176

Frost, J. (2019). *Regression analysis: An intuitive guide for using and interpreting linear models* (1st ed.). Statistics by Jim Publishing.

Gemini Team. (2024). *Gemini: A family of highly capable multimodal models* (Version 4) [Technical report]. Google DeepMind. https://arxiv.org/abs/2312.11805

Graneheim, U. H., Lindgren, B.-M., & Lundman, B. (2017). Methodological challenges in qualitative content analysis: A discussion paper. *Nurse Education Today, 56*, 29–34. https://doi.org/10.1016/j.nedt.2017.06.002

Gretzel, U., Sigala, M., Xiang, Z., & Koo, C. (2015). Smart tourism: Foundations and developments. *Electronic Markets, 25*(3), 179–188. https://doi.org/10.1007/s12525-015-0196-8

Gu, S. (2024). *A survey of Large Language Models in Tourism (Tourism LLMs)*. Qeios. https://doi.org/10.32388/8R27CJ

Hansen, K. V. (2014). Development of SERVQUAL and DINESERV for measuring meal experiences in eating establishments. *Scandinavian Journal of Hospitality and Tourism, 14*(2), 116–134. https://doi.org/10.1080/15022250.2014.886094

Hartwig, K., & Billert, M. (2018, June). *Measuring service quality: A systematic literature review*. Proceedings of the 26th European Conference on Information Systems (ECIS 2018), Portsmouth, UK. https://www.researchgate.net/publication/338502007

Hasan, M., Khan, M. N., & Farooqi, R. (2019). Service quality measurement models: Comparative analysis and application in airlines industry. *Global Journal of Enterprise Information System*, *11*(2), 30– 41. https://doi.org/10.18311/gjeis/2019

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H. (2023). *How good are GPT models at machine translation? A comprehensive evaluation*. arXiv preprint arXiv:2302.09210. https://arxiv.org/abs/2302.09210

Kaushik, V., & Walsh, C. A. (2019). Pragmatism as a research paradigm and its implications for social work research. *Social Sciences, 8*(9), 255. https://doi.org/10.3390/socsci8090255

Kim, D., Lim, C., & Ha, H.-K. (2024). Comparative analysis of changes in passenger's perception for airline companies' service quality before and during COVID-19 using topic modeling. *Journal of Air Transport Management, 115*, 102542. https://doi.org/10.1016/j.jairtraman.2024.102542

Kumar, S., & Zymbler, M. (2019). A machine learning approach to analyze customer satisfaction from airline tweets. *Journal of Big Data, 6*(1), 62. https://doi.org/10.1186/s40537-019-0224-1
Lau, T., Wang, H.-C., & Chuang, C.-C. (2011). A definition of service as base for developing service science. In *2011 International Joint Conference on Service Sciences* (pp. 49–53). IEEE. https://doi.org/10.1109/IJCSS.2011.18

Lauriola, I., Lavelli, A., & Aiolli, F. (2022). An introduction to deep learning in natural language processing: Models, techniques, and tools. *Neurocomputing, 470*, 443–456. https://doi.org/10.1016/j.neucom.2021.05.103

Lee, P., Hunter, W. C., & Chung, N. (2020). Smart tourism city: Developments and transformations. *Sustainability, 12*(10), 3958. https://doi.org/10.3390/su12103958

Libent, L., & Magasi, C. (2024). Service quality and customer satisfaction in the airline industry in Tanzania: A case of Air Tanzania Company Limited. *International Journal of Research in Business and Social Science, 13*(2), 59–71. https://doi.org/10.20525/ijrbs.v13i2.3122

Liu, X.-X., & Chen, Z.-Y. (2022). Service quality evaluation and service improvement using online reviews: A framework combining deep learning with a hierarchical service quality model. *Electronic Commerce Research and Applications, 54*, 101174. https://doi.org/10.1016/j.elerap.2022.101174

Mao, R., Liu, Q., He, K., Li, W., & Cambria, E. (2023). The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE*

Transactions on Affective Computing, 14(3), 1743–1752. https://doi.org/10.1109/TAFFC.2022.3204972

Mendieta-Aragón, E., & Garín-Muñoz, T. (2023). Comparing machine learning models on Spain's travel survey data: Logistic Regression, MLP, and Random Forest. *Tourism Economics, 29*(3), 467–481.

Mondillo, G., Frattolillo, V., Colosimo, S., Perrotta, A., Di Sessa, A., Guarino, S., Miraglia del Giudice, E., & Marzuillo, P. (2025). Basal knowledge in the field of pediatric nephrology and its enhancement following specific training of ChatGPT-4 "omni" and Gemini 1.5 Flash. *Pediatric Nephrology, 40*(1), 151–157. https://doi.org/10.1007/s00467-024-06486-3

Morgan, D. L. (2016). *Paradigms lost and pragmatism regained: Methodological implications of combining qualitative and quantitative methods*. In K. B. Jensen (Ed.), *A handbook of media and communication research: Qualitative and quantitative methodologies* (2nd ed., pp. 55–72). Routledge.

Morgan, D. L. (2007). Paradigms lost and pragmatism regained: Methodological implications of combining qualitative and quantitative methods. *Journal of Mixed Methods Research, 1*(1), 48–76. https://doi.org/10.1177/2345678906292462

Mowlaei, M. E., Saniee Abadeh, M., & Keshavarz, H. (2020). Aspect-based sentiment analysis using adaptive aspect-based lexicons. *Expert Systems with Applications, 148*, 113234. https://doi.org/10.1016/j.eswa.2020.113234

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quarterly, 34*(1), 185–200. https://doi.org/10.2307/20721420

Murugesan, R., Rekha, A. P., Nitish, N., & Balanathan, R. (2024). Forecasting airline passengers' satisfaction based on sentiments and ratings: An application of VADER and machine learning techniques. *Journal of Air Transport Management, 120*, 102668. https://doi.org/10.1016/j.jairtraman.2024.102668

Mustafa, A., Fong, J. P., Lim, S. P., & Hamid, H. A. (2005, January). *The evaluation of airline service quality using the analytic hierarchy process (AHP)*. Paper presented at the International Conference on Tourism Development, Penang, Malaysia. Retrieved from https://eprints.usm.my/429/

Namkung, Y., & Jang, S. C. (2008). Are highly satisfied restaurant customers really different? A quality perception perspective. *International Journal of Contemporary Hospitality Management, 20*(2), 142–155. https://doi.org/10.1108/09596110810852131

Neuman, W. L. (1997). *Social research methods: Qualitative and quantitative approaches* (3rd ed.). Allyn & Bacon.

Ouaddi, C., Benaddi, L., Bouziane, E. M., Naimi, L., Rahouti, M., Jakimi, A., & Saadane, R. (2025). Assessing the effectiveness of large language models for intent detection in tourism chatbots: A comparative analysis and performance evaluation. *Scientific African, 28*, e02649. https://doi.org/10.1016/j.sciaf.2025.e02649

Palese, B., & Usai, A. (2018). The relative importance of service quality dimensions in E-commerce experiences. *International Journal of Information Management, 40*, 132–140. https://doi.org/10.1016/j.ijinfomgt.2018.02.001

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1985). A conceptual model of service quality and its implications for future research. *Journal of Marketing, 49*(4), 41–50. https://doi.org/10.1177/002224298504900403

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing, 64*(1), 12–40. Pollack, B. L. (2009). Linking the hierarchical service quality model to customer satisfaction and loyalty. *Journal of Services Marketing, 23*(1), 42–50. https://doi.org/10.1108/08876040910933084

Rodrigues, L. L. R., Hussain, A., Aktharsha, U. S., & Nair, G. (2013). *Service quality measurement: Issues and perspectives*. Diplomica Verlag. https://ebookcentral.proquest.com/lib/aalborguniv-ebooks/detail.action?docID=1324041

Samir, H. A., Abd-Elmegid, L., & Marie, M. (2023). Sentiment analysis model for airline customers' feedback using deep learning techniques. *International Journal of Engineering Business Management, 15*(1), 1–23. https://doi.org/10.1177/18479790231206019

Sancho Núñez, J. C., Gómez-Pulido, J. A., & Robina Ramírez, R. (2024). Machine learning applied to tourism: A systematic review. *WIREs Data Mining and Knowledge Discovery, 14*(5), e1549. https://doi.org/10.1002/widm.154

Seth, N., Deshmukh, S. G., & Vrat, P. (2005). *Service quality models: A review*. International Journal of Quality & Reliability Management, 22(9), 913–949. https://doi.org/10.1108/02656710510625211

Silvestri, C., Aquilani, B., & Ruggieri, A. (2017). Service quality and customer satisfaction in thermal tourism. *The TQM Journal, 29*(1), 55–81. https://doi.org/10.1108/TQM-06-2015-0089

Skytrax website. *Airline quality and customer reviews*, https://www.airlinequality.com

Skytrax website. *Verified airline and airport reviews*. https://www.airlinequality.com/verified-reviews/Airline Quality

Song, H., & Liu, H. (2017). Predicting tourist demand using big data. In Z. Xiang & D. R. Fesenmaier (Eds.), *Analytics in smart tourism design: Concepts and methods* (pp. 13–29). Springer. https://doi.org/10.1007/978-3-319-44263-1_2

Stepchenkova, S., Kirilenko, A. P., & Morrison, A. M. (2009). Facilitating content analysis in tourism research. *Journal of Travel Research, 47*(4), 454–469. https://doi.org/10.1177/0047287508326509

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

Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2020). Efficient transformers: A survey. *arXiv preprint arXiv:2009.06732*. https://doi.org/10.48550/arXiv.2009.06732.

Tidake, V. S., & Sane, S. S. (2018). Multi-label classification: A survey. *International Journal of Engineering & Technology, 7*(4.19), 1045–1054. https://doi.org/10.14419/ijet.v7i4.19.28284

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need*. In Advances in Neural Information Processing Systems, 30. https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wang, Y. (2023). Factors affect airline customers' satisfaction: Data mining. In V. Gaikar et al. (Eds.), *Proceedings of the 4th International Conference on E-Commerce and Internet Technology (ECIT 2023)* (pp. 344–362). Atlantis Press. https://doi.org/10.2991/978-94-6463-210-1_43

Wooldridge, J. M. (2013). *Introductory econometrics: A modern approach* (5th ed.). South-Western Cengage Learning.

Xiang, Z., Tussyadiah, I., & Buhalis, D. (2015). Smart tourism: Foundations and developments. *Electronic Markets, 25*(3), 179–188. https://doi.org/10.1007/s12525-015-0196-8

Xiang, Z., & Fesenmaier, D. R. (2017). Big data analytics, tourism design and smart tourism. In Z. Xiang & D. R. Fesenmaier (Eds.), *Analytics in smart tourism design: Concepts and methods* (pp. 299–307). Springer. https://doi.org/10.1007/978-3-319-44263-1_19

Xu, X., Wang, X., Li, Y., & Haghighi, M. (2017). Business intelligence in online customer textual reviews: Understanding consumer perceptions and influential factors. *International*

Journal of Information Management, 37(6), 673–683. https://doi.org/10.1016/j.ijinfomgt.2017.06.004

Yan, B., Li, K., Xu, M., Dong, Y., Zhang, Y., Ren, Z., & Cheng, X. (2025). On protecting the data privacy of Large Language Models (LLMs) and LLM agents: A literature review. *High-Confidence Computing, 4*, 100300. https://doi.org/10.1016/j.hcc.2025.100300

Yang, X., Li, Y., Zhang, X., Chen, H., & Cheng, W. (2023). *Exploring the limits of ChatGPT for query or aspect-based text summarization*. arXiv preprint arXiv:2302.08081. https://arxiv.org/abs/2302.08081

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Yin, B., & Hu, X. (2023). *Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond* [Preprint]. arXiv. https://arxiv.org/abs/2304.13712

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing, 4*, 100211. https://doi.org/10.1016/j.hcc.2024.100211

Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., Cui, Y., Zhou, Z., Gong, C., Shen, Y., Zhou, J., Chen, S., Gui, T., Zhang, Q., & Huang, X. (2024). *A comprehensive capability analysis of GPT-3 and GPT-3.5 series models*. arXiv. https://arxiv.org/abs/2312.11512

Yu, M., & Hyun, S. S. (2019). The impact of foreign flight attendants' service quality on behavioral intention toward their home country—Applied SERVPERF model. *Sustainability, 11*(15), 4136. https://doi.org/10.3390/su11154136

Zadeh, A. A., Leevy, J. L., & Khoshgoftaar, T. M. (2024). A survey on the choice between binary classification and one-class classification. *Journal of Big Data*, *11*, Article 53.

Zhu, X., Li, M., & Song, H. (2019). Semantic network analysis of Airbnb reviews: A Leximancer approach. *International Journal of Hospitality Management, 82*, 244–257.

Álvarez-Carmona, M. Á., Aranda, R., Rodríguez-González, A. Y., Fajardo-Delgado, D., Sánchez, M. G., Pérez-Espinosa, H., Martínez-Miranda, J., Guerrero-Rodríguez, R., Bustio-Martínez, L., & Díaz-Pacheco, Á. (2022). Natural language processing applied to tourism research: A systematic review and future research directions. *Journal of King Saud University - Computer and Information Sciences*. Advance online publication. https://doi.org/10.1016/j.jksuci.2022.10.010

Appendix

Appendix A: LLM Prompts for ABSA

This section provides the prompt templates used to interact with large language models (LLMs) in order to extract service aspects and their associated sentiments from customer reviews. The design of these prompts reflects the theoretical foundation of the SERVPERF framework.

A1. Prompt Submitted to Gemini 1.5 Flash

Analyze the above customer review using aspect-based sentiment analysis.

Your task:
1. Read the review carefully.
2. Identify any of the predefined service categories (listed below) that are mentioned or implied in the review.
3. For each relevant category, assign one sentiment: Positive, Negative, or Neutral.
4. Output only the final list of detected aspects. Do not add any explanation or comments.

Only use the following service categories. Do not add new ones. If none are mentioned, return only:
Aspect: None

Categories:
- Service Promise Fulfillment, Accountability and Recovery
- Operation
- Maintenance/Condition
- Policy and policy transparency
- Staff Competence and Professionalism
- Communication Clarity and Accuracy
- Prompt Service Delivery and Response
- Staff Availability and Accessibility
- Personalized Attention
- Emotional Support and Sensitivity
- Airport Experience
- Aircraft Condition
- Food and Beverage Service
- Inflight Entertainment

Format your output like this:
Aspect: [Category Name], Sentiment: [Positive | Negative | Neutral]
""""),

expected_output="A list of relevant service aspects with sentiment labels only. No explanation.",
agent=agent_gemini

A2. Prompt Submitted to GPT-3.5 Turbo (Validation and Refinement)

The following is the initial structured sentiment analysis produced by another agent:

{gemini_output}

Your task:
1. Read the provided structured analysis.
2. Check if the detected aspects and assigned sentiments are accurate.
3. Correct any mistakes, and add any important missed aspects.
4. Output only the final cleaned list. Do not add any comments or explanation.

Use only the following categories (do not add new ones):
- Service Promise Fulfillment and Accountability
- Operation
- Maintenance/Condition
- Policy and policy transparency
- Staff Competence and Professionalism
- Communication Clarity and Accuracy
- Prompt Service Delivery and Response
- Staff Availability and Accessibility
- Personalized Attention
- Emotional Support and Sensitivity
- Airport Experience
- Aircraft Condition
- Food and Beverage Service
- Inflight Entertainment

Format your output like this:
Aspect: [Category Name], Sentiment: [Positive | Negative | Neutral]
"""),
expected_output="Only the cleaned and corrected list of aspects with sentiment labels. No explanation.",
agent=agent_openai

Appendix B: Analysis of Air France Airline

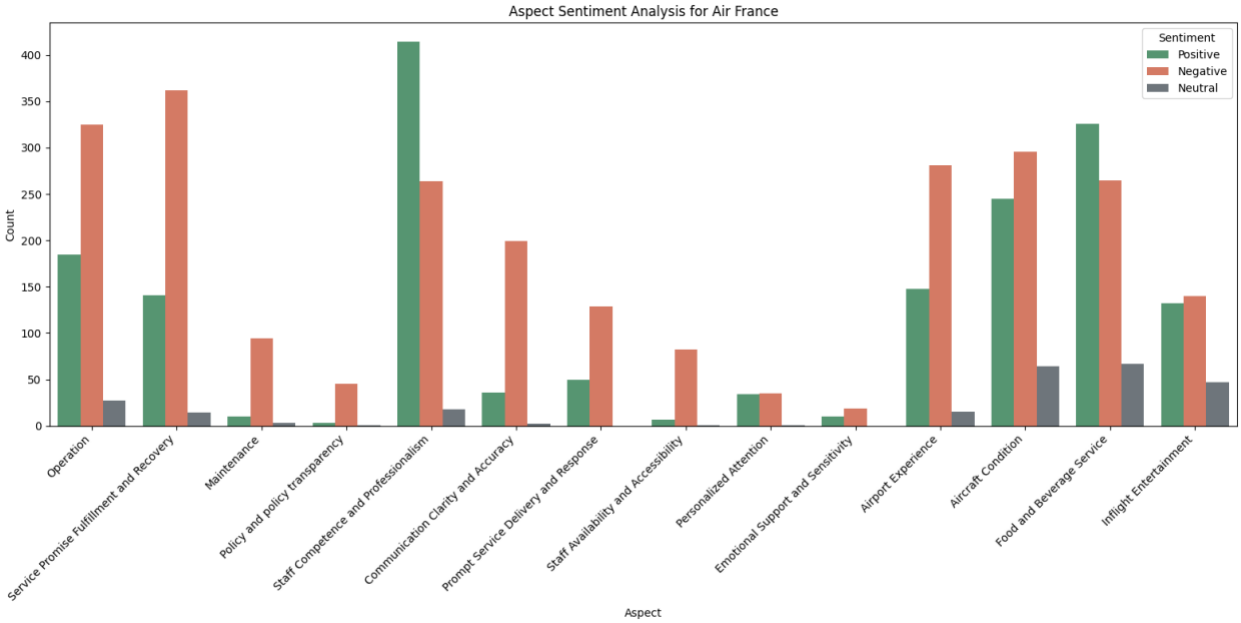

Figure B1: ABSA of service quality themes in Air France Airline

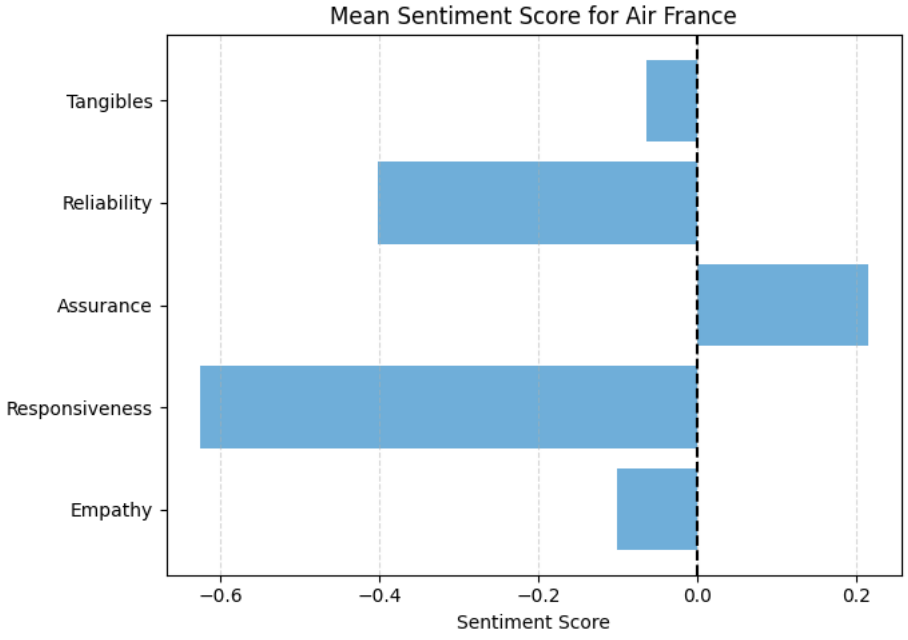

Figure B2: Mean sentiment scores of service quality dimensions in Air France Airline

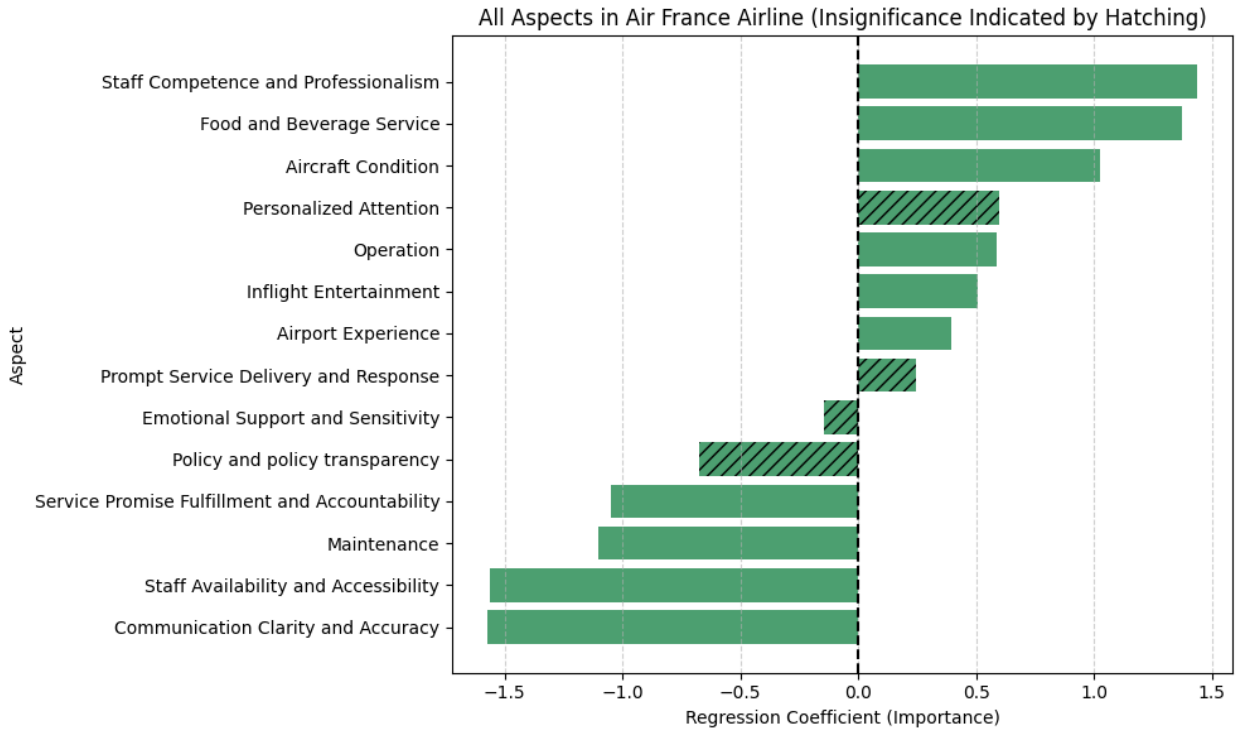

Figure B3: OLS regression analysis of all service themes in Air France Airline with indication of statistical insignificance.

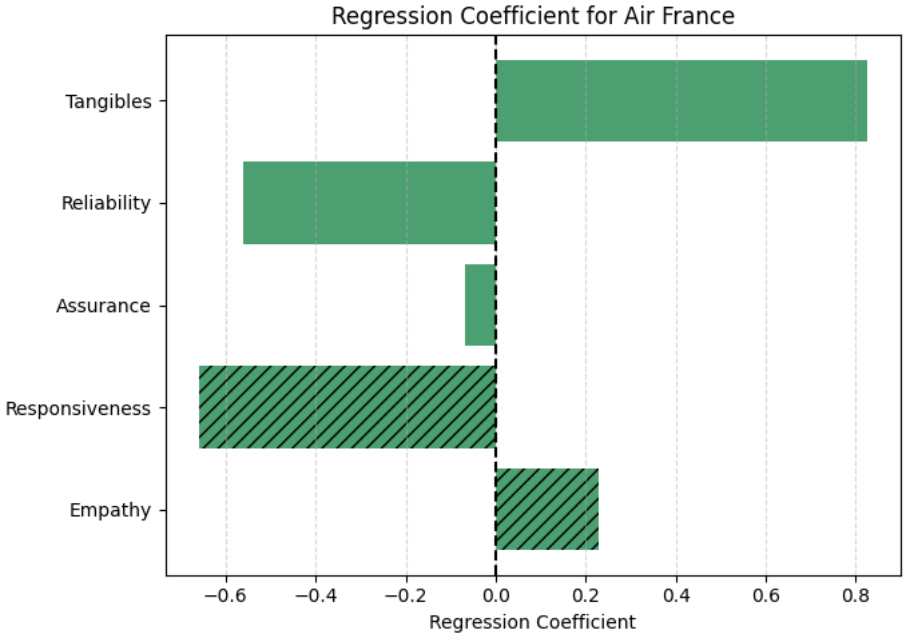

Figure B4: OLS regression analysis of service quality dimensions in Air France Airlines

Appendix C: Analysis of KLM Airline

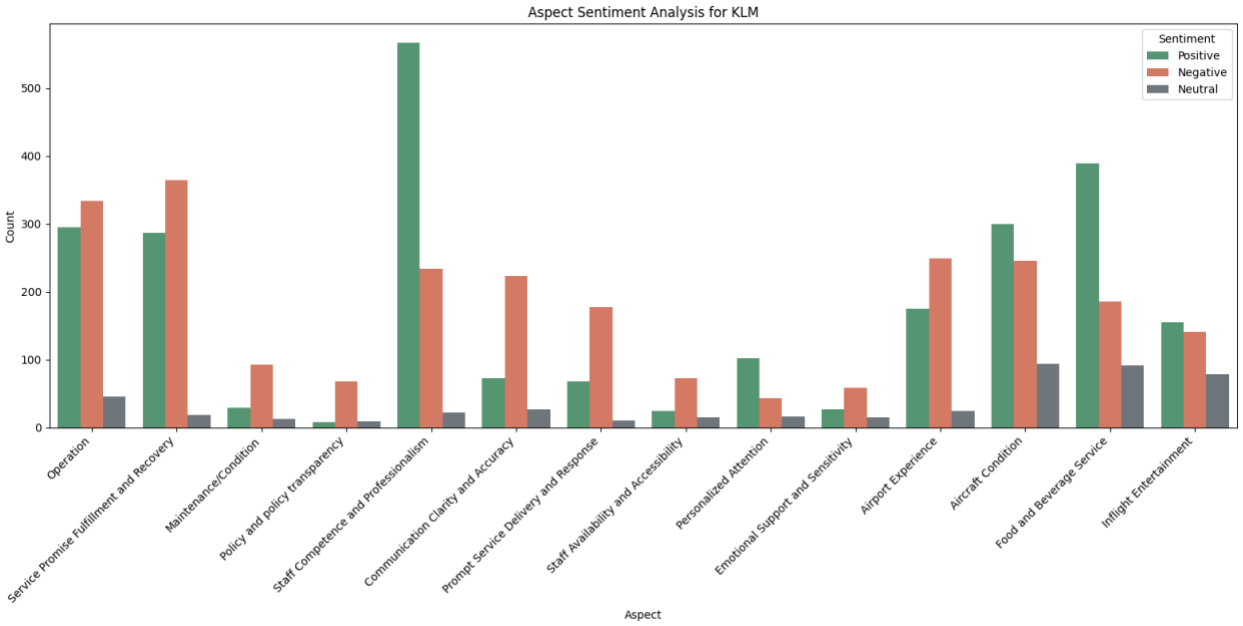

Figure C1: ABSA of service quality themes in KLM Airline

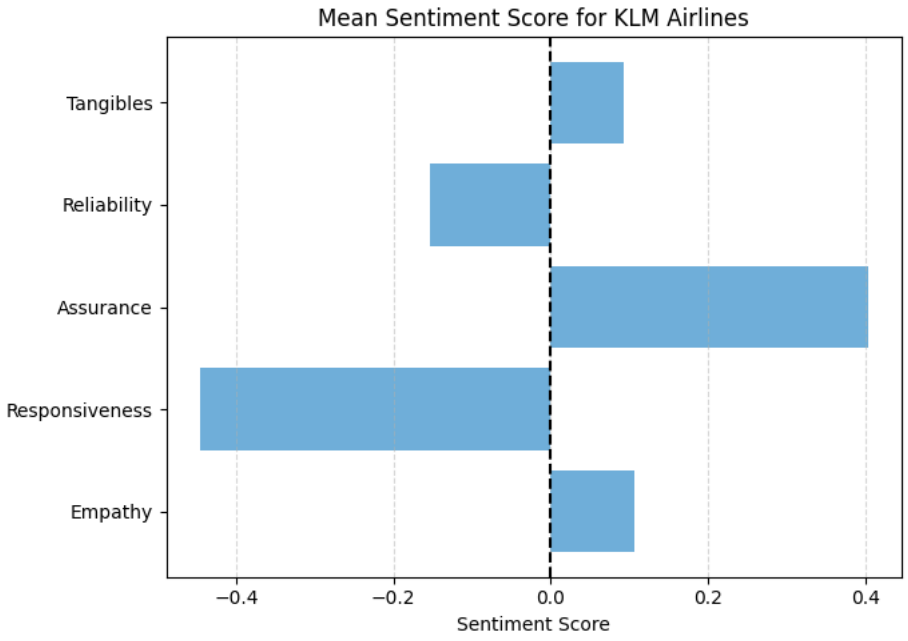

Figure C2: Mean sentiment scores of service quality dimensions in KLM Airline

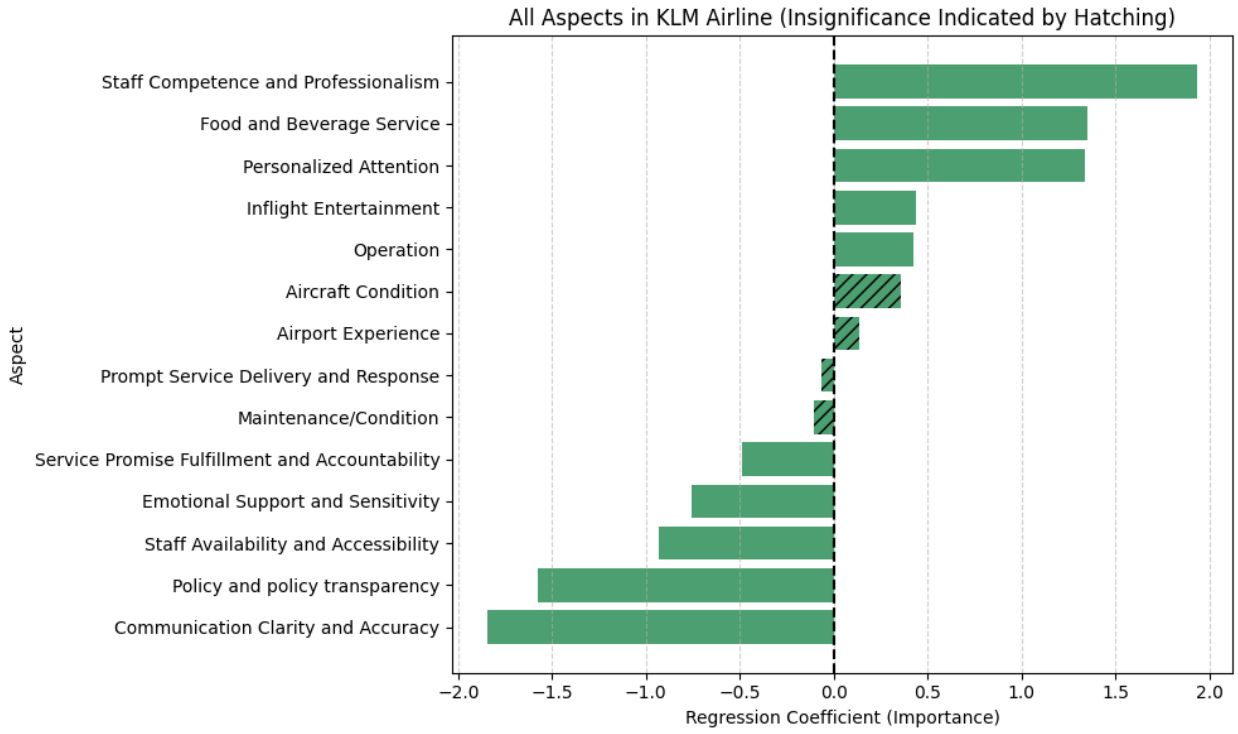

Figure C3: OLS regression analysis of all service themes in KLM Airline with indication of statistical insignificance.

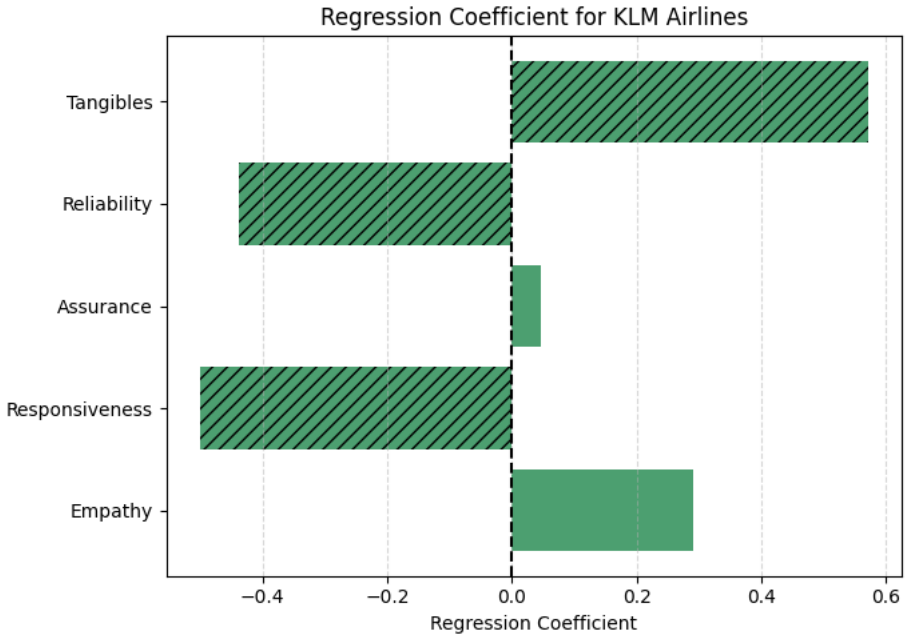

Figure C4: OLS regression analysis of service quality dimensions in KLM Airlines

Appendix D: Analysis of Finnair Airline

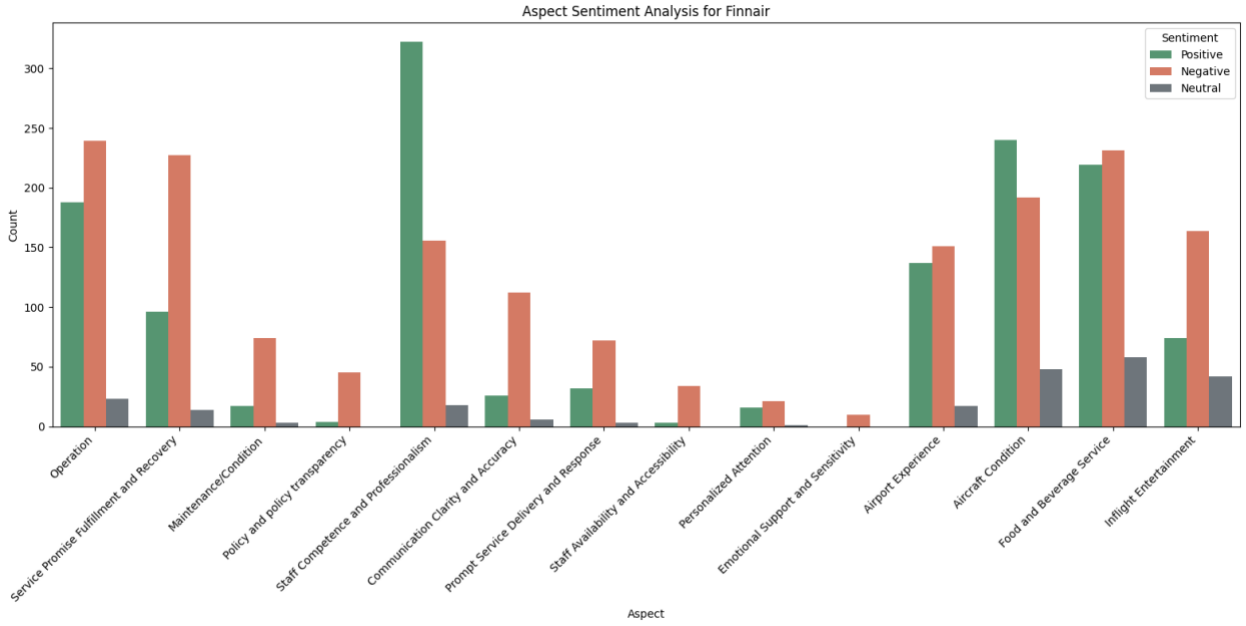

Figure D1: ABSA of service quality themes in Finnair Airline

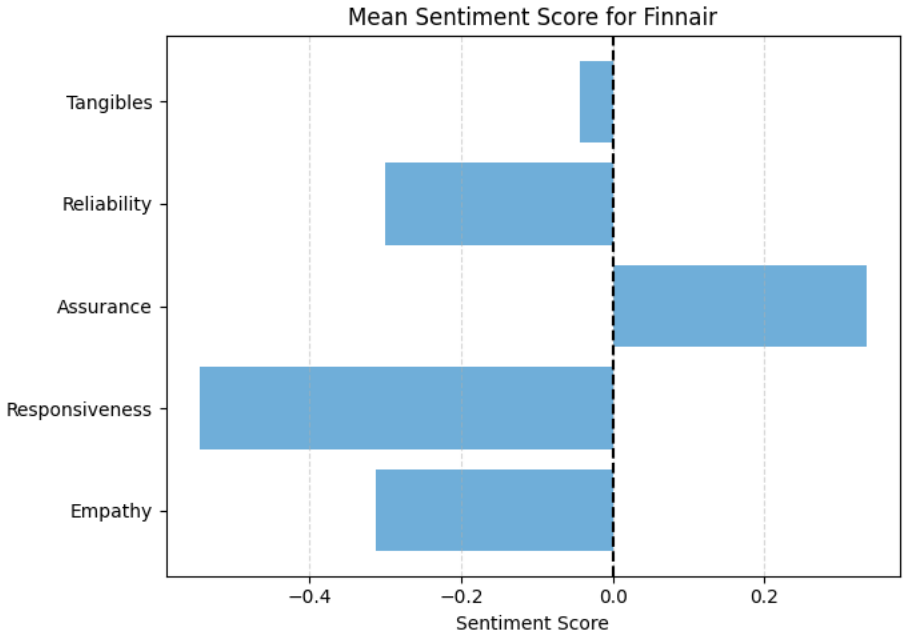

Figure D2: Mean sentiment scores of service quality dimensions in Finnair Airline

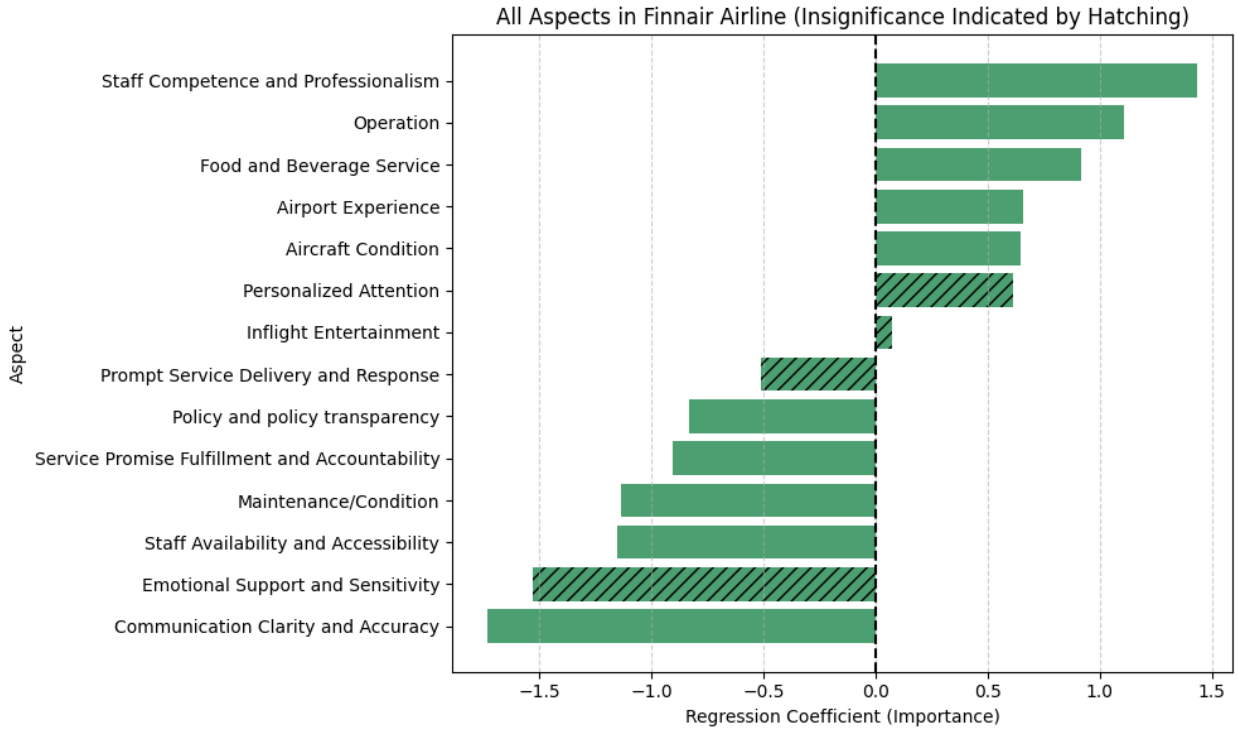

Figure D3: OLS regression analysis of all service themes in Finnair Airline with indication of statistical insignificance.

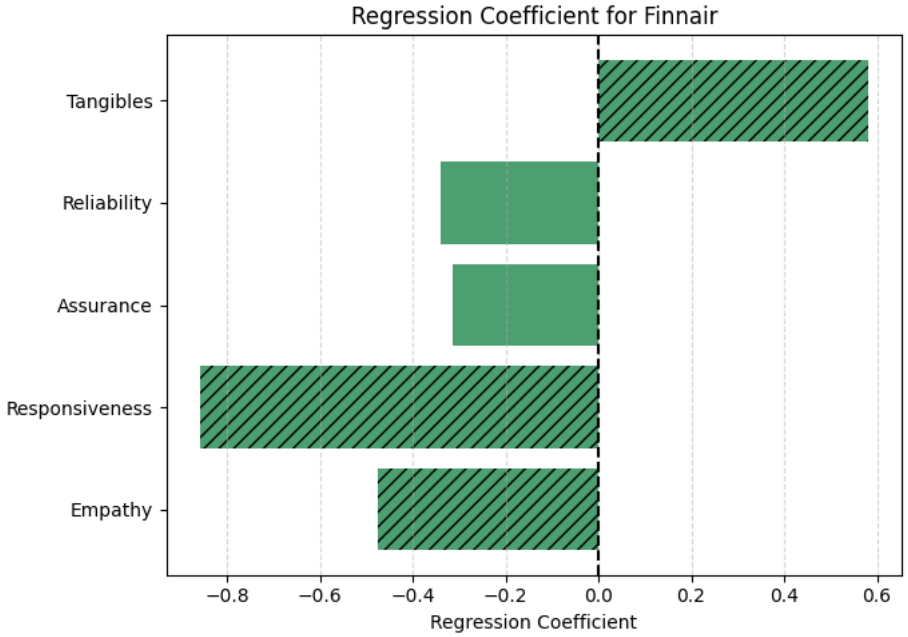

Figure D4: OLS regression analysis of service quality dimensions in Finnair Airlines

Appendix E: Analysis of SAS Airline

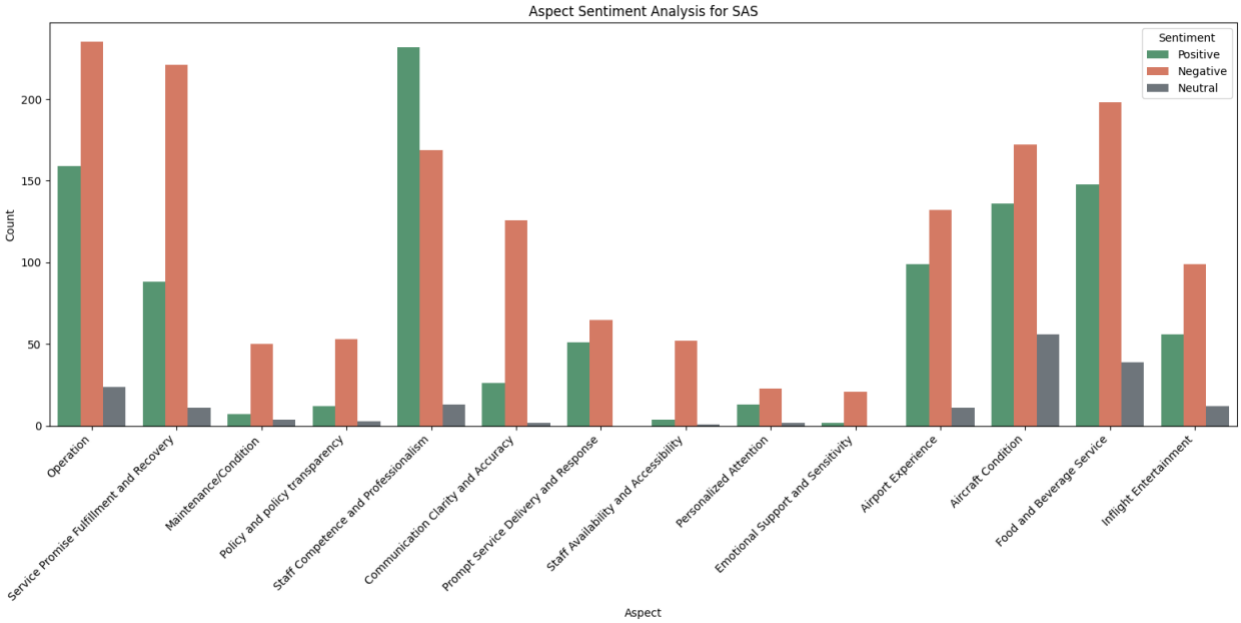

Figure E1: ABSA of service quality themes in SAS Airline

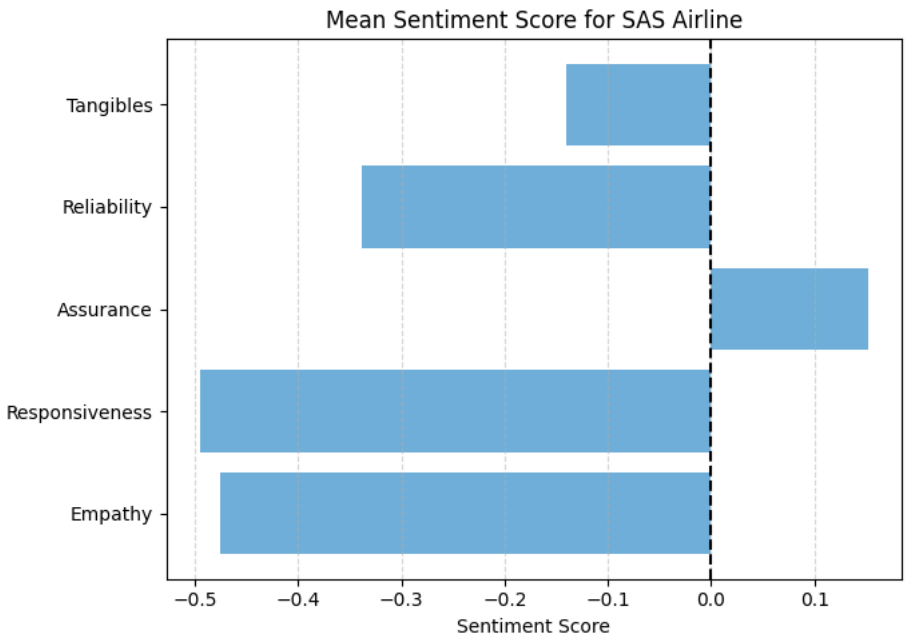

Figure E2: Mean sentiment scores of service quality dimensions in SAS Airline

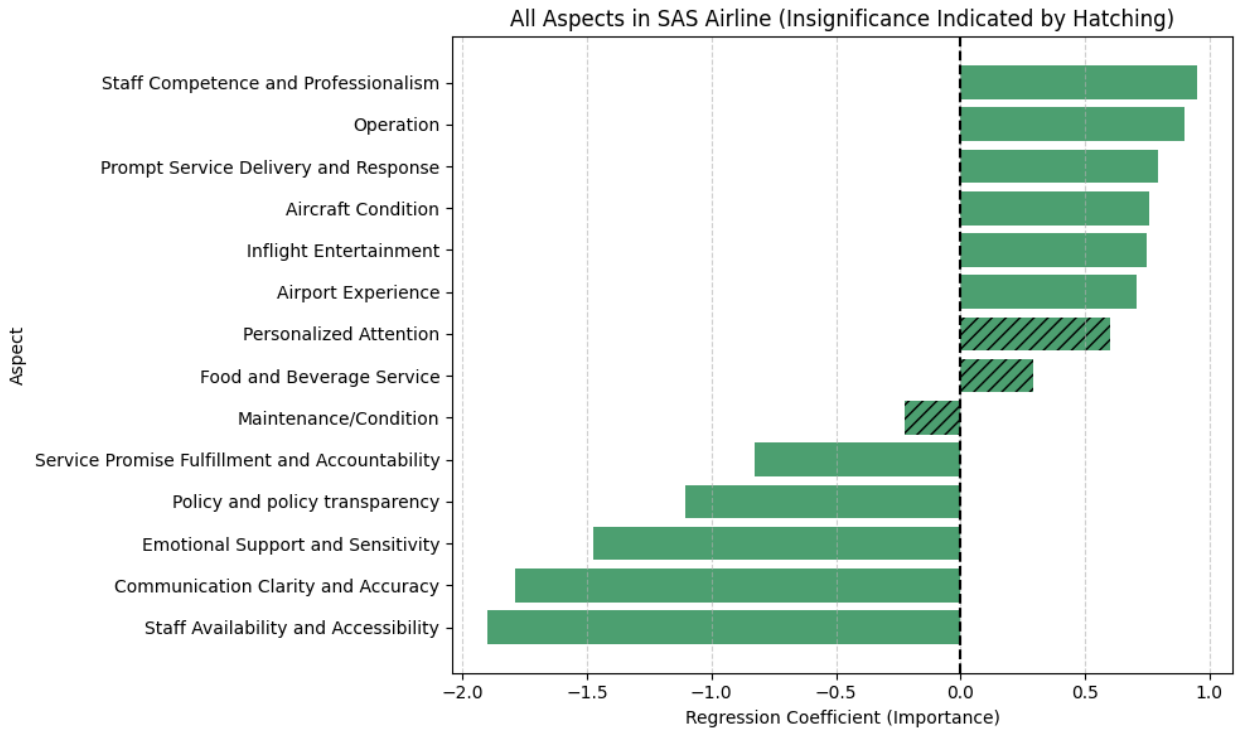

Figure E3: OLS regression analysis of all service themes in SAS Airline with indication of statistical insignificance.

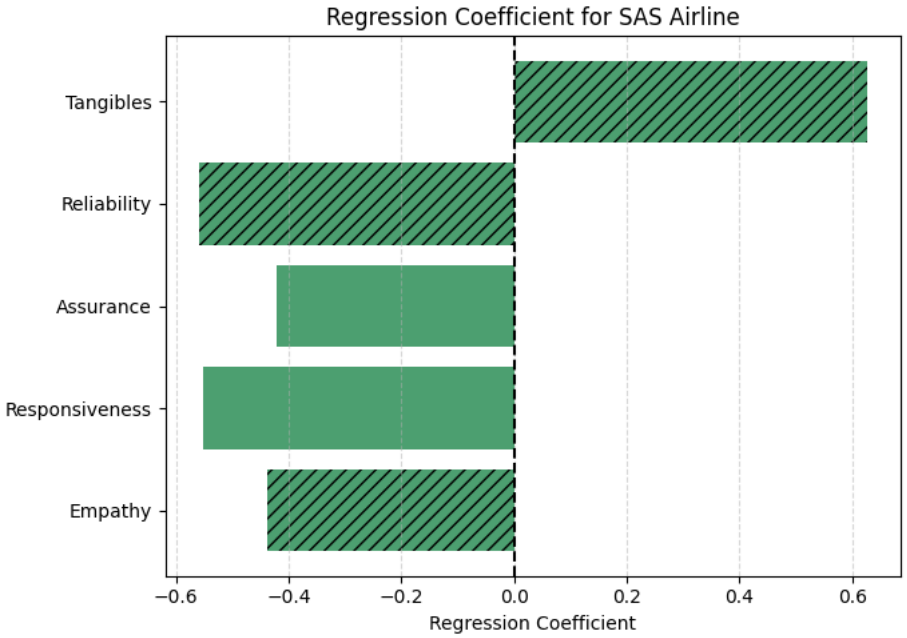

Figure E4: OLS regression analysis of service quality dimensions in SAS Airlines

Appendix F: Analysis of Swiss Airlines

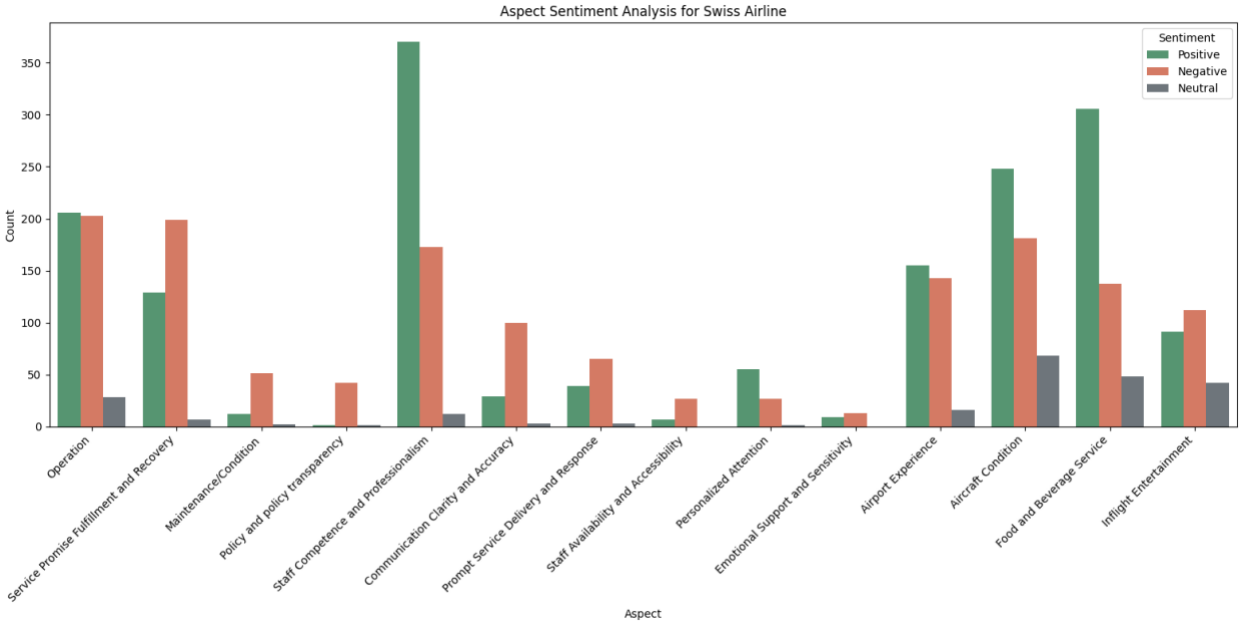

Figure F1: ABSA of service quality themes in Swiss Airlines

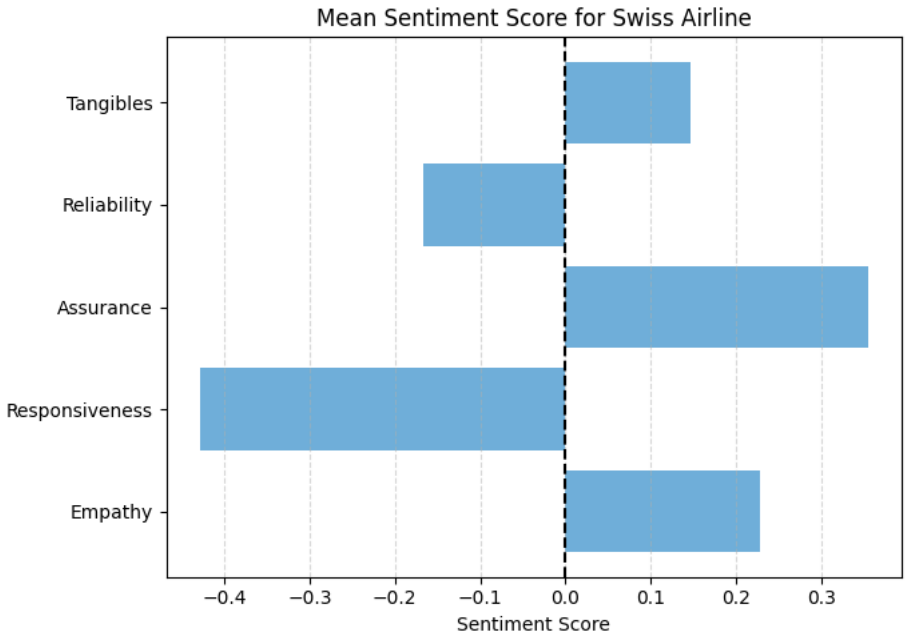

Figure F2: Mean sentiment scores of service quality dimensions in Swiss Airlines

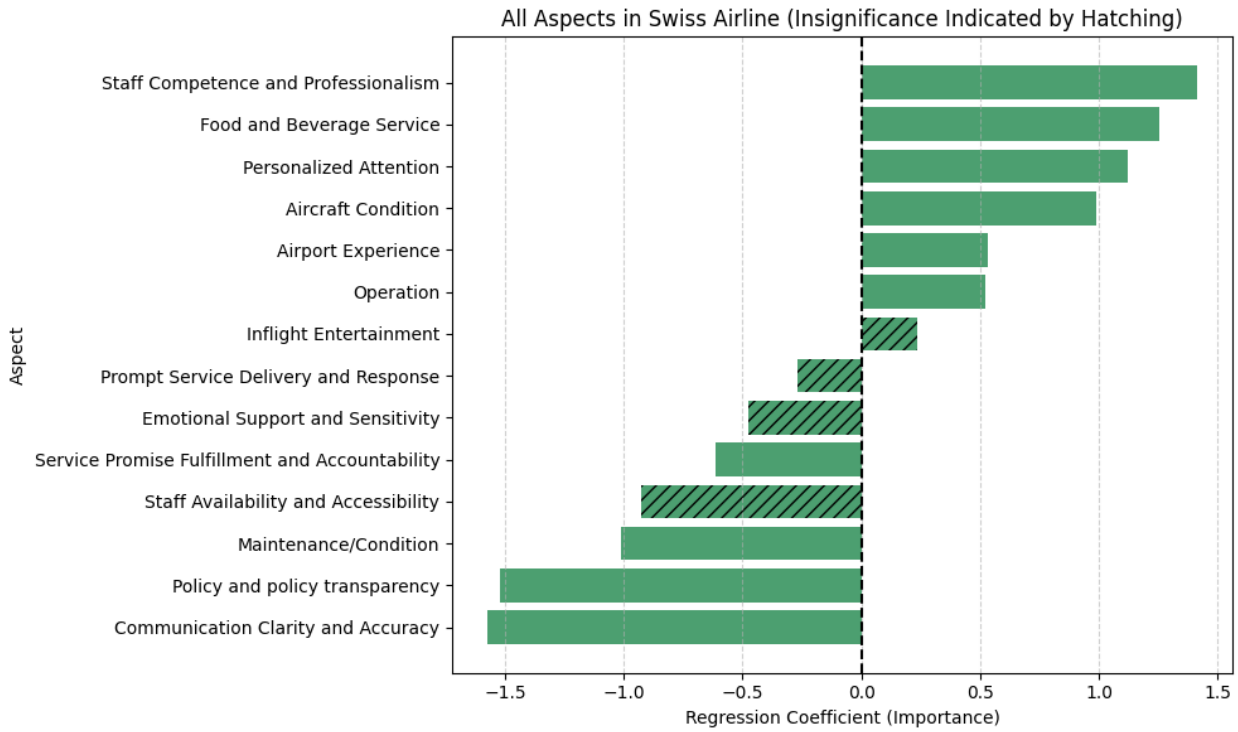

Figure F3: OLS regression analysis of all service themes in Swiss Airlines with indication of statistical insignificance.

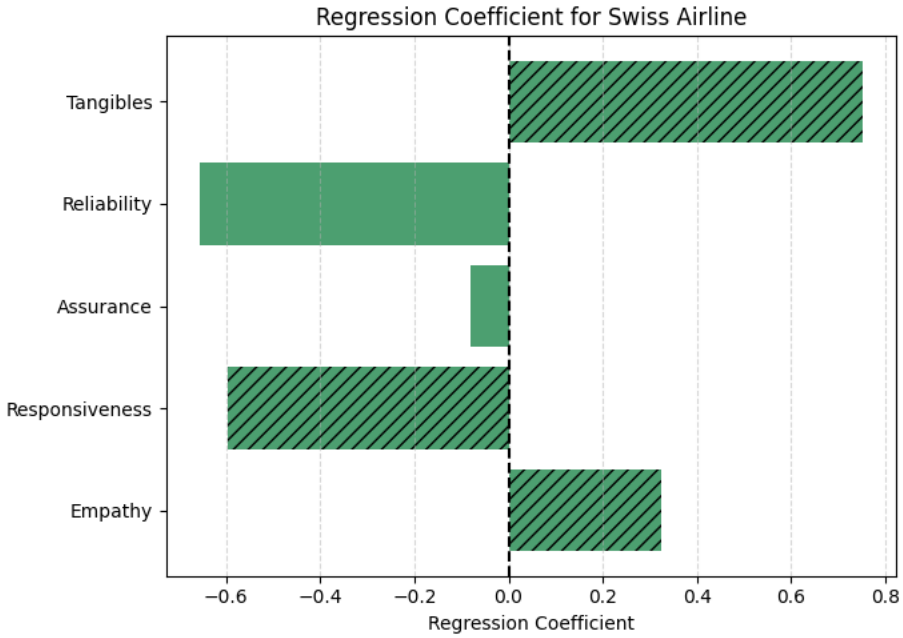

Figure F4: OLS regression analysis of service quality dimensions in Swiss Airlines

Appendix G: Analysis of TAP Portugal Airline

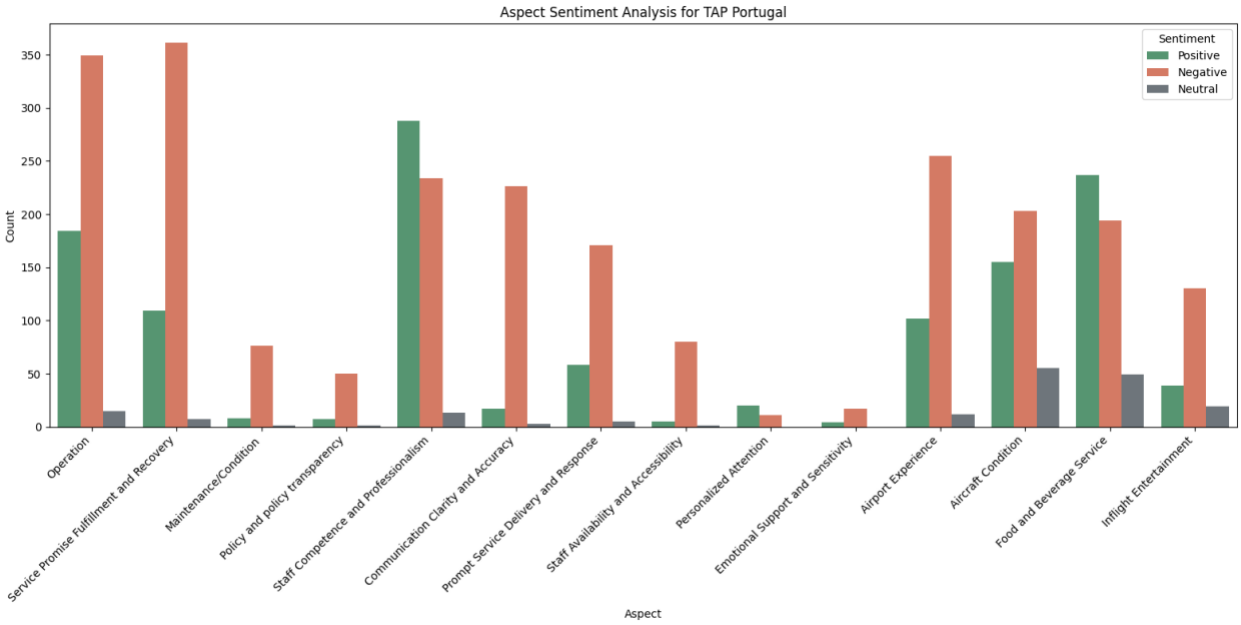

Figure G1: ABSA of service quality themes in TAP Portugal Airline

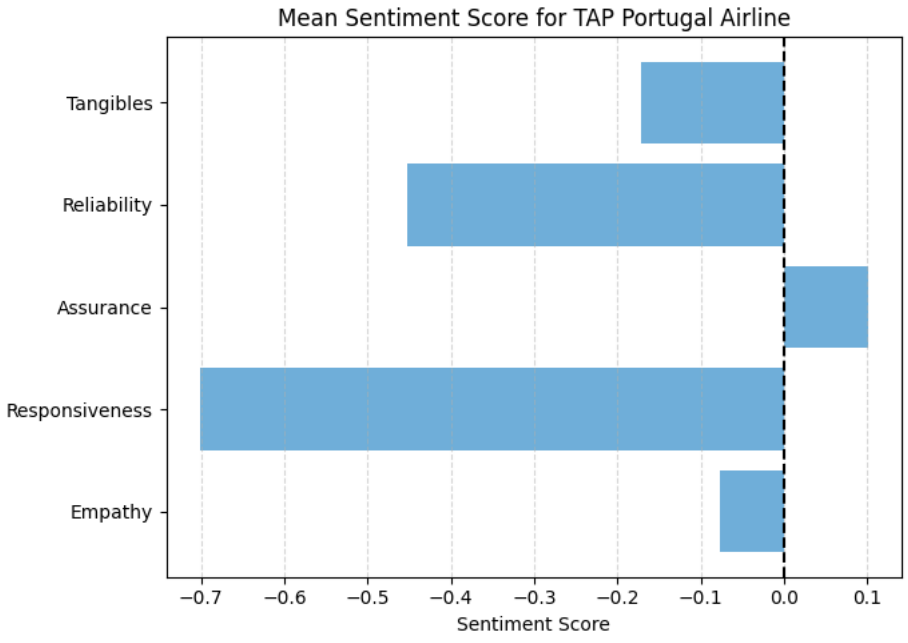

Figure G2: Mean sentiment scores of service quality dimensions in TAP Portugal Airline

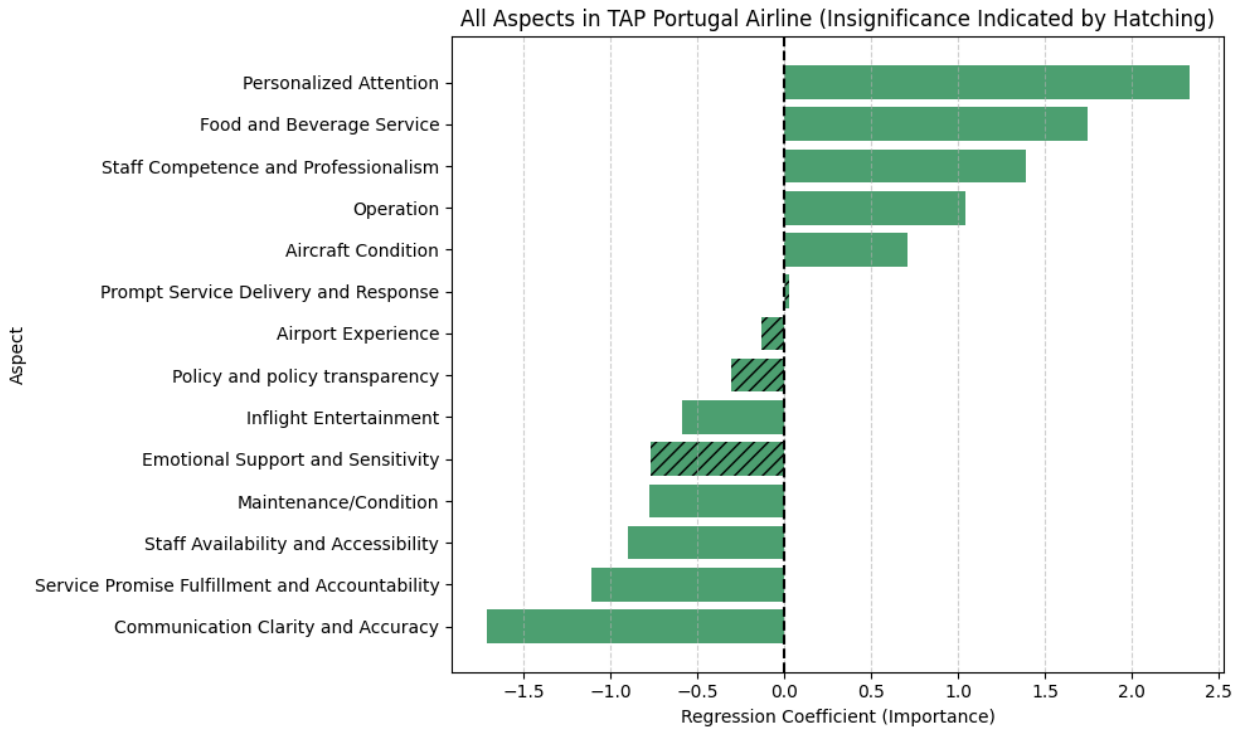

Figure G3: OLS regression analysis of all service themes in TAP Portugal Airline with indication of statistical insignificance.

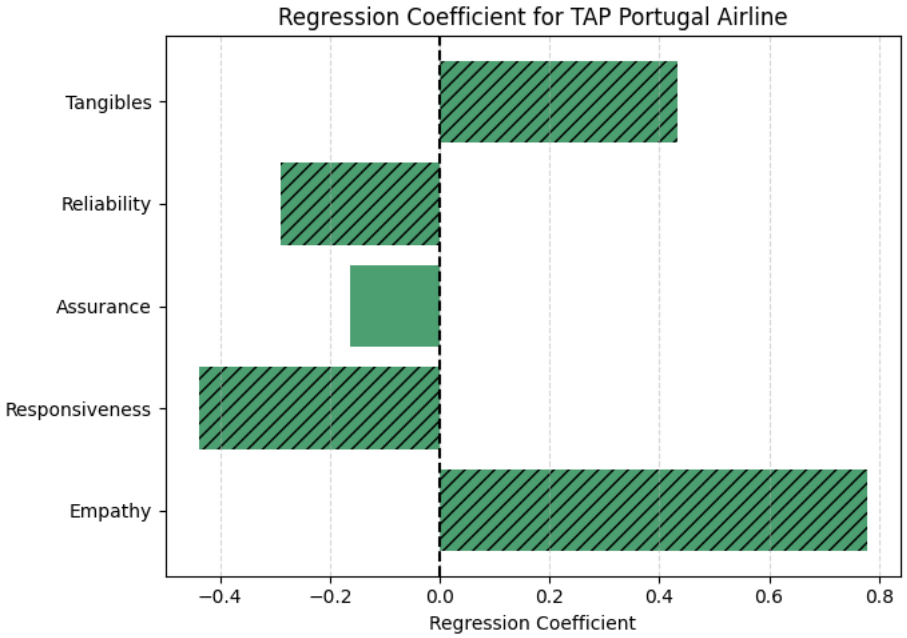

Figure G4: OLS regression analysis of service quality dimensions in TAP Portugal Airline

Appendix H: Analysis of Norwegian Airlines

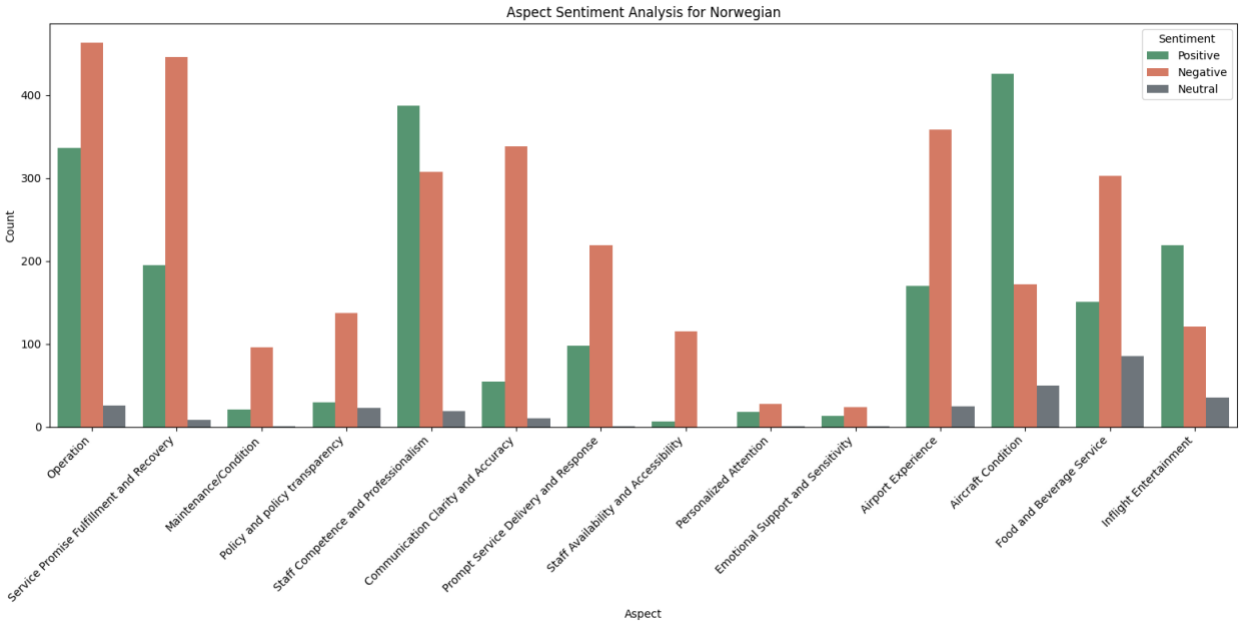

Figure H1: ABSA of service quality themes in Norwegian Airlines

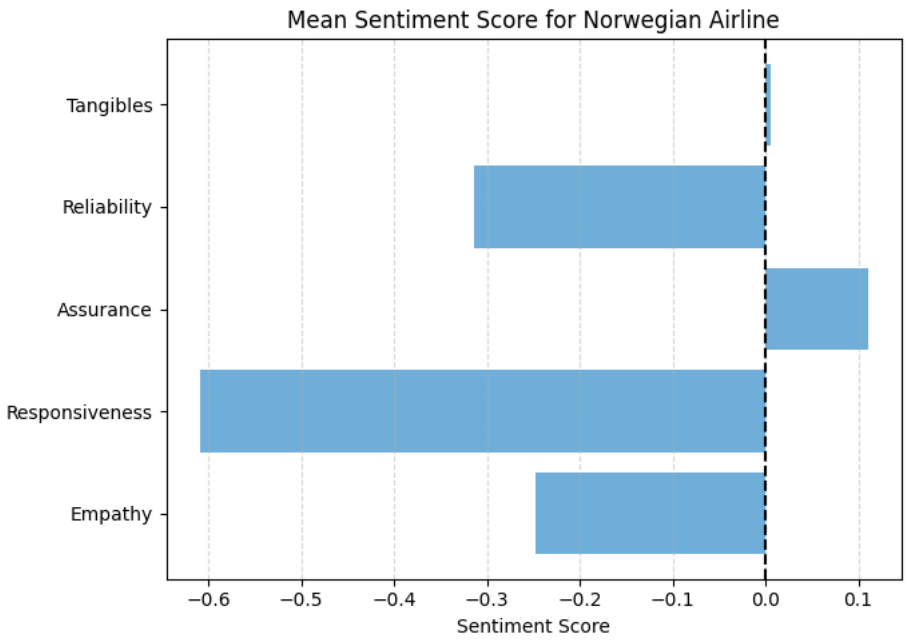

Figure H2: Mean sentiment scores of service quality dimensions in Norwegian Airlines

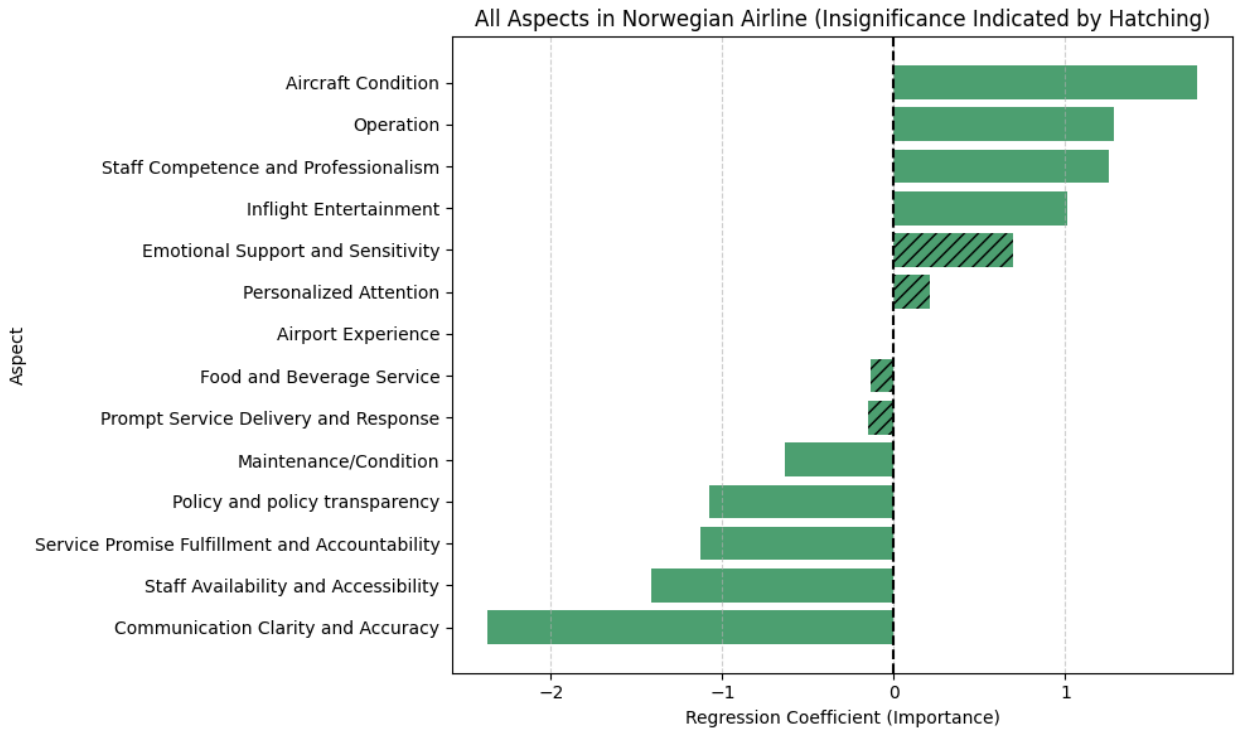

Figure H3: OLS regression analysis of all service themes in Norwegian Airlines with indication of statistical insignificance.

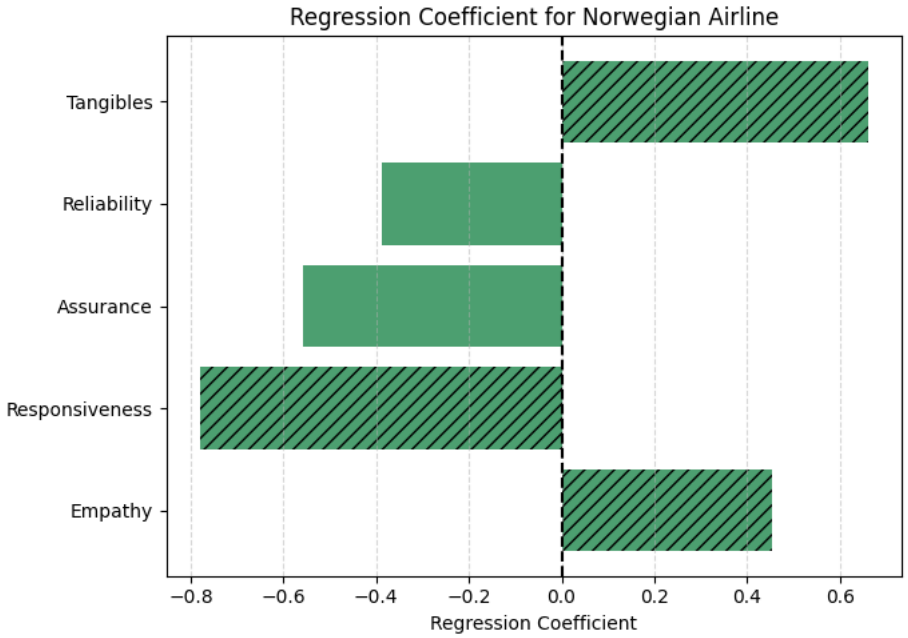

Figure H4: OLS regression analysis of service quality dimensions in Norwegian Airlines

Appendix I: Analysis of Ryanair

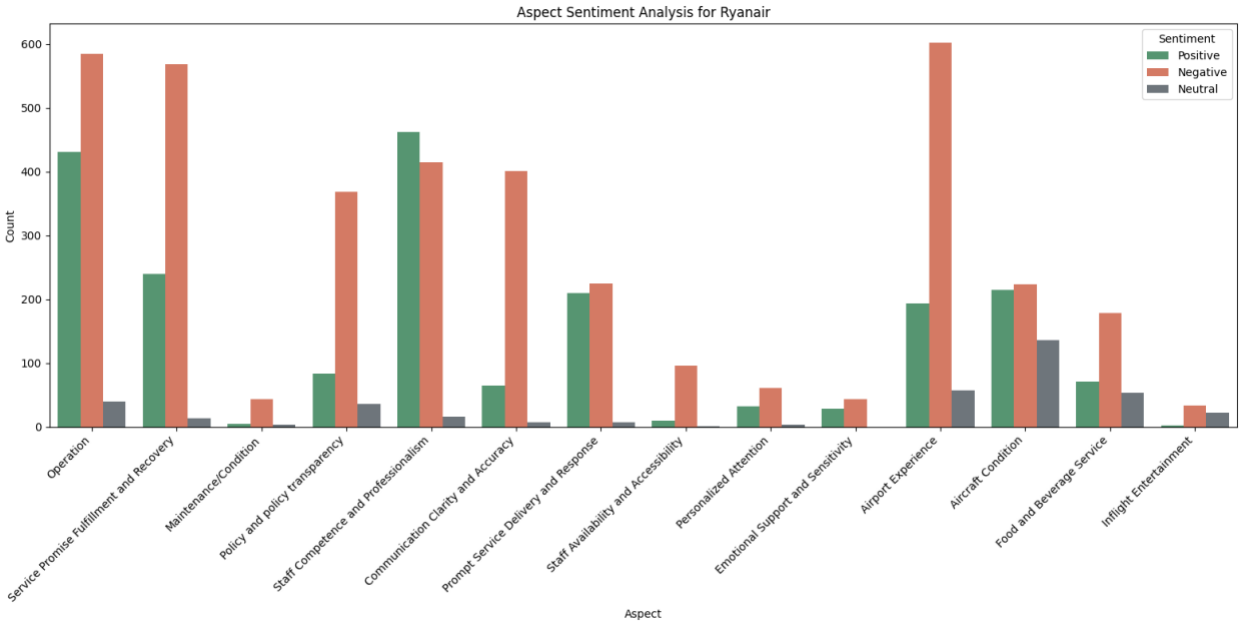

Figure I1: ABSA of service quality themes in Ryanair

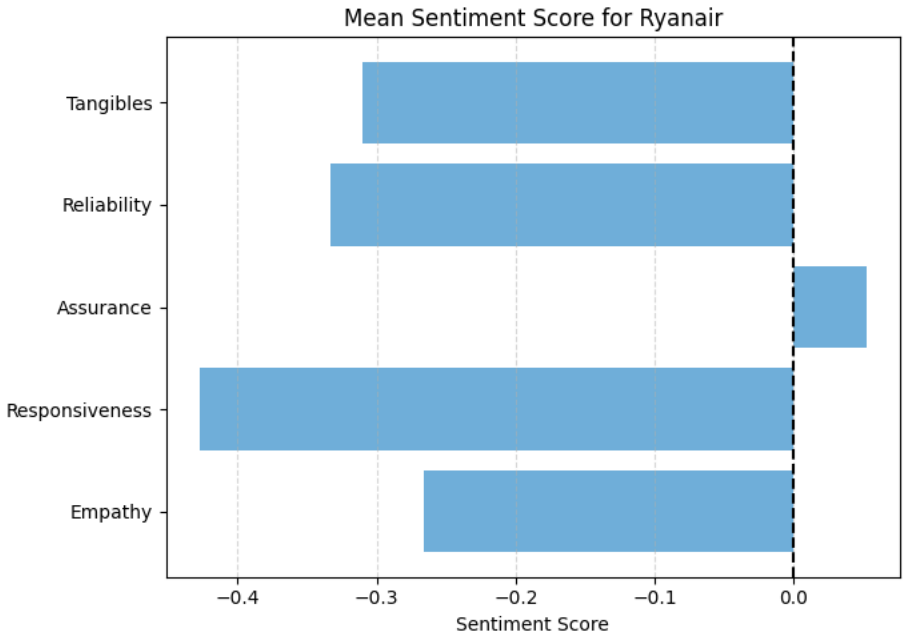

Figure I2: Mean sentiment scores of service quality dimensions in Ryanair

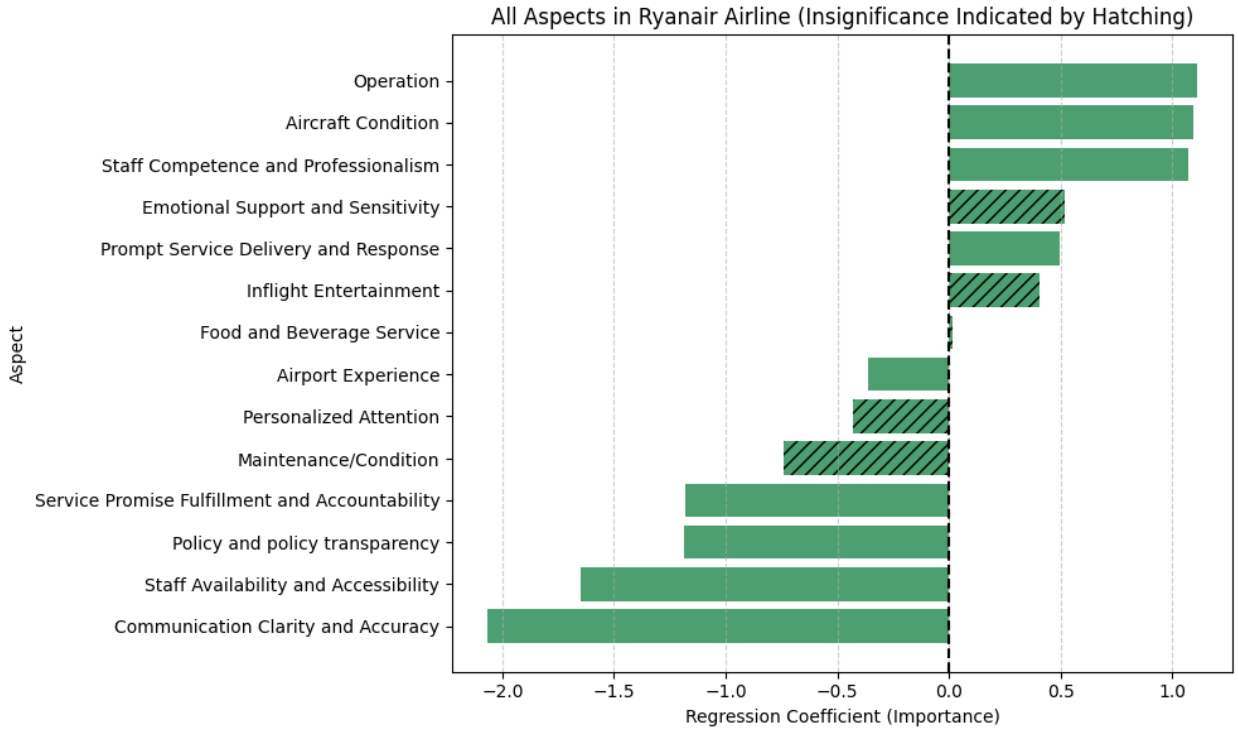

Figure I3: OLS regression analysis of all service themes in Ryanair with indication of statistical insignificance.

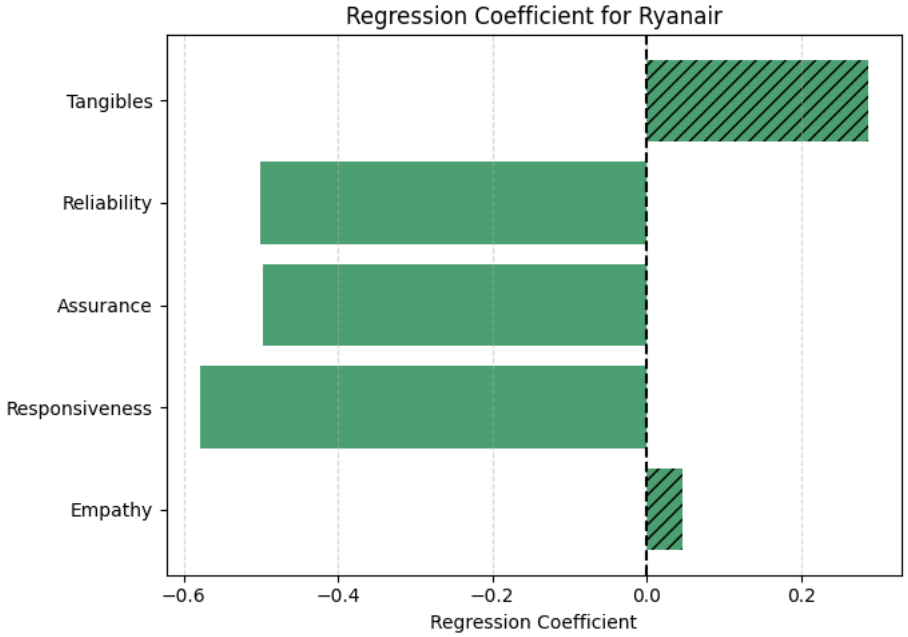

Figure I4: OLS regression analysis of service quality dimensions in Ryanair

Appendix J: Analysis of EasyJet Airline

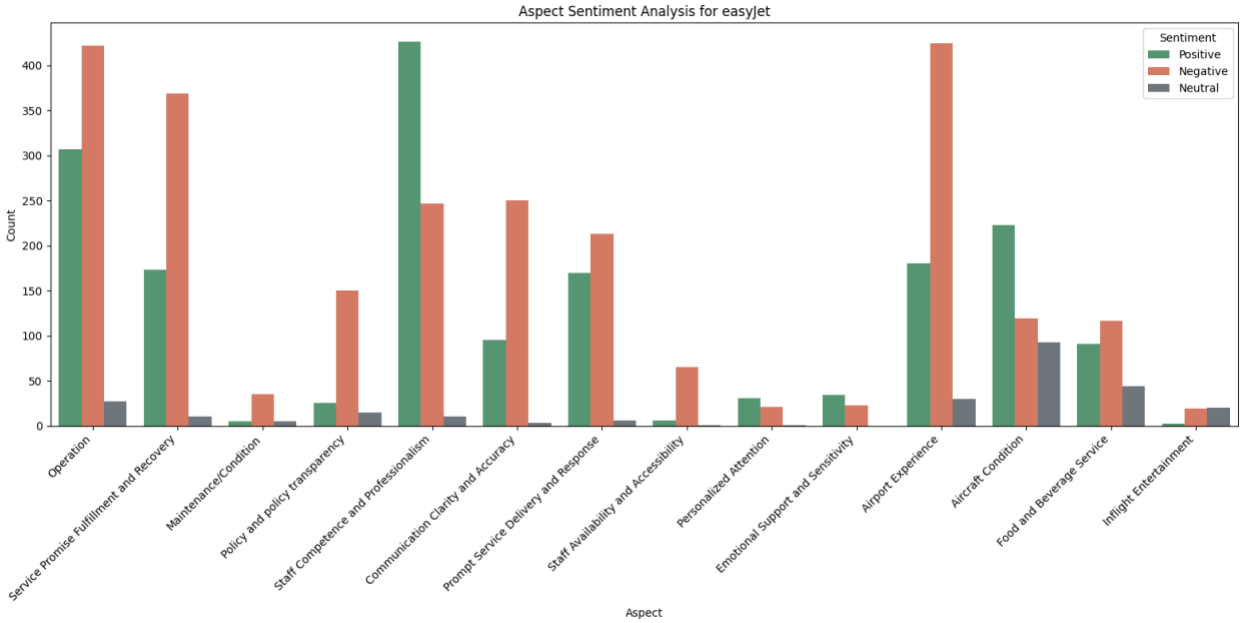

Figure J1: ABSA of service quality themes in EasyJet Airline

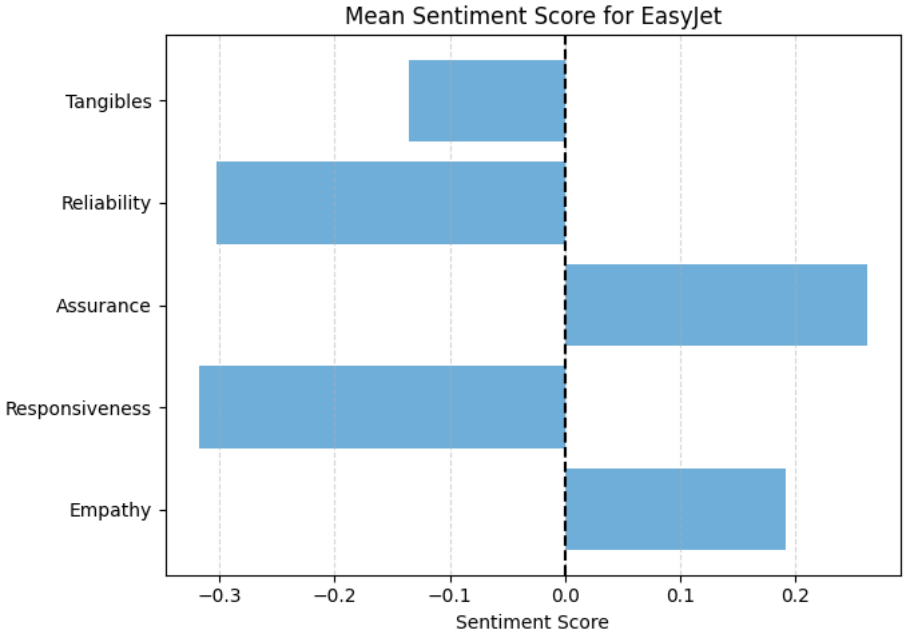

Figure J2: Mean sentiment scores of service quality dimensions in EasyJet Airline

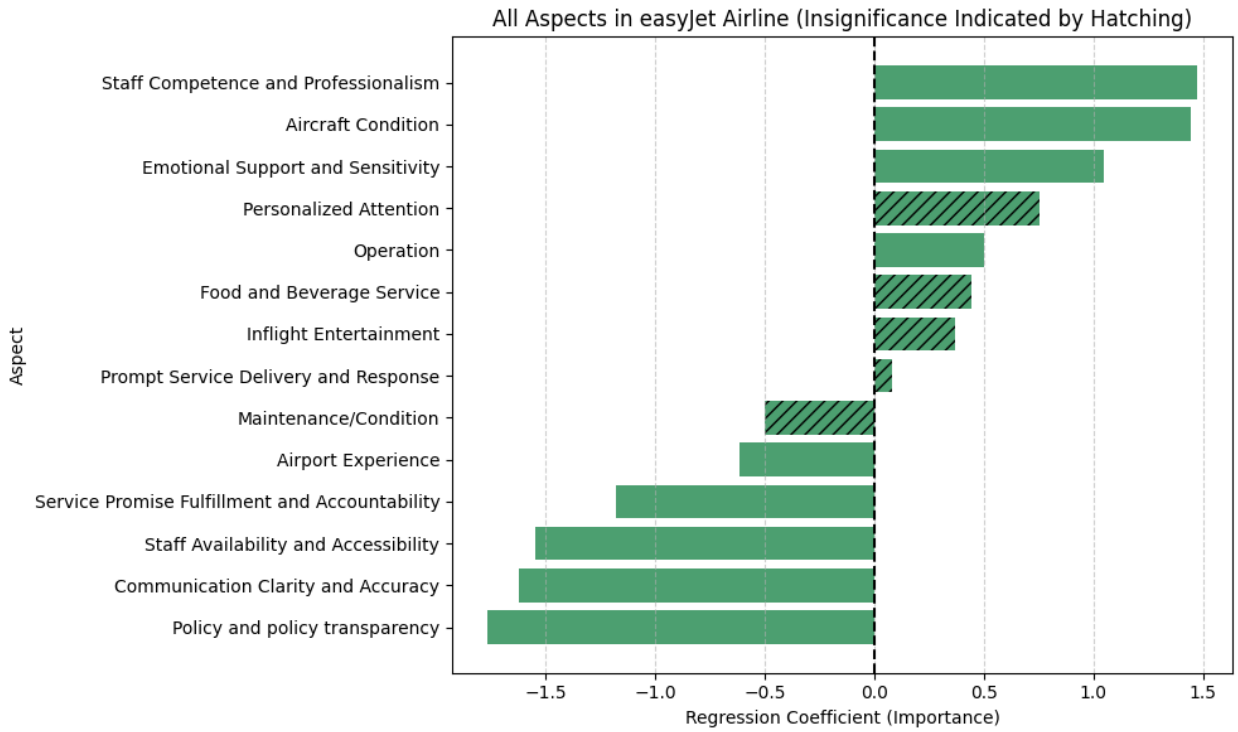

Figure J3: OLS regression analysis of all service themes in EasyJet Airline with indication of statistical insignificance.

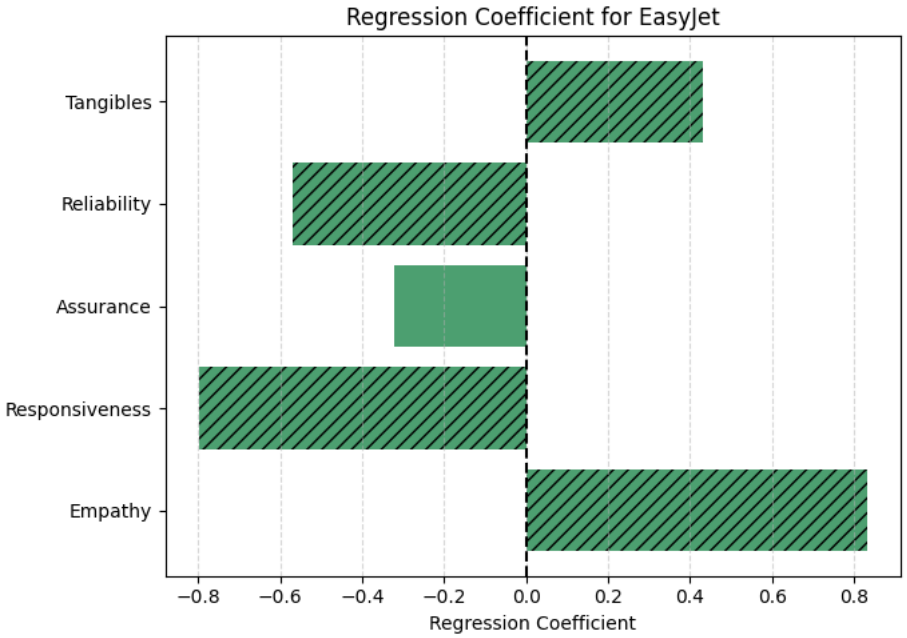

Figure J4: OLS regression analysis of service quality dimensions in EasyJet Airline

Appendix K: Number of analyzed online reviews for each airline

Airline	Number of Reviews
British Airways	1,620
Ryanair	1,358
Lufthansa	1,354
Norwegian	1,028
KLM	909
Air France	878
easyJet	936
TAP Portugal	728
Swiss Airline	670
Finnair	634