Analysing Hotel Reviews: A Text-Based Approach to Understanding Customer Feedback

by

Ritwik Singh

Student ID: 229266568

Dissertation Submitted as Part of the Requirement for Completing an MSc in Business Analytics (MN50759)

Supervisor: Dr. Maria Battara

Word Count: 14,497

Date: 01st September 2024

Acknowledgement

I would like to express my sincere gratitude to my supervisor, Dr. Maria Battara, for their guidance, support, and invaluable feedback throughout this dissertation.

I am also thankful to the faculty and staff of MSc Business Analytics for their assistance and the knowledge they have shared.

I deeply appreciate my peers and colleagues in my department for their collaboration and encouragement.

To my family and friends, especially my parents, thank you for your unwavering support and understanding during this journey.

Thank you all for making this dissertation possible.

Abstract

This dissertation explores global trends in hotel reviews using advanced text analysis techniques to understand customer sentiment and preferences. With the increasing importance of online reviews in the hospitality industry, this study aims to analyse textual data from hotel reviews across various countries to identify common themes, sentiments, and cultural influences on customer feedback.

To achieve this, a mixed-method approach was employed, combining sentiment analysis, topic modelling, and geospatial clustering. Data was collected from the Kaggle dataset "515K Hotel Reviews Data in Europe" by Jiashen Liu, which consists of reviews from the Booking.com website. This dataset provides a comprehensive collection of hotel reviews across Europe. Various text analysis tools were used to process and analyse these reviews.

The findings reveal significant variations in customer sentiment and review patterns across different regions, highlighting the influence of cultural and geographical factors on hotel reviews. Additionally, topic modelling identified key themes frequently appearing in positive and negative reviews, providing insights into customer expectations and satisfaction.

The results of this study contribute to a better understanding of global customer behaviour in the hospitality industry and offer practical implications for hotel management to enhance customer satisfaction and improve service quality.

Contents

List of Figures

List of Tables

1. Introduction

Customer satisfaction within the hospitality industry has increasingly become a focal point for researchers and practitioners alike, as it serves as a critical determinant of success in a highly competitive environment. Satisfied customers has long been closely linked to the hospitality business, particularly the hotel industry. The emergence of digital platforms has allowed customers to freely express their experiences, which is highly influential for a hotel's success and image (Xiang et al., 2015). Numerous details are available in these internet reviews, which can be examined to discover more about the viewpoints and fulfilment of customers. It is still challenging to assess and extract actionable insights from the volume of data, though, because the reviews are unstructured (Pang & Lee, 2008).

The dataset used in this study includes Booking.com customer reviews from 2015 to 2017. This period is important because it is considered a “normal” one in the travel and hospitality industries, undisturbed by interruptions to global travel patterns and consumer behaviour (Martins et al., 2018). For comparative studies in different environments, knowing consumer input from this era offers a baseline for average customer expectations and satisfaction (Gössling et al., 2020).

The importance of understanding visitor feedback cannot be overstated, as it directly affects a hotel's reputation and, consequently, its financial performance (Hay, 2024). Online reviews are becoming an essential part of the decision-making process because of their accessibility and extensive use by potential customers (Chevalier & Mayzlin, 2006). Therefore, analysing these reviews provides valuable insights into consumer preferences, expectations, and areas needing improvement. Despite substantial research, there is still an enormous gap in the literature on sentiment analysis and customer satisfaction.

Several research have shown that positive evaluations are generally associated with better satisfaction ratings, but negative assessments usually draw attention to areas where service delivery is lacking (Liu, 2012). In the analysis of customer fulfilment, sentiment analysis has become an important tool. However, a significant amount of prior research has focused on components of sentiment analysis, including the polarity of reviews, largely ignoring the overall context of review features (Pang & Lee, 2008). The goal of this dissertation is to close this gap by offering a comprehensive analysis of customer satisfaction across various regions, considering not only sentiment scores but also other important variables such as reviewer demographics, length, and frequency of specific keywords.

Accurately analysing the sentiments stated in reviews is a difficult process, which presents a significant obstacle for this research. Contextual implications, linguistic challenges, and cultural differences can all complicate sentiment analysis (Thelwall et al., 2010). Moreover, it is important to employ robust data processing techniques as bias may arise from differences in the volume and quality of evaluations among different hotels (Zhou et al., 2019). Modern machine learning algorithms present substantial hurdles when processing large datasets, requiring the use of efficient data management solutions.

It is anticipated that hotel management would benefit greatly from the research's findings, which will give them practical knowledge about client preferences and opportunities for development. This dissertation adds to the body of knowledge already available on customer satisfaction by utilising sophisticated analytics and machine learning. It also provides a foundation for further research in this area. The knowledge gathered from this research may potentially have larger implications for other service-oriented sectors where customer input is vital in determining corporate strategy.

The research holds significance as it can improve decision-making in the hospitality sector. Understanding the fundamental factors that contribute to positive or negative customer experiences becomes increasingly important as customer reviews continue to play a significant part in influencing public opinion and driving business performance (Anderson, 2012). This dissertation offers a fresh perspective on customer satisfaction by combining machine learning and sentiment analysis. Ultimately, this study can help develop more focused marketing plans, customised customer experiences, and increased brand loyalty, all of which can boost overall company performance in the fiercely competitive hospitality sector.

Additionally, the study will lay the groundwork for future research into the use of machine learning to analyse consumer feedback in various businesses. The approach and conclusions of this dissertation could be modified for use in other settings where the importance of consumer input is similar, such retail or healthcare (Nguyen et al., 2019). This study advances our knowledge of the connection between sentiment and satisfaction, which benefits the wider field of customer experience management by providing insightful information that is useful in a variety of industries.

2. Literature Review

The exponential growth of user-generated content on digital platforms has revolutionized the landscape of hospitality and tourism management (Leung et al., 2013). Online reviews have become essential elements in consumers' decision-making processes, impacting their perceptions and booking habits in a time when digital interactions are frequently occurring (Ye, Law and Gu, 2009). With a special emphasis on Booking.com, this literature review summarises the key findings and concepts from relevant academic papers and research projects on consumer reviews and rating systems. It focusses into how user-generated content has changed over time in the travel and hospitality sector, as well as the approaches used in earlier studies and the effects of online reviews on customer behaviour and company outcomes. This review offers a thorough overview of the present state of research and proposes areas for additional exploration by highlighting the significance of earlier findings. These revelations served as the basis for the dissertation, which intends to explore the nuances of subject and sentiment analysis of customer reviews on Booking.com in further detail to provide an integrated method for comprehending and improving consumer satisfaction in the hospitality sector.

2.1 Historical Context and Evolution

From the early days of the internet, there has been a long history of analysing customer feedback. Online travel forums started to appear in the early 2000s, enabling users to discuss and exchange experiences on accommodation and other travel-related services (Vermeulen & Seegers, 2009). With the advent of all these platforms, a new era in hospitality and tourism management began, one in which consumer opinion played a crucial role in determining how businesses should be run.

Ratings and reviews from customers soon proved to be very helpful information sources. In contrast to marketing teams produced promotional content, these reviews offered real, unscripted input from customers. Because of their genuineness, the reviews gained a lot of credibility and were quite reliable, providing unbiased information about the advantages and disadvantages of lodging. According to Leung et al. (2013), hotel management can use genuine client reviews as a crucial tool to enhance their services.

The significance of customer reviews has grown over time. According to (Phillips et al., 2016), reviews on sites such as Booking.com now include thorough evaluations of a range of hotel

experience factors, such as location, facilities, level of service, quality of accommodation, and overall value. A quantifiable indicator of overall satisfaction is provided by the variety of ratings that are frequently included with these evaluations. According to Ye et al. (2009), a thorough grasp of consumer experiences is provided by the integration of qualitative and quantitative data.

2.2 The Importance of Online Reviews and Rating Systems

Positive reviews can significantly enhance a hotel's reputation, attract new customers, and increase bookings (Vermeulen & Seegers, 2009). Conversely, negative reviews can deter potential customers and damage the hotel's reputation. Empirical studies have demonstrated that online reviews influence customer decision-making processes and can affect the financial performance of hotels (Mariani & Borghi, 2018). Therefore, analysing these reviews is critical for hotel management to improve service quality and customer satisfaction (Anderson, 2012).

The influence of online reviews extends beyond individual customer experiences. Aggregated review scores and the overall sentiment expressed in reviews can substantially impact a hotel's ranking on booking platforms like Booking.com. Higher rankings enhance visibility, which in turn can lead to increased booking rates. This makes it imperative for hotels to maintain positive reviews and high ratings (Öğüt & Onur Taş, 2012).

Additionally, online reviews reassure prospective guests about a hotel's quality and dependability by acting as a type of social evidence. This is especially crucial in the very competitive hospitality sector, where customers have a wide range of choices. As a result, monitoring and addressing internet evaluations has become an essential part of hotel administration. While noting good feedback can promote customer loyalty, promptly and thoughtfully responding to unfavourable evaluations might lessen their impact (Mariani & Borghi, 2018).

The establishment of confidence between the hotel and prospective guests is also greatly aided by online reviews. When potential customers see a lot of great reviews, they become more confident in the hotel and are more inclined to make a reservation. On the other hand, an abundance of unfavourable reviews or an absence of positive reviews may cast doubt on a product or service and encourage potential buyers to look elsewhere. The previously mentioned dynamic highlights the imperative for hotels to not only deliver exceptional customer care but

also foster post-stay customer engagement to incentivise positive online experience sharing (Leung et al., 2013).

The impact of rating systems on customer reviews and hotel performance is another critical area of research. Mariani and Borghi (2018) examined how the Booking.com rating system influences hotel performance. Their findings indicate that customer evaluations and ratings heavily affect booking decisions and hotel revenues. Understanding the factors that drive these ratings, such as room cleanliness, service quality, and value for money, can help hotels enhance their services and improve their ratings, ultimately boosting their financial performance.

Ratings influence prospective guests' expectations and offer a quantitative indicator of customer satisfaction. Increased trust and greater booking rates are typically correlated with higher ratings. Öğüt and Onur Taş (2012) discovered, for example, that hotels with higher ratings on Booking.com saw noticeably increased booking rates, underscoring the significance of preserving high ratings.

Rating systems have an impact on hotel operations management as well. A positive feedback cycle where great customer satisfaction generates high ratings, which in turn draw more guests and increase bookings, is created when hotels are incentivised to maintain high levels of service and amenities (Mariani & Borghi, 2018). On the other hand, concentrating only on ratings while neglecting to solve underlying service problems may cause a discrepancy between what customers expect and what they really receive, which could negatively impact the hotel's standing and bottom line (Vermeulen & Seegers, 2009).

2.3 Overview of Review Platforms

Websites with customer reviews, like Yelp, Expedia, Booking.com, TripAdvisor, and Airbnb, have become essential to the hospitality sector. Every platform has advantages. For example, TripAdvisor is well known for its thorough coverage of lodging, dining, and attractions, which makes it a useful tool for visitors looking for in-depth information (O'Connor, 2010). Yelp is well-known for its community-driven reviews of local businesses, including restaurants and services. Travellers can enjoy a seamless experience with Expedia as it integrates booking capabilities with customer feedback (Browning et al., 2013).

These platforms provide a range of useful data types for analysis:

- Ratings: Numerical ratings of the hotel experience based on several factors, including staff, value for money, cleanliness, comfort, and location.
- Customer Reviews: Detailed written reviews that use the customers' own words to describe their experience.
- Owner Responses: Hotel management's comments to reviews from guests reveal how hotels handle criticism.
- Metadata: Details include the review date, the kind of traveller (family, alone, etc.), and the duration of stay.

An extensive examination of consumer feedback is made possible by these data types. While textual evaluations offer deep insights into individual features that customers like or criticise, ratings offer a rapid overview of customer satisfaction across multiple characteristics of the hotel experience (Filieri, 2015). Owner comments can disclose the hotel's commitment to providing excellent customer service as well as how well they manage compliments and complaints (Gössling et al., 2018).

TripAdvisor has attracted significant attention for its ability to impact traveller decisions because of the vast number of reviews and in-depth customer feedback it offers. The platform is a major participant in the online review ecosystem due to its large user base and quantity of information. Yelp has grown to be a major informational resource for tourists, especially in urban areas, despite its initial concentration on eateries and small businesses (Luca, 2016). Its community-driven methodology guarantees reviews with a high degree of participation and authenticity.

Airbnb has grown to be a major force in the hospitality sector by offering an alternative for hotels. Reviews on Airbnb provide a distinctive perspective on what travellers have experienced with a range of accommodations, from studio apartments to full homes (Zervas et al., 2017). The platform's comprehensive review system, which includes ratings on certain areas of the stay, such cleanliness, communication, and the check-in procedure, reflects its emphasis on community and trust (Guttentag, 2013). Studies conducted on Airbnb have demonstrated that customer opinions and booking decisions are significantly shaped by reviews. According to studies, properties that receive better ratings and reviews typically have greater occupancy rates and may charge more money (Ert et al., 2016). Additionally, the presence of detailed reviews and ratings helps build trust between hosts and guests, which is essential for the platform's success (Ert et al., 2016).

Due to its global reach and ability to host reviews from a wide range of nationalities, Booking.com facilitates a more comprehensive investigation of customer preferences and experiences across the globe (Mariani & Borghi, 2018). Additionally, Booking.com is a perfect resource for thorough academic research due to its organised data and multitude of filtering possibilities (Anderson, 2012). Booking.com is unique in that it gathers qualitative information through in-depth text reviews in addition to quantitative information through ratings. Users' reviews are more credible and reliable because the site requires them to have finished their stay before posting one (Mariani & Borghi, 2018). Furthermore, the availability of owner replies adds another level of complexity to the analysable data by offering insights into how hotels interact with customers and maintain their online image (Gössling, Hall and Andersson, 2016).

2.4 Past Research in Tourism Management Related to User Reviews and Rating Systems

Over the past few decades, there has been a tremendous evolution in tourism management, mostly brought about by improvements in technology and the widespread use of the internet. To draw in and keep consumers, tourism management first concentrated on conventional strategies including print media, travel agencies, and word-of-mouth recommendations (Buhalis and Law, 2008). But as digital platforms and the internet have grown in popularity, the sector has changed, becoming more data-driven and agile (Xiang et al., 2015). Researchers started examining the effects of internet reviews on customer behaviour and business performance as they grew more common. Early research centred on figuring out how reviews affected reservations and hotels' general reputation. Vermeulen and Seegers (2009) discovered, for example, that whilst bad reviews can discourage potential guests and harm a hotel's reputation, favourable ratings improve a hotel's image, draw in new business, and increase bookings.

Previous research on user reviews and rating systems in tourism management have aimed to:

- Understand Customer Satisfaction: Determine elements that impact both positive and negative reviews to enhance customer satisfaction and service quality (Phillips et al., 2016).
- Predict Hotel Performance: Mariani & Borghi (2018) use reviews and ratings to forecast revenue and reservation rates.

- Enhance Marketing Strategies: Using information gathered from reviews to focus marketing and draw in more customers (Anderson, 2012).

These studies have advanced the knowledge of the essential role that online reviews play in the hospitality sector. Researchers can find areas for improvement and important drivers of customer satisfaction by assessing customer feedback. This will eventually improve service quality and boost customer loyalty for hotels (Phillips et al., 2016).

Research has also looked at how reviews and ratings alter depending on a country's expectations and degree of satisfaction. For instance, that guests from various areas give different emphasis to various components of their visit, which affects their evaluations. By being aware of these variations, hotels may better serve a wide range of guests and increase their appeal internationally. Language limitations and cultural differences can influence how reviews are written and understood, according to research (Bosangit et al., 2012). This emphasises the necessity for culturally sensitive service techniques.

The total scores from a set of evaluations have a big influence on what consumers decide to buy. Xiang et al. (2015) used big data analytics to investigate how aggregated review scores affected customer booking behaviour. They found that aggregated evaluations, often the primary deciding element, allow guests to quickly and efficiently assess hotel quality. This highlights how important it is to maintain a high overall rating to attract new reservations.

Another topic of interest is how managers react to internet reviews. Research by Gössling, Hall, and Andersson (2016) revealed that hotel management can improve overall customer impression by responding to unfavourable evaluations in a proactive and pleasant manner. Their research revealed that proficient management responses show a dedication to providing excellent customer service and can transform possibly negative feedback into chances for improvement.

Even with these developments, there are still difficulties in evaluating internet reviews. There are several obstacles, including the large volume of data, linguistic diversity, and the subjective character of reviews. To overcome these issues and improve the validity of their findings, researchers are still creating increasingly complex models. Further developments in machine learning and natural language processing technology could lead to more thorough and precise analysis of user reviews in the future.

2.4.1 Success Stories and Failures

Hotels that have successfully used client feedback to improve their services and increase reservations and ratings are among the success tales found in the literature. Mariani and Borghi (2018), for example, identified hotels that improved customer satisfaction and revenue by taking a proactive approach to responding to evaluations. On the other hand, failures frequently involve hotels that disregarded patron criticism, which damages

Hotel management can learn from the past and create ways to interact with customers more effectively and provide better services by studying both successful and unsuccessful scenarios. Active review management is crucial since it has been demonstrated that proactive interaction with customer reviews increases customer satisfaction and loyalty (Mariani & Borghi, 2018).

2.4.2 Research Based on Nationality

Understanding differences in consumer expectations is essential for international hotel companies serving a broad audience. Hotels can improve their appeal and satisfaction across a range of markets by customising their offerings to match the unique requirements of distinct consumer segments. Further highlighting the necessity for culturally sensitive service techniques is research that indicates language limitations and cultural quirks might impact how reviews are written and interpreted (Bosangit, Dulnuan and Mena, 2012).

Customer expectations and language used in reviews on websites such as Booking.com are influenced by nationality, which has a big impact on how people feel about their hotel experiences. Research indicates that, depending on their cultural origins, international guests frequently give varying priorities to different parts of hotel services, which may have an impact on their overall satisfaction and feedback (Ali et al., 2021; Sheng-Hshiung, Gwo-Hshiung and Kuo-Ching, 1997). Cultural disparities could cause certain countries to prioritise cleanliness and hotel quality, while others could prioritise location and customer service.

Because of these cultural differences, examining customer reviews from the standpoint of a customers' country of origin can reveal important information about how various groups see and rate hotel services. Hotel managers who want to raise customer retention and service quality in a variety of marketplaces must be aware of these subtleties. Additionally, hotels can create tailored strategies that meet the unique needs and expectations of various cultural groups by recognising trends in the sentiment and matter of reviews from a variety of nationalities (Reisinger and Turner, 2024).

Future research in this report will examine the ways in which nationality and other regional characteristics affect customer satisfaction ratings and reviews. Through an analysis of changes in review content, word count, and attitude according to the country of the reviewers, the study seeks to identify cultural variations in consumer expectations and feedback patterns. According to Chen and Starosta (2005), such a study will shed light on the various needs of tourists from across the world and emphasise the significance of developing marketing and service strategies that are culturally appropriate.

Understanding the expectations and experiences of international tourists will be improved by including a nationality-based perspective into customer review analysis. By using this strategy, hotels can better tailor their offerings to the varied tastes of their foreign client base, which will increase overall customer satisfaction.

2.5 Challenges and Difficulties in Analysing Online Reviews

When it comes to assessing internet reviews, researchers encounter a few fundamental obstacles, including subjectivity and bias, linguistic and cultural variations, and data volume and variety. Resolving these issues is essential to extracting precise and useful information from feedback from customers.

The data is varied and includes details like review dates and traveller kinds in addition to numerical ratings and in-depth text reviews. Robust computational algorithms and sophisticated data processing frameworks are necessary to handle such massive volumes of data (He, Zha, & Li, 2013). To maintain high standards of accuracy and efficiency while accommodating the constant flow of new data, researchers must make sure their methods are scalable.

Reviews from customers are written in a variety of languages and cultural backgrounds. Since every language has distinct idioms, linguistic nuances, and cultural references that might influence how one interprets sentiment and the subject matter, this diversity complicates the study (Bosangit, Dulnuan, and Mena, 2012). These variations must be taken into consideration for analysis to be effective and prevent misunderstandings that could provide false conclusions.

There are constant issues in ensuring objectivity in sentiment analysis and resolving any biases in consumer reviews. Reviews are by their very nature subjective, representing unique viewpoints and experiences that may be shaped by the author's prejudices, prior experiences, and the particulars of the stay (Mudambi & Schuff, 2010). The outcomes of analyses can be

distorted by these biases, thus it's critical to create techniques that can recognise and lessen their effects.

2.6 Present Research

The analysis of feedback from customers has undergone substantial development, resulting in the hotel business becoming increasingly data-driven. While the industry first relied on straightforward textual feedback and star ratings, it today employs advanced analytical approaches to obtain deep understanding into how customers feel. This change has been made possible by the emergence of websites like Booking.com, TripAdvisor, and Airbnb, which enable a more sophisticated analysis of user input.

Reviews have an impact on hotel reputation and booking decisions, according to early study. Negative evaluations had the opposite effect on hotel image and bookings as positive reviews did (Vermeulen & Seegers, 2009). Techniques such as sentiment analysis and topic modelling were used to derive precise insights from the growing volume of reviews. Subjectivity, linguistic diversity, and data bulk continue to be obstacles despite progress. As NLP and machine learning continue to progress, researchers are creating increasingly sophisticated approaches to deal with these problems.

Building upon the detailed exploration of customer satisfaction through sentiment analysis and topic modelling, this dissertation aims to address several key research questions that will guide the study. These questions are designed to delve deeper into the nuances of customer reviews and their implications for the hospitality industry. Specifically, this research seeks to answer the following:

1. How do sentiment scores from customer reviews correlate with overall hotel satisfaction ratings?
2. How do the interrelations between sentiment, cultural background, and specific hotel attributes contribute to predicting customer loyalty and repeat bookings across different geographic regions?
3. How can insights derived from analysing customer reviews be leveraged by hotels to enhance customer satisfaction and improve their competitive position in the market?
4. How do the predominant themes and sentiment distributions identified in customer reviews vary across different hotel locations, and what do these variations reveal about

the specific areas of improvement needed for different hotel attributes in the UK hospitality industry?

These research questions will guide the subsequent analysis and form the basis for the methodological approach, ensuring that the study remains focused on its objectives while contributing valuable insights to the field of hospitality management.

3. Methodology

The methodology section describes the research design and methods employed to examine customer reviews and hotel performance across Europe, utilising data sourced from Booking.com. This section provides a thorough explanation of the research methodology, including the methods used for gathering data, the analytical instruments used, and the strategies used to ensure the validity and reliability of the findings. This methodology, which emphasises the methodical gathering and examination of extensive review data, guarantees solid and repeatable results and raises the study's general credibility.

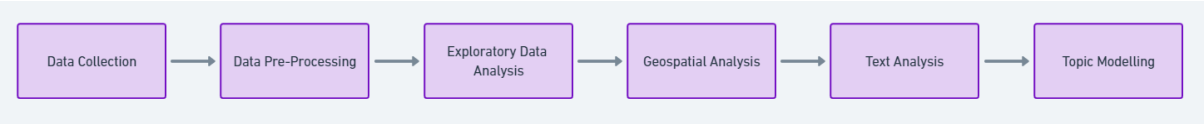

Figure 1: Flowchart of Methodology

3.1 Data Description

Customer reviews for hotels in different countries and areas are included in the dataset. These reviews offer insightful information about customer preferences, opinions, and satisfaction levels. The dataset includes scores of 1,493 hotels in Europe as well as a thorough compilation of 515,738 customer reviews, positive, negative and neutral. Every hotel's physical location is also given, allowing for additional regional research and insights into the distribution of accommodations and related ratings.

The dataset is comprised of 17 features, each providing valuable information about the hotels and reviews. Below is a detailed description of each feature:

Feature	Description
Hotel_Address	Full Address of the hotel. Used for geographical analysis.
Review_Date	Date when the reviewer posted the corresponding review.
Average_Score	Average score of the hotel, calculated based on the latest comments in the past year. Indicates overall performance.
Hotel_Name	Name of the hotel. Identifies and categorises reviews.
Reviewer_Nationality	Nationality of the reviewer. Analyses satisfaction across demographics.
Negative_Review	Highlights areas of dissatisfaction.
Review_Total_Negative_Word_Counts	Total number of words in the negative review.

Positive_Review	Highlights areas of satisfaction.
Review_Total_Positive_Word_Counts	Total number of words in the positive review.
Reviewer_Score	Score the reviewer has given to the hotel based on their experience. Key metric for evaluating satisfaction.
Total_Number_of_Reviews_Reviewer_Has_Given	Number of reviews the reviewer has given in the past. Indicated popularity and engagement.
Total_Number_of_Reviews	Total number of valid reviews the hotel has.
Tags	Context tags the reviewer gave to the hotel. Used for segmenting and analysing reviews.
days_since_review	Time since the review was posted.
lat	Latitude of the hotel.
lng	Longitude of the hotel.

Table 1: Summary of Data Fields and Descriptions for Hotel Reviews Dataset

3.2 Software and Tools

This section outlines the software and tools used for data processing, visualisation, and analysis. R, a well-established software environment for statistical computation, was chosen for its robust capabilities in data manipulation, statistical analysis, and visualization.

3.3 Data Pre-Processing and Transformation

Data pre-processing and transforming data are crucial procedures to guarantee that the dataset is suitably ready for analysis. These processes are crucial for cleaning, standardising, and structuring the data to make it suitable for statistical analysis and modelling (Han, Kamber and Pei, 2011). Due to the dataset's mixed nature—it contains both structured data in the form of numerical and categorical variables, along with unstructured data in the form of text—several R pre-processing and transformation techniques were used.

i. **Handling Missing Data:** The first step in data pre-processing was identifying and addressing missing data, which can significantly impact the accuracy and reliability of analyses (Little and Rubin, 2019). Missing values in the dataset were handled by removing rows or columns with a significant proportion of missing entries. This approach was chosen to maintain data integrity and avoid potential biases that could arise from imputing missing values, especially when the number of missing values was small relative to the dataset size. Instead of imputation, missing data were removed to ensure that only complete and reliable data were used in the analysis. This decision was based on the relatively low proportion of

missing values, which allowed for data removal without significantly impacting the overall dataset size.

ii. **Handling Duplicate Values:** In addition to handling missing data, it was essential to identify and remove duplicate values from the dataset. Duplicate entries can distort analysis results by skewing distributions and introducing redundancy. To ensure the accuracy of the data, any rows that were exact duplicates of another were removed. This step was crucial for maintaining the integrity of the dataset and ensuring that each entry was unique, thereby improving the reliability of subsequent analyses.

iii. **Combining Review Texts:** To streamline text data processing, all reviews, regardless of sentiment, were consolidated into a single column named "Review." This consolidation was essential for simplifying subsequent text processing steps and ensuring that all review data was centrally located, enhancing the efficiency of text analysis.

iv. **Text Cleaning:** Several cleaning procedures were used to get the text data ready for analysis:

- **Converting to Lower Case:** All text was standardised to lower case to reduce redundancy and ensure consistency.
- **Removing Punctuation and Numbers:** Punctuation and numerical characters were removed to focus on meaningful words, thereby reducing noise in the dataset and improving the quality of text analysis.
- **Stopword Removal:** Stopwords are common words (e.g., "and" "is", "in") that typically do not add significant meaning to text analysis and are often removed to focus on more informative words. A customised list from the stopwords package was used to remove common stopwords, except for words such as "not," "no," "nor", "neither," and "cannot," which are important for sentiment analysis. This step enhanced the analysis by concentrating on the most meaningful words, reducing noise, and improving the quality of insights derived from the data.
- **Lemmatisation:** Lemmatisation was performed to reduce words to their base or root forms, which is crucial for normalising text data. For example, words like "running," "ran," and "runs" are reduced to "run." This process helps group similar words, improving the quality and coherence of text analysis by reducing the dimensionality of the data.

v. **Reconstructing Cleaned Reviews:** After cleaning, the individual words were recombined into complete reviews using grouping and summarising functions. This step ensured that

the cleaned text was in a format suitable for further analysis while retaining the benefits of the previous cleaning steps.

vi. **Integrating Cleaned Text:** The cleaned and lemmatized reviews were reintegrated into the original dataset as a new column. This integration allowed for comparative analysis and modelling by ensuring that the cleaned text data was available alongside the original data.

vii. **Tokenization and Tidy Formatting:** The combined review text was formatted into a tidy format using tokenization, which involves splitting the text into individual words. This format, facilitates detailed word-level text analysis, allowing each word to be represented as a single row in the dataset. This method supports various text mining tasks, such as frequency analysis and sentiment analysis.

3.4 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an essential phase in the data analysis process that involves using visual approaches to summarise a dataset's key features. EDA's main objectives are to analyse data distributions, identify trends, find anomalies, and determine how variables relate to one another (Tukey, 2019). This stage allows for a thorough understanding of the data before applying complex statistical analyses or machine learning models. These techniques can highlight significant trends and insights that could otherwise go unnoticed and help to guarantee that the data is appropriate for the intended studies (Behrens, 2024).

3.4.1 Word Frequency Calculation

Word frequencies were determined by counting the occurrences of unique words in the tokenised dataset to determine the most mentioned words in reviews from customers. This stage yielded useful insights into the main subjects and themes that consumers frequently address in their reviews by quantifying word frequency.

3.4.2 Word Cloud Visualisation

A word cloud was generated to graphically depict the terms that customers used most frequently. A word cloud is a widely used data visualisation method that shows words in different sizes according on how frequently they appear in a text dataset. A term's frequency of occurrence in the data increases with its size in the word cloud. Within a significant amount of textual data, this technique is especially helpful for locating major themes and important subjects (Heimerl et al., 2014).

3.4.3 N-gram Extraction

N-gram analysis is a text mining technique that looks at word sequences (called n-grams) that are contiguous in a dataset. Using n-gram analysis, common phrases and expressions in customer evaluations were found in this study. This allowed for greater understanding of the sentiments and topics that consumers commonly mentioned. By capturing multi-word phrases that represent additional aspects of customer feedback, this approach goes beyond single-word analysis (Manning, Raghavan, and Schütze, 2008).

i. **Bigram Extraction:** Bigrams, or two-word sequences, were extracted from the cleaned review text to identify commonly recurring phrases that describe key elements of the hotel experience. Specific bigrams that did not contribute meaningful insights, such as "no negative" and "no positive," were filtered out to focus on more informative phrases.
ii. **Trigram Extraction:** Trigrams, or three-word sequences, were also analysed to uncover more specific expressions in customer reviews. Irrelevant trigrams were filtered out to enhance the focus on meaningful phrases.

More detailed insights into customer sentiment and preferences can be gained by identifying common phrases that customers commonly use to describe their experiences. The analysis improves the comprehension of customer reviews and identifies specific areas of customer satisfaction and dissatisfaction by looking at bigrams and trigrams, which helps to capture the relational context between words that single-word analysis could miss (Jurafsky and H. Martin, 2018).

3.4.4 Analysing Descriptive Language through Part-of-Speech Tagging

A natural language processing (NLP) technique called Part-of-Speech (POS) tagging is used to tag each word in a text with a part of speech, such as a noun, verb, adjective, etc. This analysis is valuable for understanding the grammatical structure of sentences and identifying specific word types that contribute to the sentiment and descriptive qualities in textual data (Jurafsky and H. Martin, 2018).

POS tagging was carried out on the cleaned review texts by applying the loaded model to the dataset. With the help of this, every word in the text is given a part of speech, giving extensive grammatical information that can be used for additional analysis. The satisfaction and dissatisfaction of customers is frequently directly correlated with adjectives like "good," "great," "small," and "poor".

Adjectives are important markers of attitude and perception in customer reviews, and the POS tagging analysis makes it possible to examine them in-depth. Through adjective isolation, the study revealed hotel qualities that customers often bring up, both favourably and unfavourably. This strategy helps determine areas for improvement and comprehend the satisfaction of customers (Bird, Ewan Klein and Loper, 2009).

3.4.5 Location-Based Analysis

Understanding how geographic and cultural factors affect customer ratings and satisfaction requires location-based analysis. Using R programming, this study carried out a thorough geographic analysis. It processed and visualised spatial data using packages including ggplot2, sf, dplyr, and rnaturalearth (Pebesma, 2018; Wickham et al., 2019).

i. **Aggregation of Hotels by Country:** The dataset was grouped by country to count the number of unique hotels in each location, providing a clear view of hotel distribution across regions. According to the data, France has the most hotels (458), followed by the United Kingdom (400) and Spain (211). Assessing regional trends in customer satisfaction and review patterns requires an understanding of hotel distribution (Chhetri et al., 2020). A uniform mapping of country names was utilised to guarantee consistency across the entire set.

ii. **Distribution of Reviews by Reviewer Nationality:** The dataset was aggregated by nationality of reviewers to determine the regions with the highest volume of reviews. To visualise this distribution, a map was made, showing that, with roughly 246,209 reviews, the United Kingdom had the most, followed by the United States (35,634 reviews) and Australia (21,688 reviews). The understanding of regional participation and cultural influences on consumer feedback is aided by this visualisation.

iii. **Word Count Analysis by Reviewer Nationality:** The verbosity of reviews from different countries was shown by calculating the word count for both positive and negative reviews. To investigate cultural variations in feedback, the average word count for both positive and negative reviews was calculated. Bar charts were used to compare the feedback styles of the top 10 countries based on the average word count for both forms of analyses.

iv. **Average Hotel Review Scores by Country:** The dataset was divided into countries, and the average review score for each “ was determined to investigate the effect of geographic location on review scores. These scores were visualised as a line graph using ggplot2, which

allowed for cross-national comparisons of customer satisfaction levels and highlighted differences in average review scores (Wickham, 2016).

v. **Geospatial Clustering of Countries Based on Average Review Word Count:** Prior to clustering, the dataset was meticulously prepared to ensure accuracy. Different representations of the same location, such as "United States of America" and "United States," were standardised to a single form to facilitate accurate aggregation and analysis. Rows with missing latitude or longitude were removed to prevent errors in spatial processing. Additionally, territories and dependencies (e.g., "Gibraltar" to "United Kingdom") were mapped to their parent countries to avoid misclassification, ensuring the integrity of the analysis. The cleaned dataset was aggregated by nationality to compute average word counts for reviews.

- Determining the Optimal Number of Clusters:

  To determine the optimal number of clusters for grouping countries based on average review word count, the Elbow Method was applied. This method involves plotting the total within-cluster sum of squares (WCSS) against the number of clusters (Bholowalia and Kumar, 2014). The WCSS is calculated as:

$$WCSS = \sum_{i=1}^{n} (x_i - \mu)^2$$

  where:

  - n is the number of data points in the cluster,
  - x_i represents each data point,
  - μ is the centroid of the cluster.

  For multiple clusters, the total WCSS is the sum of the WCSS values for all clusters. By plotting the WCSS for different values of **K** (number of clusters), the point where the rate of decrease sharply diminishes (the "elbow point") is identified. This point represents the optimal number of clusters because adding more clusters beyond this does not provide additional meaningful value or significantly reduce the within-cluster sum of squares (WCSS), indicating that further clustering would only capture minor variations and not enhance the overall understanding of the data.

- K-means Clustering Algorithm:

  After determining the optimal number of clusters, the K-means clustering algorithm was employed to group countries based on the computed average word counts. K-means

clustering is an unsupervised learning algorithm that partitions data into **K** clusters by minimising the sum of squared distances between data points and their corresponding cluster centroids.

The objective function minimised by K-means clustering is given by:

$$J = \sum_{j=1}^{K} \sum_{x_i \in C_j} ||x_i - \mu_j||^2$$

where:

- J is the total within-cluster sum of squares,
- K is the number of clusters,
- x_i represents a data point (average word count for a country in this context),
- μ_j is the centroid of cluster j,
- C_j is the set of data points in cluster j.

Implementation of K-means Clustering:

a) **Initialization**: The centroids of the clusters were initialised randomly.
b) **Assignment Step**: Each country (data point) was assigned to the nearest cluster centroid based on the average word count.
c) **Update Step**: The centroids of each cluster were recalculated by taking the mean of all data points assigned to the cluster.
d) **Convergence**: Steps 'b' and 'c' were repeated until the centroids no longer changed significantly, indicating convergence.

Using the K-means clustering algorithm, countries were grouped into three clusters based on their average review word counts. This clustering revealed geographic groupings and trends in review styles, identifying clusters with high, moderate, and low average word counts. This analysis highlights how geographic and cultural contexts influence customer feedback (Jain, 2010).

vi. **Analysis of Trip Types:** A tag dictionary was developed to classify the different trip types that reviewers mentioned to assess customer travel preferences. To highlight the frequency of each trip type, this required breaking down tags into their component parts, eliminating unnecessary characters, cleaning the "Tags" column to extract relevant data, and organising

the tags into a dictionary. The dictionary revealed information about consumer behaviour and preferences. “Leisure trip” and “Couple” were the most often stated tags, demonstrating a significant inclination towards leisure and romantic experiences.

3.5 Text Analysis

3.5.1 Sentiment Analysis

Sentiment analysis of customer reviews was performed using the ‘sentimentr’ package and NRC Lexicon to categorise sentimental tones expressed in the feedback. To determine the relationship between text sentiment and review scores, this procedure included tokenisation, sentiment categorisation, visualisation, and correlation analysis.

Tokenisation and Sentiment Categorisation Using the NRC Lexicon

The reviews were tokenised into individual words to facilitate detailed sentiment analysis. This process involved breaking down the text into discrete words, which were then matched with the NRC Lexicon. The NRC Lexicon is a widely used tool for sentiment analysis that categorises words into various sentiment categories such as joy, trust, fear, anger, and sadness (Mohammad and Turney, 2013). The analysis was able to provide an in-depth understanding of the text's sentimental content by mapping terms from the analyses to these sentiments.

For this analysis, positive and negative reviews were examined separately. To enable a more focused analysis of sentiments found in the evaluations, general sentiment categories such as “positive” and “negative” were eliminated (Kiritchenko, Zhu and Mohammad, 2014). With the use of this method, it was possible to figure out the various sentimental tones that reviewers were expressing, which led to a better comprehension of customer attitudes and their possible effects on feedback and overall satisfaction.

Sentiment Score Analysis Using the ‘sentimentr’ Package

The sentiment scores of reviews were calculated using the ‘sentimentr’ package, which provides a robust method for quantifying the sentimental tone of text data. This package allows for a deeper understanding of sentiment than just word counts by computing sentiment scores based on the polarity of words and phrases inside sentences.

Density plots were used to visualise the analysis of these sentiment scores, showing how the sentiment scores varied amongst positive and negative reviews. The positive and negative

sentiment plots showed clearly defined peaks, indicating an obvious difference in the sentimental tone of the two types of reviews. This distinction offers important insights into the dynamics of consumer feedback by capturing the overall sentimental impact that the reviews are likely to have on readers (Taboada et al., 2011).

Pearson Correlation Analysis

The Pearson correlation coefficient was employed to evaluate the relationship between sentiment scores and review scores. The Pearson correlation coefficient is a statistical measure used to determine the strength and direction of the linear relationship between two continuous variables. This method was selected due to its ability to quantify the degree of association between sentiment scores, derived from textual reviews, and numerical review scores, providing insights into how changes in sentiment scores relate to changes in review scores (Nicewander, 1988).

The Pearson correlation coefficient, denoted as r, is calculated using the following formula:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

where:

- X_i *and* Y_i represent individual data points for sentiment scores and review scores, respectively,
- $\bar{X}$ is the mean of the sentiment scores,
- $\bar{Y}$ is the mean of the review scores,
- $\sum$ indicates the summation over all data points.

To assess the statistical significance of the correlations, p-values were calculated for each Pearson correlation coefficient. A very small p-value ($p < 0.001$) indicates that the observed correlation is highly unlikely to have occurred by chance, confirming a statistically significant relationship between sentiment scores and review scores.

In this study, sentiment scores were compared with numerical review scores to identify any correlations between these variables. The results revealed strong positive correlations, with coefficients of 0.70 for positive sentiment scores and 0.80 for negative sentiment scores, suggesting that higher sentiment scores are associated with higher review scores.

These correlations were further illustrated using scatter plots, which visually demonstrated the positive linear relationship between sentiment scores and review scores. This analysis provides valuable insights into customer satisfaction by revealing the connection between customer sentiments expressed in textual reviews and their corresponding numerical ratings.

3.5.2 Topic Modelling

Topic modelling was applied to customer reviews from the United Kingdom using the Latent Dirichlet Allocation (LDA) algorithm. LDA is a popular method for identifying hidden topics within large text datasets by grouping words that frequently appear together, thus revealing underlying themes (Blei et al., 2003).

Data Preparation

The entire dataset consisted of 515,738 entries, encompassing reviews from various nationalities. For this study, topic modelling was focused specifically on the subset of reviews from UK citizens. This targeted approach enabled a detailed examination of feedback from UK reviewers, allowing for insights that are particularly relevant to the UK market. To get the text data ready for analysis, a thorough preparation step was conducted before topic modelling. Through the assurance of consistent input data and the reduction of noise, the preprocessing techniques were essential to improving the analysis results.

A Document-Term Matrix (DTM) was constructed following preprocessing to show the frequency of terms in the reviews. The term frequency-inverse document frequency (TF-IDF) weighting technique was applied to the DTM to improve the model's performance even more. TF-IDF helps identify terms that are more informative by emphasising less frequent, but more significant words, and de-emphasising common words that occur frequently across documents.

The TF-IDF for a term *t* in a document *d* is calculated as:

$$TF - IDF(t, d) = TF(t, d) \times IDF(t)$$

where:

- *TF(t, d)* (Term Frequency) is the number of times term *t* appears in document *d*.
- *IDF(t)* (Inverse Document Frequency) is calculated as:

$$IDF(t) = \log\left(\frac{N}{DF(t)}\right)$$

with:

- *N* being the total number of documents in the corpus,
- *DF(t)* (Document Frequency) being the number of documents containing the term *t*.

Determination of the Optimal Number of Topics

To determine the optimal number of topics, coherence scores were calculated. Coherence scores are a measure of the interpretability and meaningfulness of the topics generated by the model, with higher scores indicating more coherent and distinguishable topics (Röder et al., 2015). Coherence is calculated based on the degree of semantic similarity between high-scoring words in the topic and is defined by the following formula:

$$Coherence = \sum_{w_i, w_j \in T} \log \frac{D(w_i, w_j) + \epsilon}{D(w_j)}$$

where:

- T is the set of words in the topic,
- w_i and w_j are words within the topic,
- $D(w_i, w_j)$ is the number of documents containing both words w_i and w_j,
- $D(w_j)$ is the number of documents containing word w_j,
- ϵ is a small constant to prevent division by zero.

Based on these scores, a model with seven topics was chosen, as it provided the most interpretable and distinct set of topics from the UK customer reviews.

After determining the ideal number of topics, each topic was given a name that reflected unique themes of the hotel experience as experienced by UK citizens and was derived from the terms that appeared most frequently in each category. To guarantee that the themes were both culturally and statistically meaningful, a manual naming procedure was essential.

Model Selection as Latent Dirichlet Allocation (LDA)

LDA was chosen as the topic modelling algorithm due to its effectiveness in identifying latent topics within large collections of documents. With the ideal number of topics selected, the LDA model was executed. The LDA model's parameters were estimated using Gibbs

sampling, a Markov Chain Monte Carlo technique. The LDA model yields two important matrices:

- **Document-Topic Matrix**: Represents the distribution of topics for each document (review), indicating which topics are most prevalent in each review.
- **Topic-Term Matrix**: Lists the top terms associated with each topic, providing insight into the main themes captured by the model.

Key Topics Identified in UK Reviews

The top terms and their coherence scores for each of the seven subjects found by the LDA model were found. These ratings help in demonstrating each topic's uniqueness and clarity. Out of all the topics, Topic 1 has the highest coherence score and is the easiest to understand and distinguish. The overall coherence ratings, however, are somewhat low, indicating that there may be some topic overlap and that the themes might not be totally different.

Distribution of Themes Across Reviews

The distribution of review themes across customer reviews written by UK nationals was analysed to understand the different aspects of hotel experiences highlighted by reviewers. This analysis reveals that UK nationals emphasize different aspects of their hotel experiences depending on the country, reflecting how local standards, cultural differences, or personal preferences in hospitality might influence their perceptions.

Sentiment Distribution Across Topics

The relationship between sentiment and the identified topics was further investigated after completing sentiment analysis and topic modelling to obtain a deeper understanding of customer reviews. To guarantee a targeted study on a specific population, the dataset was restricted to reviews from UK citizens alone. The reviews were categorised according to their relevance to the topic by using the LDA model's highest posterior probability. To understand the variation in sentiment across these topics, the proportions of positive and negative reviews for each topic were calculated. The distribution of sentiment throughout the specified themes was clearly shown by the bar chart that used these proportions to visualise it. Each bar represented a topic and was further broken into segments for positive and negative sentiments. This study made it feasible to correlate specific topics with the sentiments conveyed, which improved our comprehension of the responses from customers.

4. Results

The results section is entirely devoted to examining the wide range of customer reviews that Booking.com has gathered. This section uses a combination of text mining techniques, and sentiment analysis to identify the main themes and sentiments that customers have expressed. It also investigates how these factors affect customer satisfaction in different hotel categories and geographical areas.

4.1 Word Frequency Analysis

The Word Frequency Analysis was conducted to identify the most frequently mentioned words in customer reviews on Booking.com. This analysis highlights the key topics that guests frequently discuss.

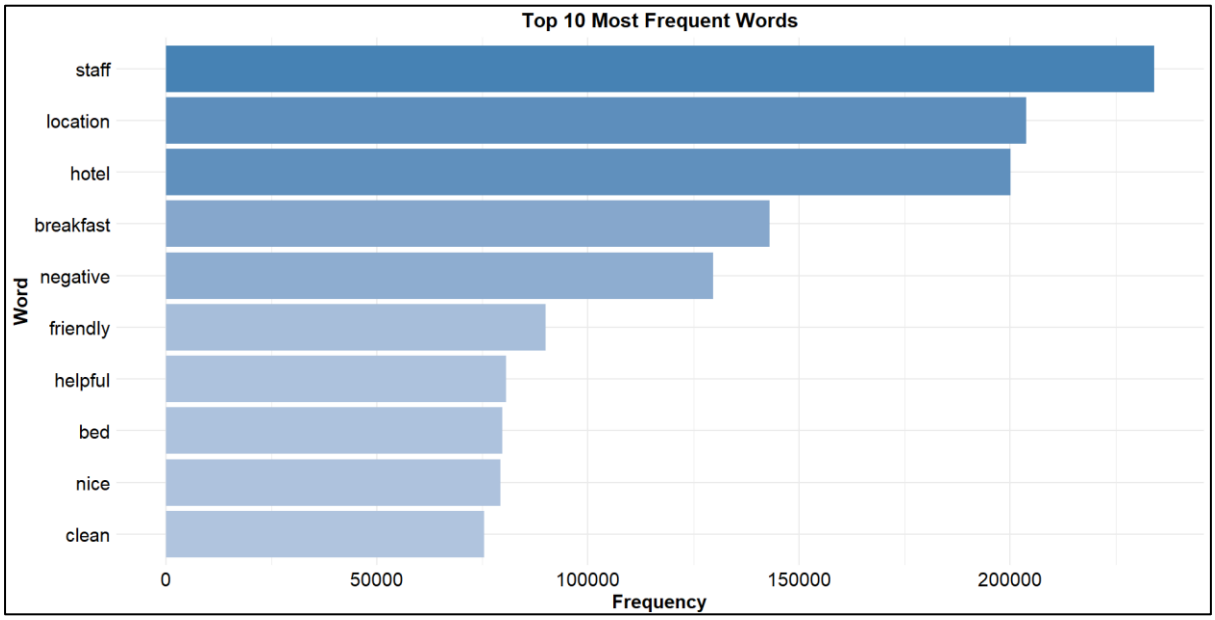

Figure 2: Top 10 Most Frequent Words in Customer Reviews

These frequently mentioned words reveal the main aspects of hotel experiences that guests discuss in their reviews. The prominence of words like “staff” and “location” indicates that customer service and the hotel's geographical position are vital to guest satisfaction. Words such as “friendly,” “helpful,” and “clean” suggest that positive experiences with staff and cleanliness are highly valued by guests. Conversely, the mention of “negative” indicates areas where improvements might be necessary. This analysis helps identify the key elements that influence guest perceptions and satisfaction.

4.2 Word Cloud

The Word Cloud visualisation was created to give the most mentioned words in the customer evaluations a clear, visual depiction. Larger words indicate more frequency, and this kind of visualisation emphasises the relative importance of various terms by showing them in variable sizes. This word cloud illustrates that the terms “hotel”, “breakfast”, “staff” and “location” are most frequently used by customers.

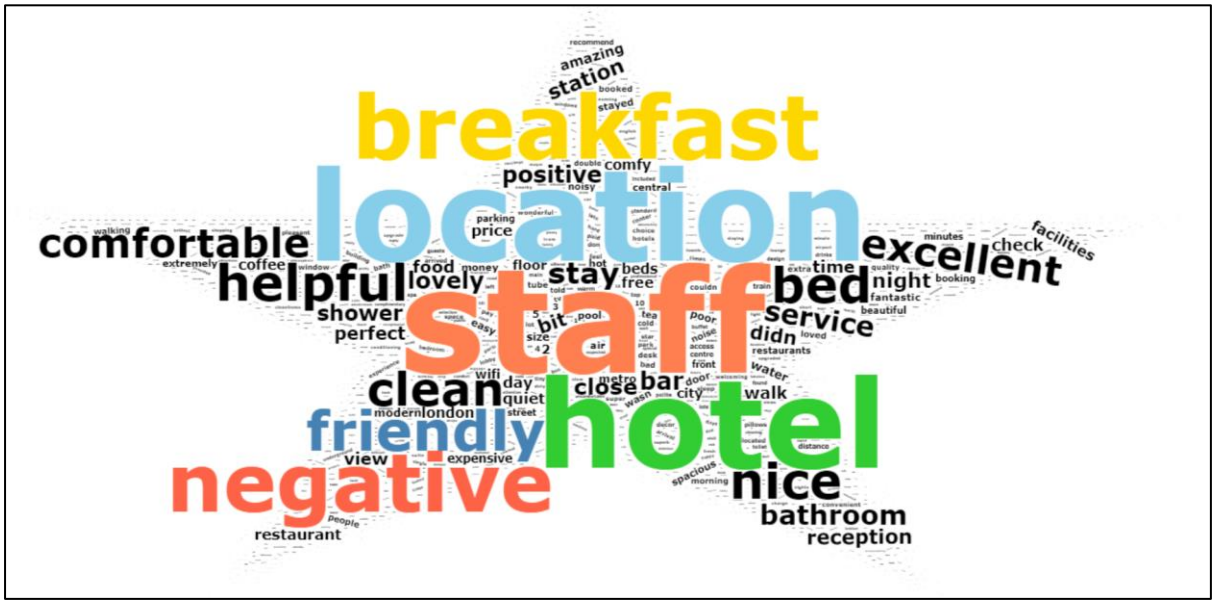

Figure 3: Word Cloud of Most Frequent Words in Customer Reviews

4.3 N-gram Analysis

N-gram analysis was applied to examine common word sequences that regularly occur together in customer reviews, notably bigrams (two-word sequences) and trigrams (three-word sequences). Deeper understanding of the particular words and sentiments that consumers frequently express is offered by this study, which may not be possible with a single-word analysis.

4.3.1 Bigrams

Bigrams, or two-word combinations, were analysed to identify frequently recurring phrases that reflect specific aspects of the hotel experience.

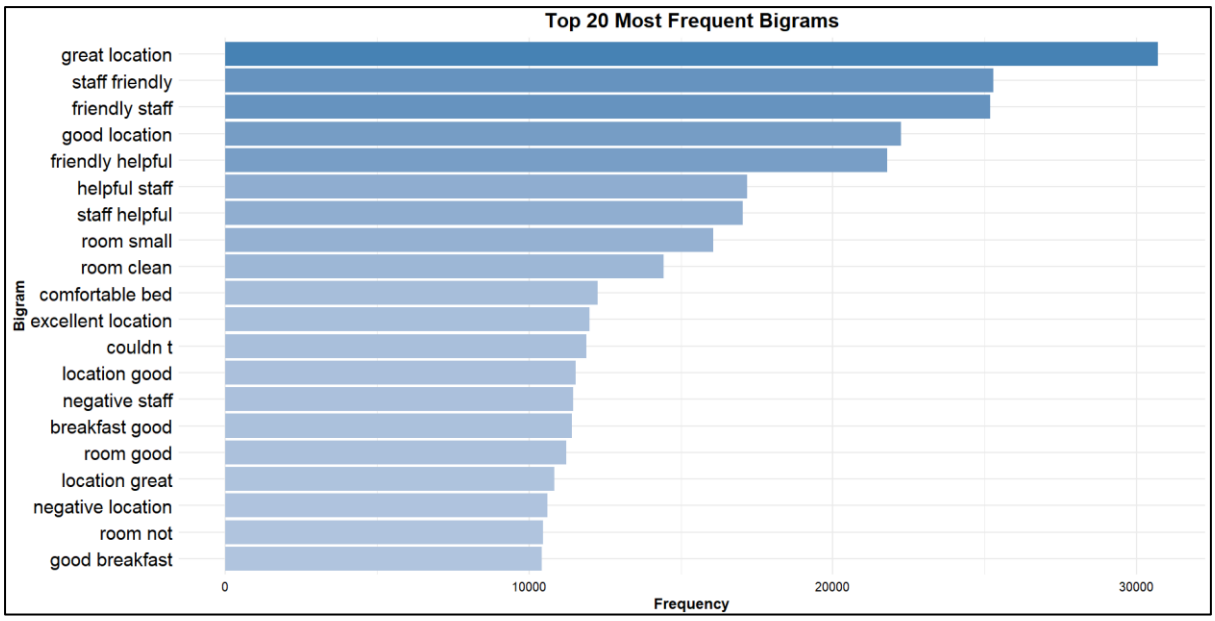

Figure 4: Top 20 Most Frequent Bigrams in Customer Reviews

The top ten bigrams that were most frequently found in the reviews are displayed in the graph in Figure 4. Some of the most significant components of customer experience that these terms capture include “friendly staff”, “comfortable bed”, and “great location”.

4.3.2 Trigrams

Trigrams, or three-word combinations, were also analysed to uncover more specific expressions and sentiments in the reviews.

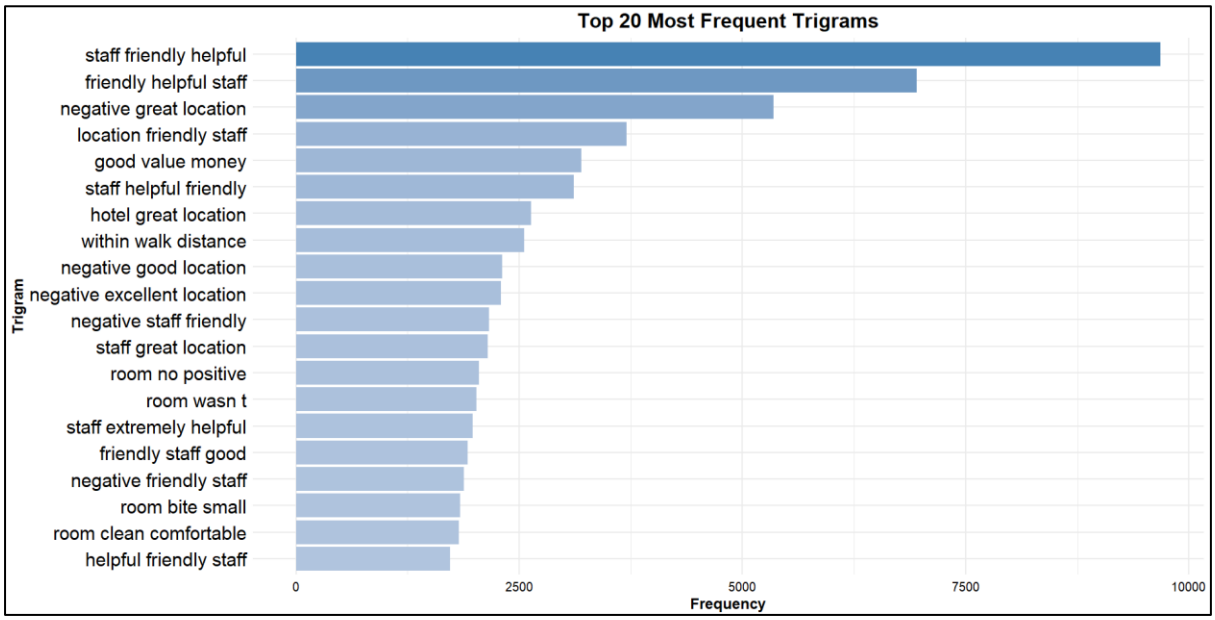

Figure 5: Top 20 Most Frequent Trigrams in Customer Reviews

The top 20 trigrams that showed up most frequently in the data set are shown in Figure 5. Customer feedback is typically more understood in context when it is expressed in these ways: “would stay again”, “great customer service”, and “friendly helpful staff”.

4.4 POS Tagging Analysis

Part-of-Speech (POS) tagging was performed to identify and analyse the adjectives used in customer reviews. Adjectives are particularly useful in understanding the sentiment and descriptive qualities that customers attribute to their experiences. By focusing on adjectives, the analysis aimed to uncover the specific attributes of hotels that customers frequently highlight, both positively and negatively.

4.4.1 Overall Adjective Frequency

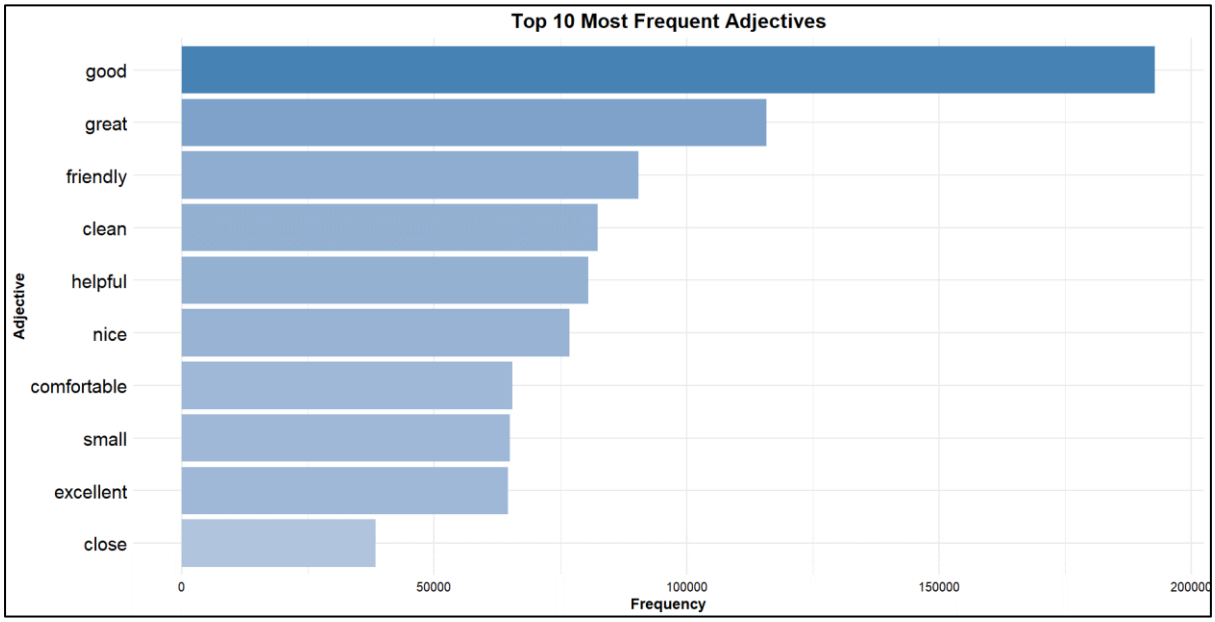

Figure 6: Top 10 Most Frequent Adjectives in Customer Reviews

Figure 6 shows the top 10 adjectives most frequently used across all customer reviews. Words like “good” and “great” have high frequency which means more positive reviews than the negative ones.

4.4.2 Adjectives in Positive Reviews

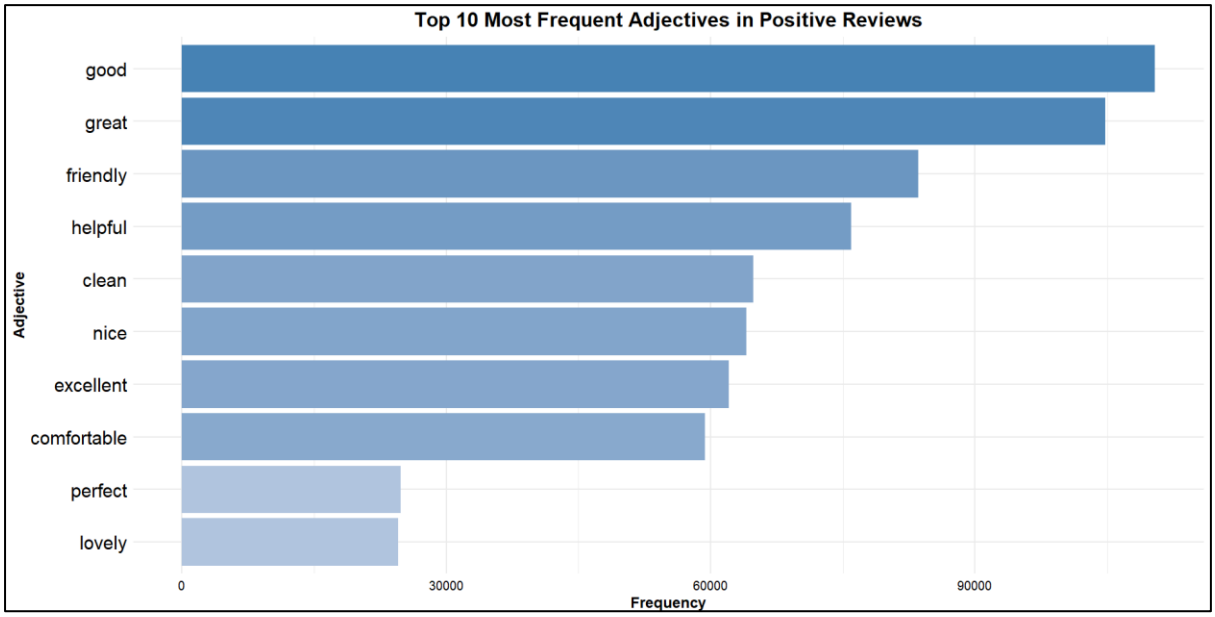

Figure 7: Top 10 Most Frequent Adjectives in Positive Reviews

Figure 6 focuses specifically on the adjectives found in positive reviews. The adjectives "good", "great", "friendly", and "helpful" are particularly prominent in positive feedback, indicating the attributes that customers most appreciate.

4.4.3 Adjectives in Negative Reviews

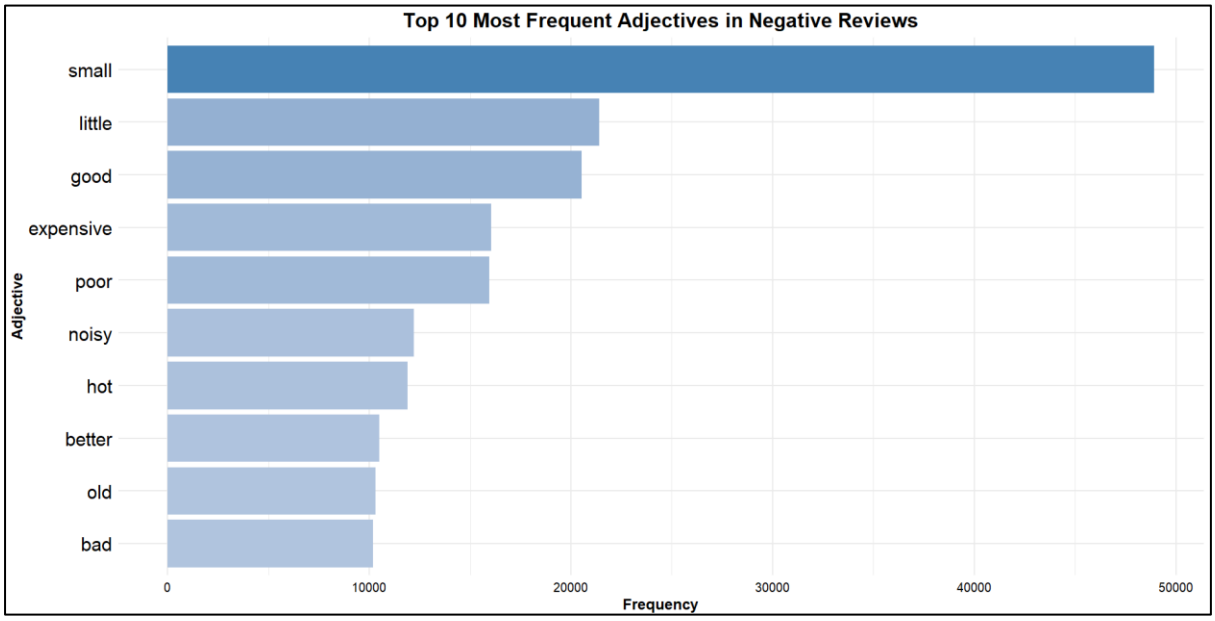

Figure 8: Top 10 Most Frequent Adjectives in Negative Reviews

Figure 8 highlights the adjectives most found in negative reviews. Adjectives such as “small”, “little”, “expensive”, and “poor” are frequently used in negative feedback, pointing to the aspects of the hotel experience that customers are dissatisfied with.

4.5 Location Analysis

The impact of geographic factors on customer reviews is examined in this part. These aspects include the nationality distribution of reviews, the relationship between hotel location and review scores, and the presence of geographic patterns in review content. The investigation sheds light on how cultural and geographic factors affect consumer satisfaction and opinions.

4.5.1 Distribution of Hotels by Country

Understanding the geographic distribution of hotels across different countries is crucial for assessing the impact of regional variables on customer feedback. The number of hotels in each of the countries included in the dataset is summarised in the table below. In the dataset, hotels are most prevalent in France.

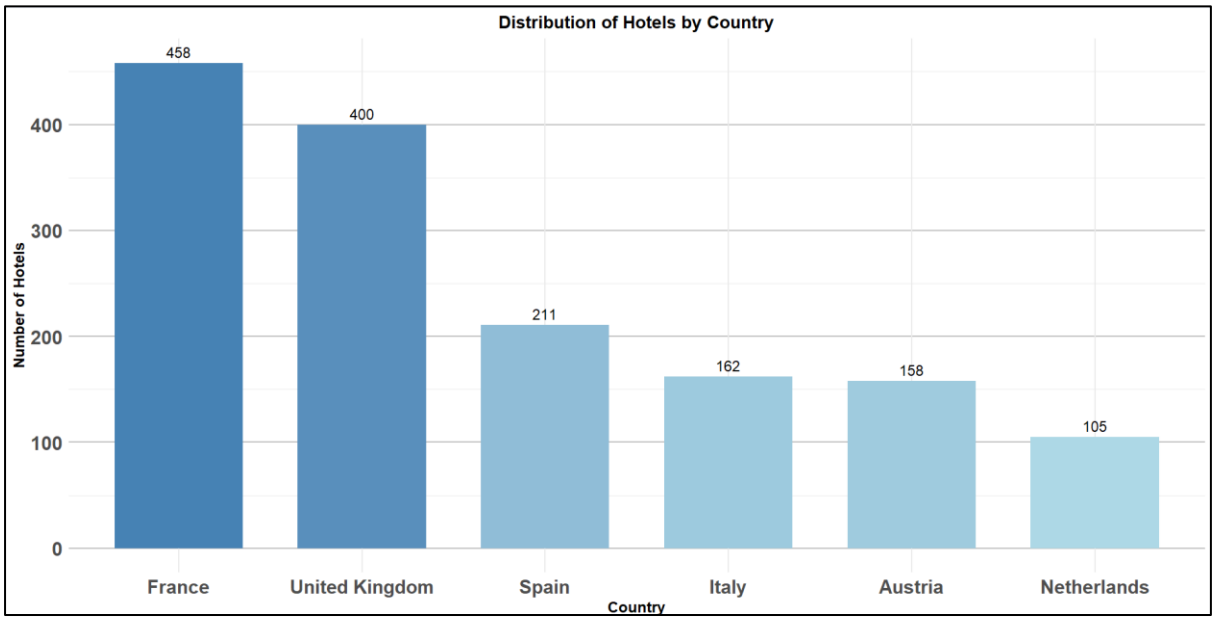

Figure 9: Distribution of Hotels by Country

4.5.2 Distribution of Reviews by Reviewer Nationality

Gaining insight into the cultural and geographical factors influencing hotel reviews can be achieved by analysing the geographic distribution of customer reviews by country. The map in

this section shows how reviews are distributed according to the nationality of the reviewers, emphasising the top 5 countries with the most reviews.

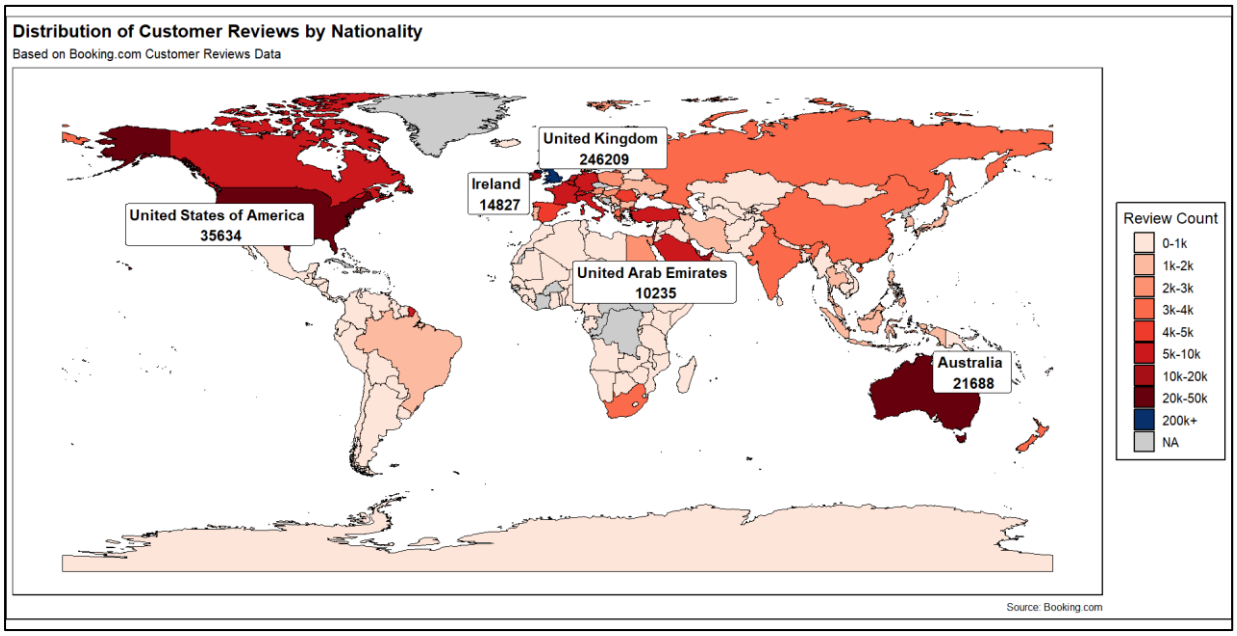

Figure 10: Geographic Distribution of Reviews by Reviewer Nationality (Top 5 Countries)

United Kingdom has the highest number of customer reviews of ~240000 which signifies that majority of the users giving reviews in the dataset are from the United Kingdom.

4.5.3 Word Count Analysis by Reviewer Nationality

Analysing the word count of reviews based on the nationality of the reviewers provides insights into how different cultural groups express their experiences. This subsection presents the average word counts for both positive and negative reviews across various nationalities.

i. Word Count in Positive Reviews by Reviewer Nationality

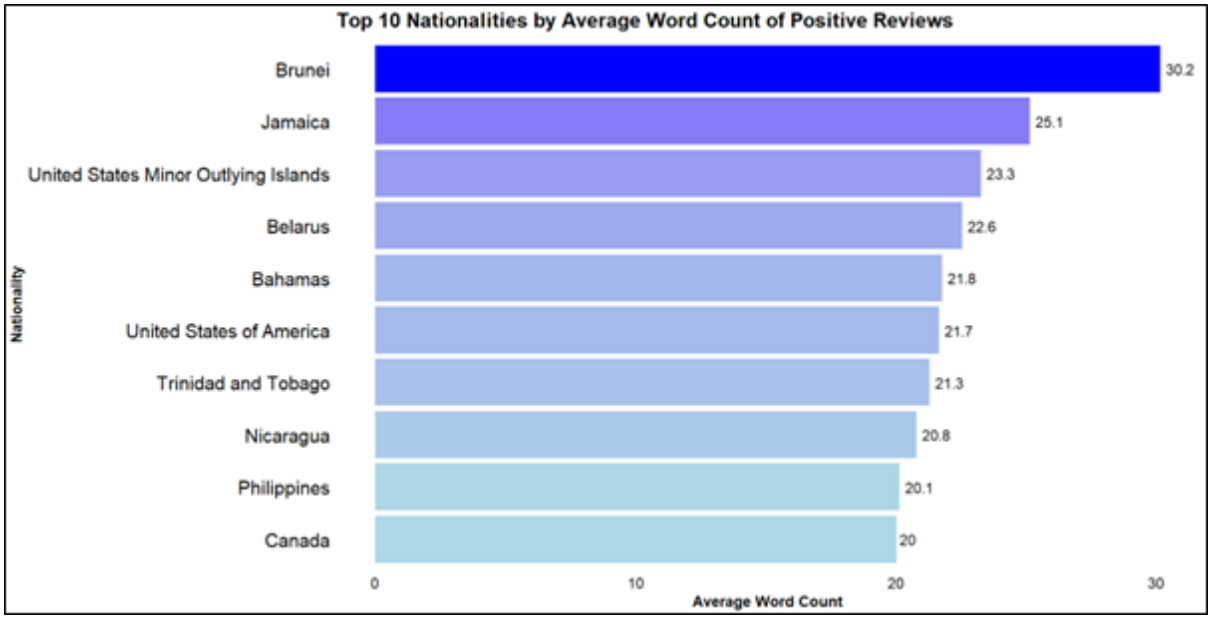

Figure 11: Top 10 Nationalities by Average Word Count of Positive Reviews

The top 10 nationalities by average word count for positive reviews with at least 10 reviews are shown in the graph in Figure 11. With an average of 30.2 words per review, Brunei write the most reviews (30.2 words) followed by Jamaica (25.1 words). These results suggest that when reviewers from these countries have a pleasant experience, they are more likely to provide thorough feedback.

ii. Word Count in Negative Reviews by Reviewer Nationality

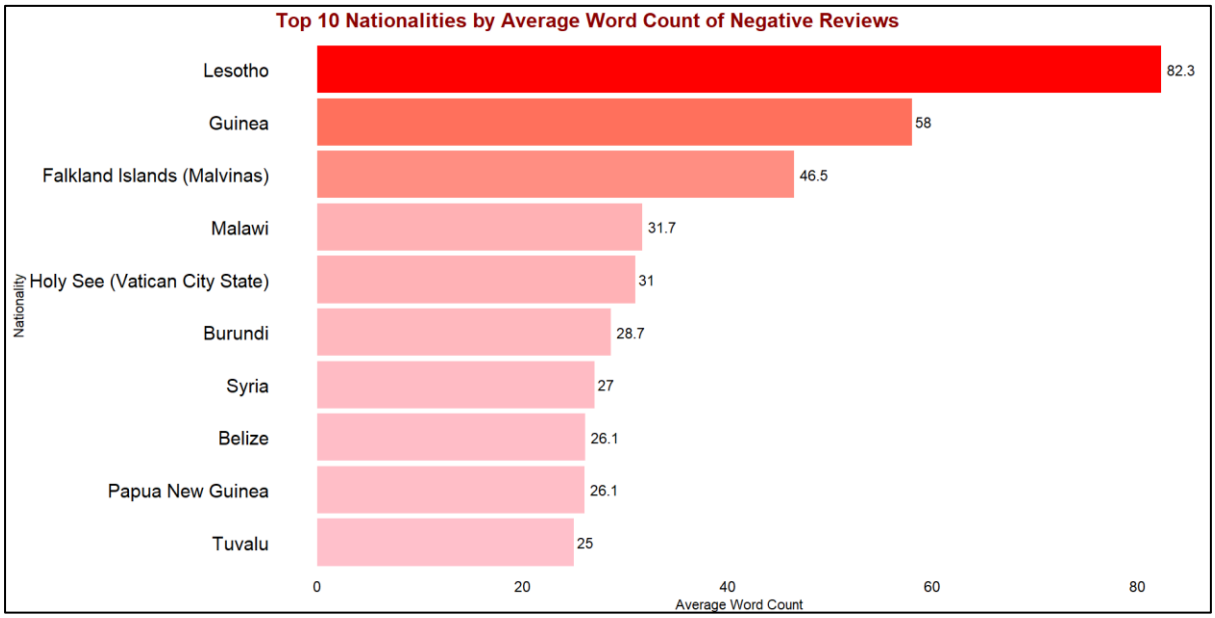

Figure 12: Top 10 Nationalities by Average Word Count of Negative Reviews

The graph in Figure 12 shows the top 10 nationalities by average word count for negative reviews with minimum 10 reviews. Reviewers from Lesotho have the highest average word count for negative reviews at 82.3 words per review, significantly higher than other nationalities. This is followed by Guinea (58 words) and the Falkland Islands (Malvinas) (46.5 words), indicating that reviewers from these countries tend to provide more detailed feedback when their experiences are negative.

4.5.4 Average Hotel Review Scores by Country

The accessibility of a hotel to attractions, convenience, and the surrounding area can all have a substantial impact on how satisfied customers are with their overall stay. This section examines the relationship between hotel locations—which are determined by their latitude and longitude—and average review scores.

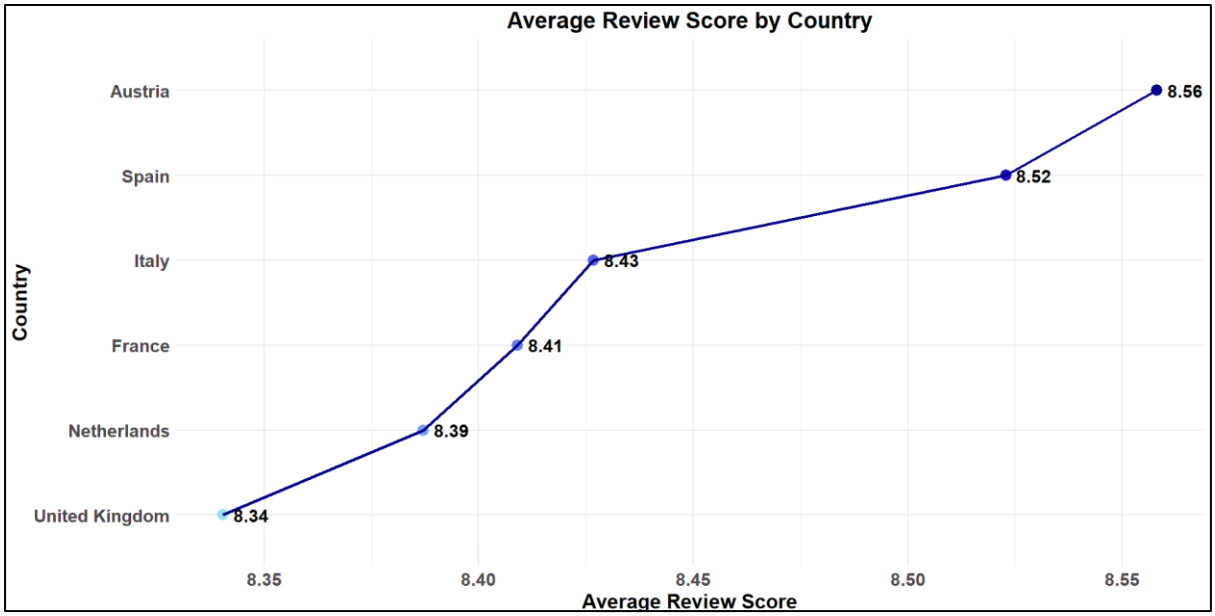

Figure 13: Average Hotel Review Scores by Country

The average review scores of hotels in different countries are shown in Figure 13. With an average review score of 8.56, Austrian hotels lead the field, followed by those in Spain (8.52), Italy (8.43), France (8.41), the Netherlands (8.39), and the United Kingdom (8.34). These variations in average scores demonstrate how a customer's satisfaction and perception may be influenced by their geographic location.

4.5.5 Geospatial Clustering of Countries Based on Average Review Word Count

Geospatial clustering is a technique used to group countries based on similarities in review characteristics, such as average review word count. This analysis reveals patterns and trends in customer feedback influenced by geographic and cultural factors.

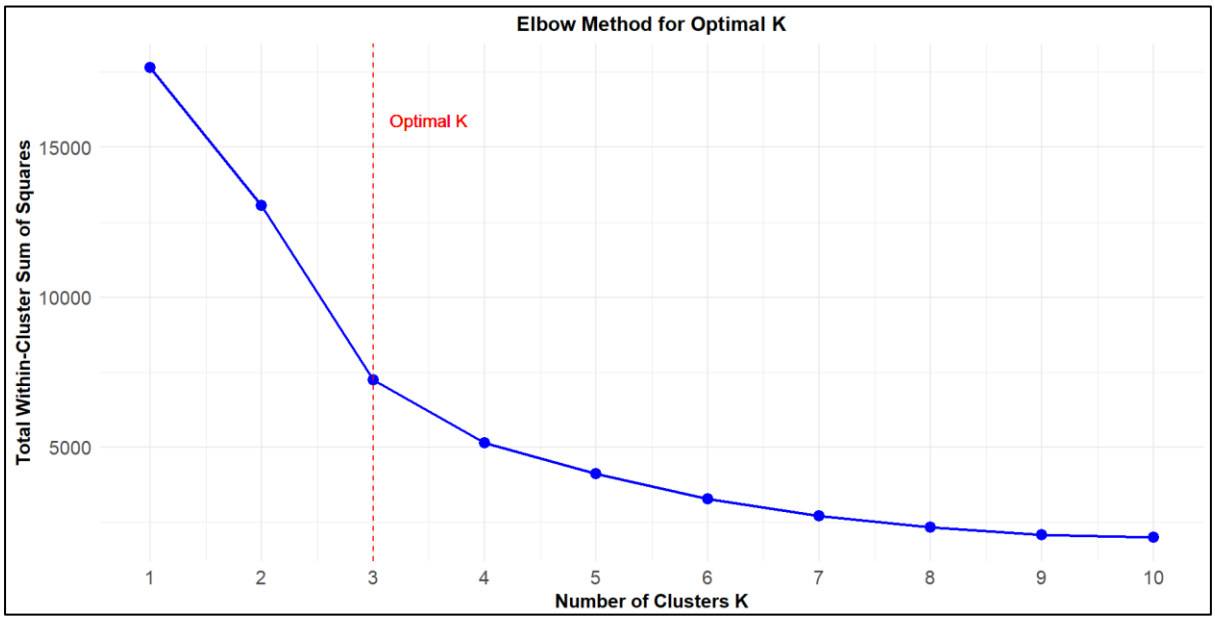

Figure 14: Determining Optimal Clusters with the Elbow Method

The Elbow Method graph in Figure 14, showed a clear elbow at three clusters, indicating that this was the optimal number of clusters for the data. This result suggests that three distinct clusters effectively capture the variation in review word counts across different countries.

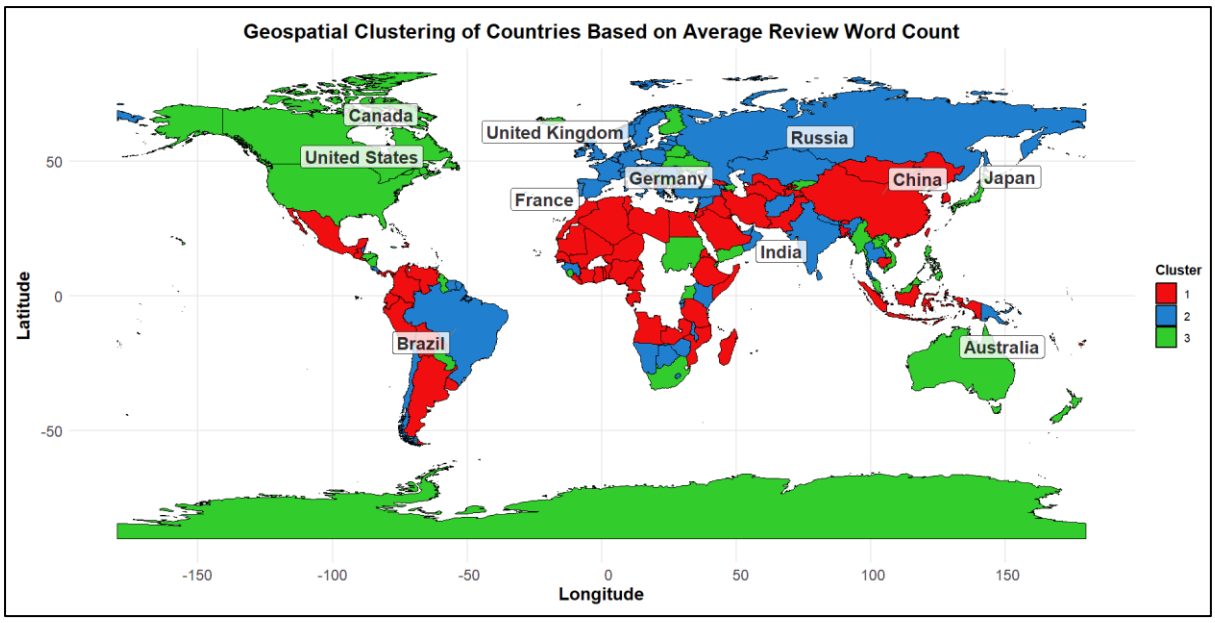

Figure 15: Geospatial Clustering of Countries by Average Review Word Count

The map in Figure 15 illustrates the results of the geospatial clustering analysis. Countries are color-coded based on their assigned cluster, representing groups of countries with similar average review word counts. The analysis identified three distinct clusters:

- The countries in Cluster 1 (red) have the greatest average word counts per review, suggesting that consumers in these areas typically submit reviews that are more in-depth. Major countries including China and Brazil are included in this cluster.
- The countries in Cluster 2 (green) have average word counts that are moderate, indicating a well-rounded approach to review writing. This group includes countries like Australia and the United States.
- The countries in Cluster 3 (blue) had the shortest average reviews, suggesting that consumers in these areas provide shorter feedback. Among the countries in this cluster are Russia, India, etc.

4.6 Analysis of Trip Types

Identifying the different needs and preferences connected to each travel category requires an understanding of the kinds of travels that clients take. Based on the data, this section looks at the most typical trip kinds and shows how the goal of a trip might affect consumer behaviour and the content of hotel reviews.

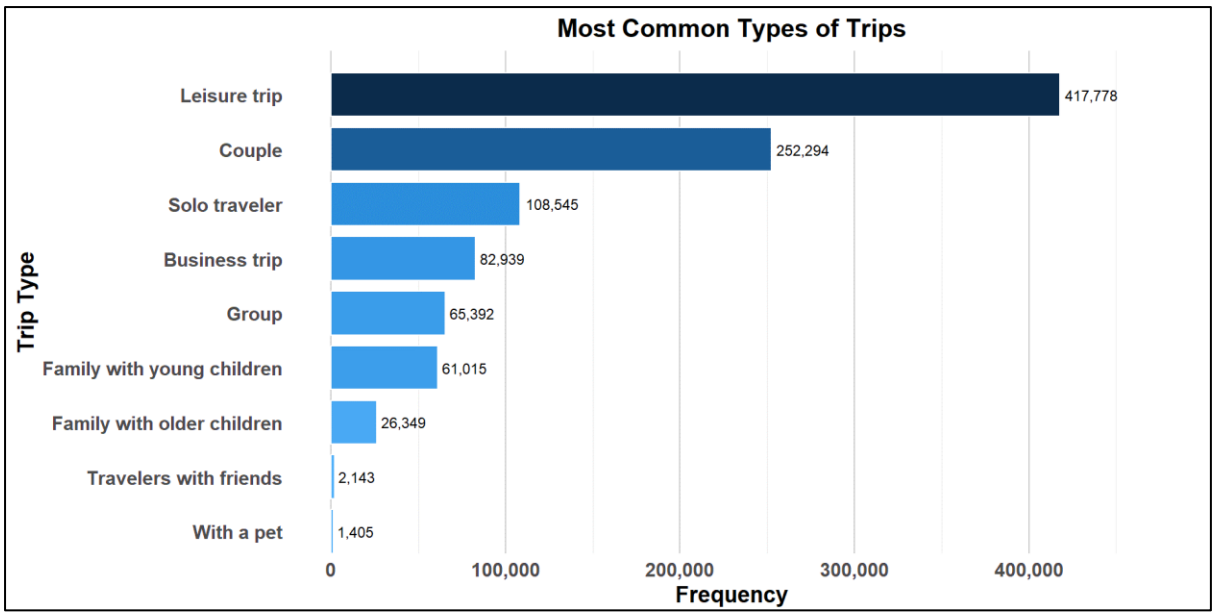

Figure 16: Most Common Types of Trips

This analysis of trip types indicates that leisure and couple trips dominate the dataset, suggesting that many customers are looking for relaxing and romantic experiences.

4.7 Sentiment Analysis

This section presents the results of the sentiment analysis performed on customer reviews, categorized using the NRC Lexicon. The analysis provides a breakdown of the various sentiments expressed in both positive and negative reviews, as well as the distribution of sentiment scores.

4.7.1 Sentiment Distribution in Positive and Negative Reviews (NRC Lexicon)

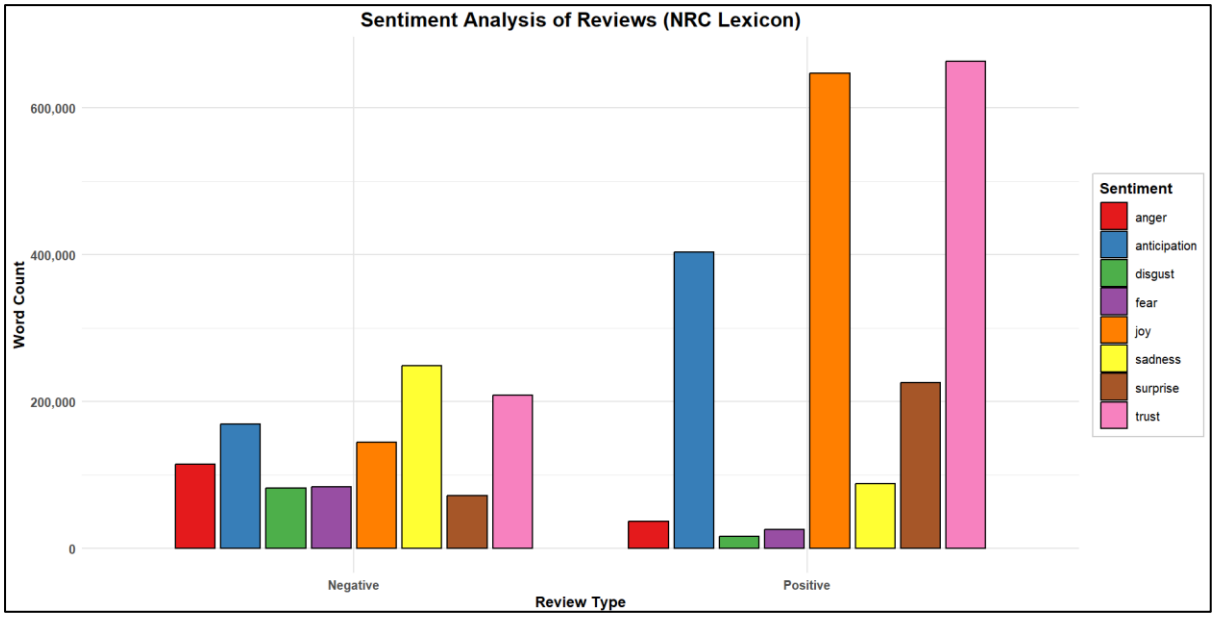

Figure 17: Sentiment Analysis of Reviews (NRC Lexicon)

Figure 17 displays the word count distribution of different sentiments in both positive and negative reviews as categorized by the NRC Lexicon. In positive reviews, the most frequent sentiments are **trust** and **joy**, each exceeding 600,000 words. Negative reviews are primarily associated with **sadness** and **anger**, both with substantial word counts around 200,000.

4.7.2 Distribution of Word Counts by Sentiment

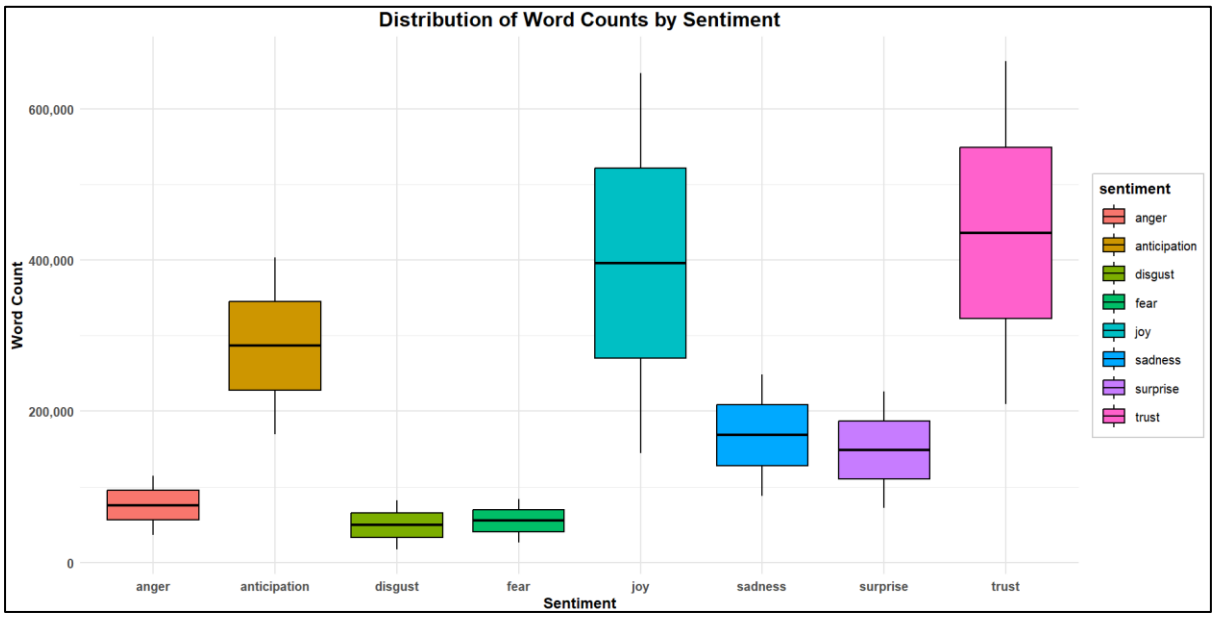

Figure 18: Distribution of Word Counts by Sentiment

Figure 18 shows a box plot of word counts for each sentiment category across all reviews. The box plot highlights the range and variability of word counts for different sentiments. **Joy** and **trust** sentiments show higher median word counts and a wider spread, while **disgust**, **fear**, and **anger** have lower median word counts and less variability.

4.7.3 Distribution of Sentiment Scores for Reviews

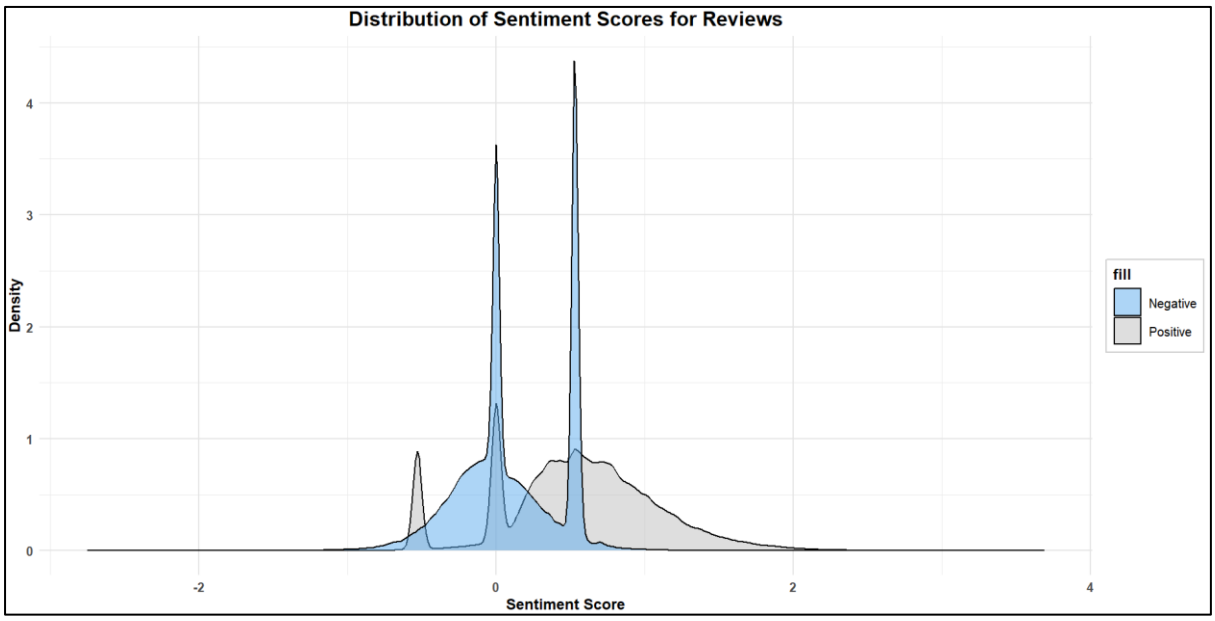

Figure 19: Distribution of Sentiment Scores for Reviews

Figure 19 illustrates the distribution of sentiment scores for both positive and negative reviews using a density plot. The plot reveals two main peaks, one representing positive sentiment scores and the other negative sentiment scores, suggesting a clear distinction between the two types of reviews.

4.7.4 Correlation Between Average Sentiment Scores and Review Scores

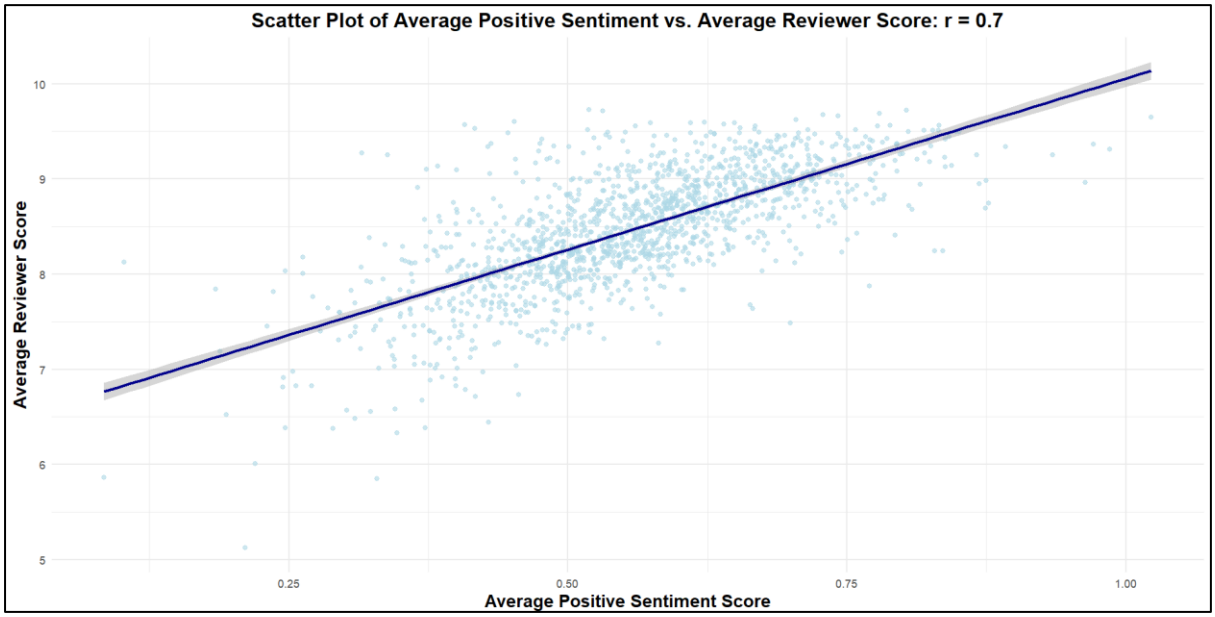

Figure 20: Scatter Plot of Average Positive Sentiment Scores vs. Average Review Scores

Figure 20 shows a scatter plot depicting the relationship between average positive sentiment scores and average review scores for each hotel. There is a positive correlation, indicating that hotels with higher average positive sentiment scores tend to have higher average review scores.

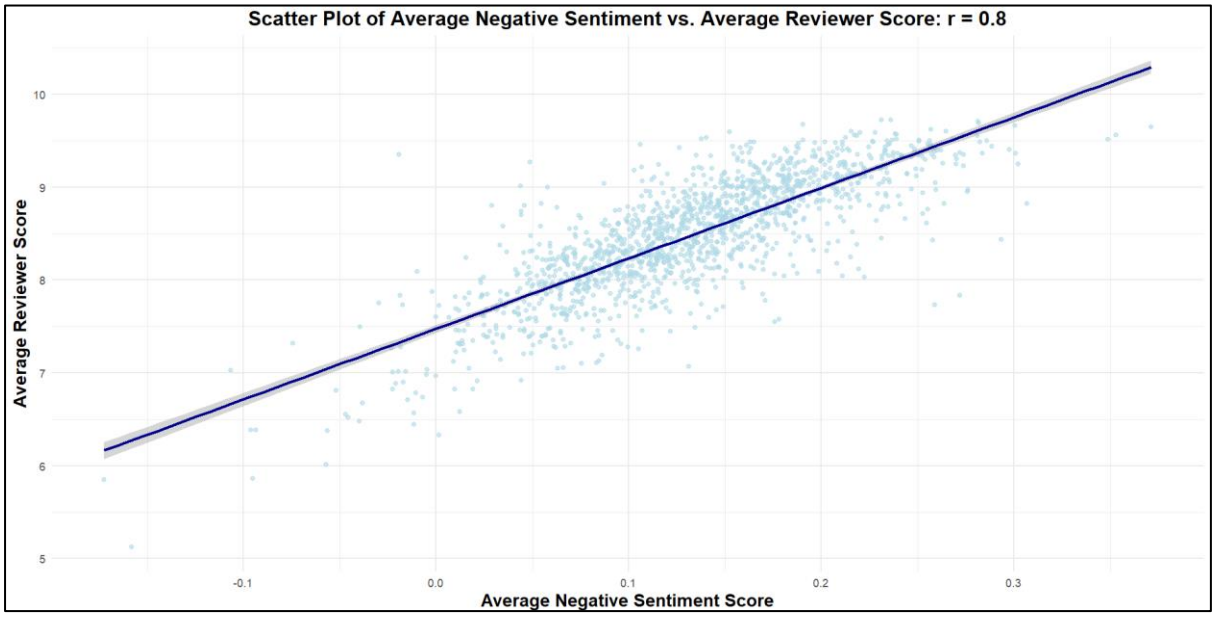

Figure 21: Scatter Plot of Average Negative Sentiment Scores vs. Average Review Scores

Figure 21 presents a scatter plot showing the relationship between average negative sentiment scores and average review scores for each hotel. This plot indicates a positive correlation

between average negative sentiment scores (higher values indicating more positivity) and average review scores.

Sentiment Type	Correlation Coefficient	P-value
Positive Sentiment	0.7	3.74E-215
Negative Sentiment	0.8	< 2.2e-16

Table 2: Correlation Coefficients and P-values Between Sentiment Scores and Review Scores

Table 2 above shows the correlation coefficients and corresponding p-values, confirming a statistically significant relationship between sentiment scores and review scores. The p-values are extremely small ($p < 0.001$), indicating that these correlations are highly unlikely to be due to random chance. A positive correlation for both positive sentiment and negative sentiment suggests that both types of feedback (textual sentiment and numerical rating) align well in reflecting customer satisfaction.

4.9 Topic Modelling

4.9.1 Determination of the Optimal Number of Topics

To determine the optimal number of topics, coherence scores were calculated for various numbers of topics. Coherence scores measure the interpretability and meaningfulness of topics, with higher scores indicating better-defined topics. After evaluating the coherence scores, seven topics were identified as ideal for capturing the main themes in the UK customer reviews.

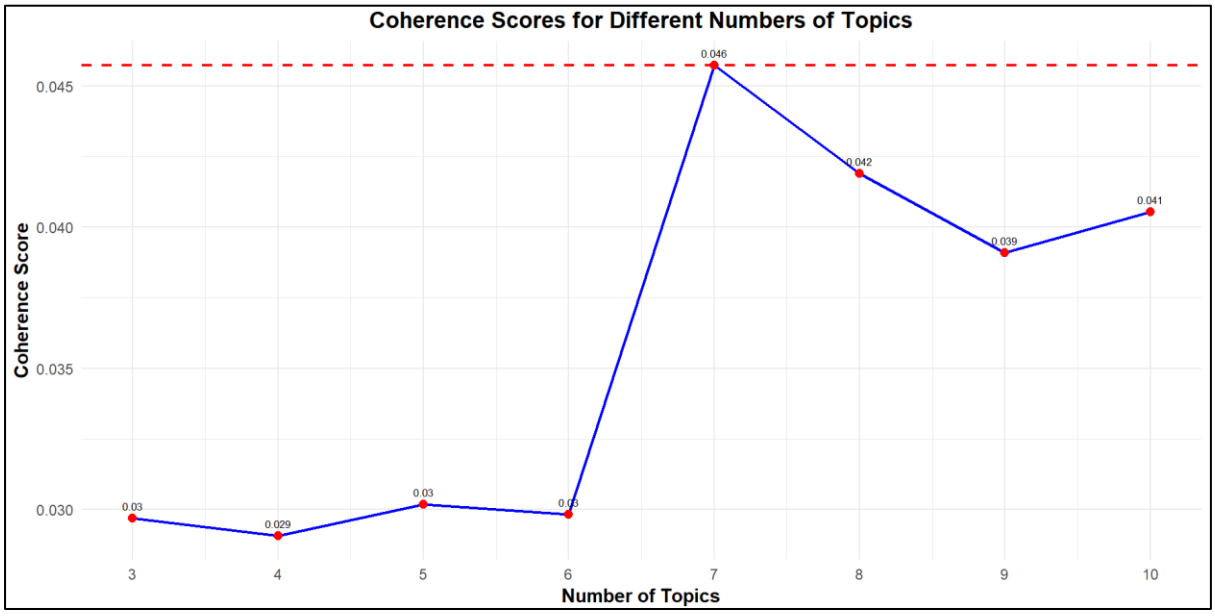

Figure 22: Coherence Score for Different Number of Topics

A total of seven topics were identified through the LDA model. The number of topics was determined by assessing the coherence scores and selecting the model that provided the most interpretable and distinct topics.

4.9.2 Coherence Scores of Topics

Figure 23 illustrates the coherence scores for seven topics generated from the LDA model. Topic 1 has the highest coherence score of 0.094, indicating it is the most interpretable and cohesive topic. Topic 4 also shows a relatively high coherence score of 0.078, suggesting strong internal consistency.

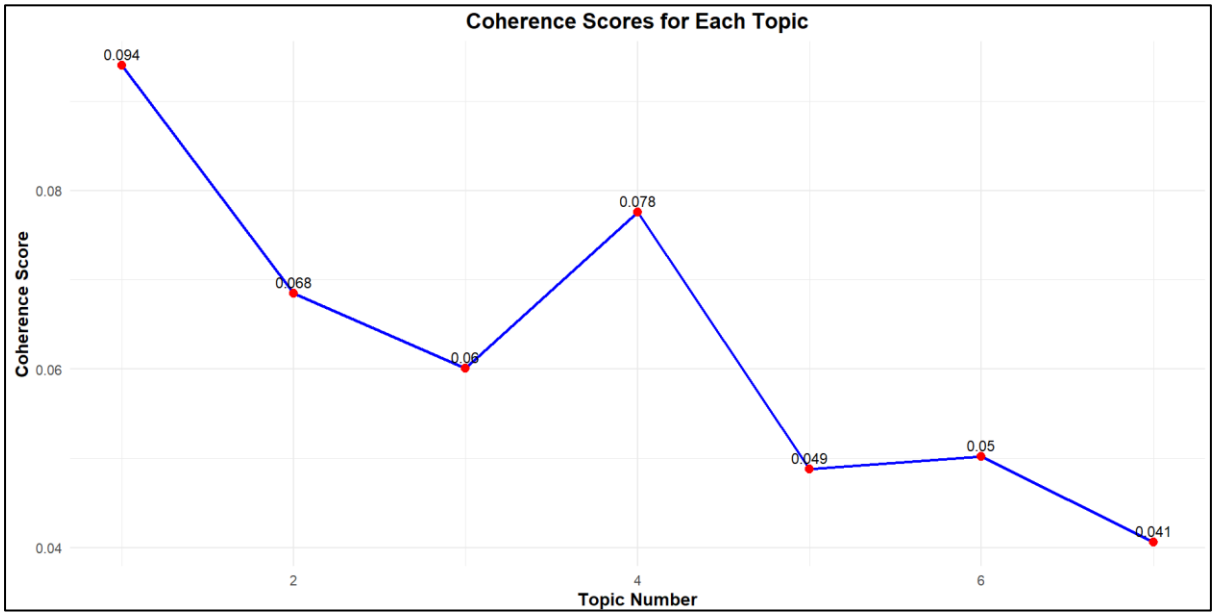

Figure 23: Coherence Scores for Each Topic

In contrast, Topics 5, 6, and 7 have the lowest coherence scores, ranging from 0.041 to 0.050, indicating that these topics are less coherent and potentially harder to interpret. The variability in coherence scores highlights differences in the clarity and interpretability of each topic, with some topics providing clearer thematic insights than others.

4.9.3 Key Topics Identified in UK Reviews

The top 10 terms from each topic, together with the top words connected with each topic and their accompanying coherence scores, are displayed in Table 3. This helps highlight the main points of focus in the reviews and the LDA model's identification of each topic's clarity.

Topic	Coherence Score	Top Terms	Label
1	0.09	get, one, checkin, booking, time, reception, staff, service, waiting, day, arrival	Check-in Process
2	0.07	room, bed, clean, comfort, nice, view, comfy, spacious, great, quiet	Room Comfort & Quality
3	0.06	hotel, location, walk, convenient, close, station, park, city, area, restaurant	Location & Accessibility
4	0.07	staff, experience, location, friend, help, great, excellent, polite, attentive, service	Staff & Service Quality
5	0.05	good, breakfast, bar, service, food, price, amenities, coffee, free, tea	Food & Beverage
6	0.05	hotel, stay, love, comfort, like, everything, experience, really, amazing, perfect	Overall Experience
7	0.04	room, small, bathroom, shower, location, amenities, night, floor, door, cold	Room Size & Amenities

Table 3: Top Words and Coherence Scores in Identified Topics for UK Reviews

Notably, out of all the themes, Topic 1 has the greatest coherence score, making it the most comprehensible and unique. It should be highlighted, however, that all the topics' coherence ratings are somewhat low, indicating that although they all reflect different concepts, there might be some overlap or less clear division between them.

4.9.4 Distribution of Review Themes Across Hotels by UK Nationals

The terms that appeared the most frequently in each category were used to name the topics, which represented different themes of the hotel experience as reported by UK citizens in their reviews. To ensure that the subjects are understandable and pertinent to the examination of customer feedback, each name was selected to best reflect the main idea of the related phrases.

Table 3 below shows the distribution of review themes from customer reviews written by UK nationals for hotels across various countries. The reviews have been categorized into seven distinct themes identified after conducting topic modelling analysis: Check-in Process, Room Comfort & Quality, Location & Accessibility, Staff & Service Quality, Food & Beverage, Overall Experience, and Room Size & Amenities. Each theme captures different aspects of hotel experiences that were frequently highlighted by UK reviewers.

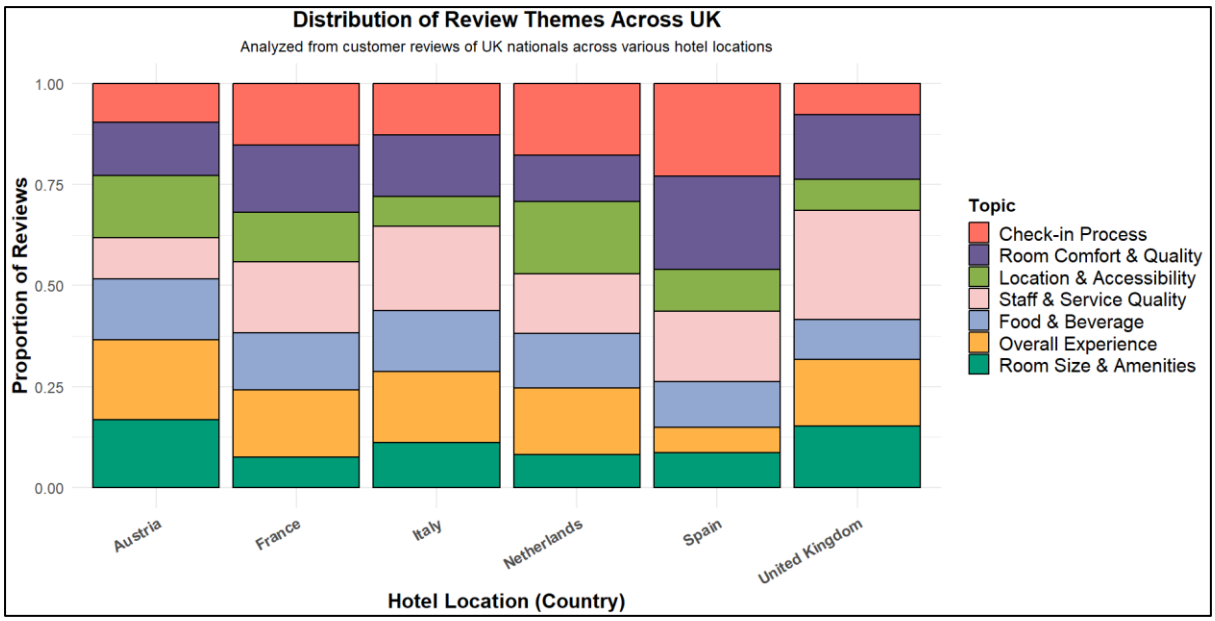

Figure 24: Distribution of Review Themes Across UK Reviewers

This research shows that, depending on the UK citizens have different expectations and experiences. These variations may be caused by regional norms, cultural variances, or individual preferences for hospitality. The thematic differences across countries provide valuable insights for hoteliers looking to cater to the needs and expectations of UK travellers in different regions.

4.10 Sentiment Distribution Across Topics

Figure 25 illustrates the sentiment distribution across different topics in reviews from UK nationals. Each bar represents a topic, divided into positive and negative sentiment proportions.

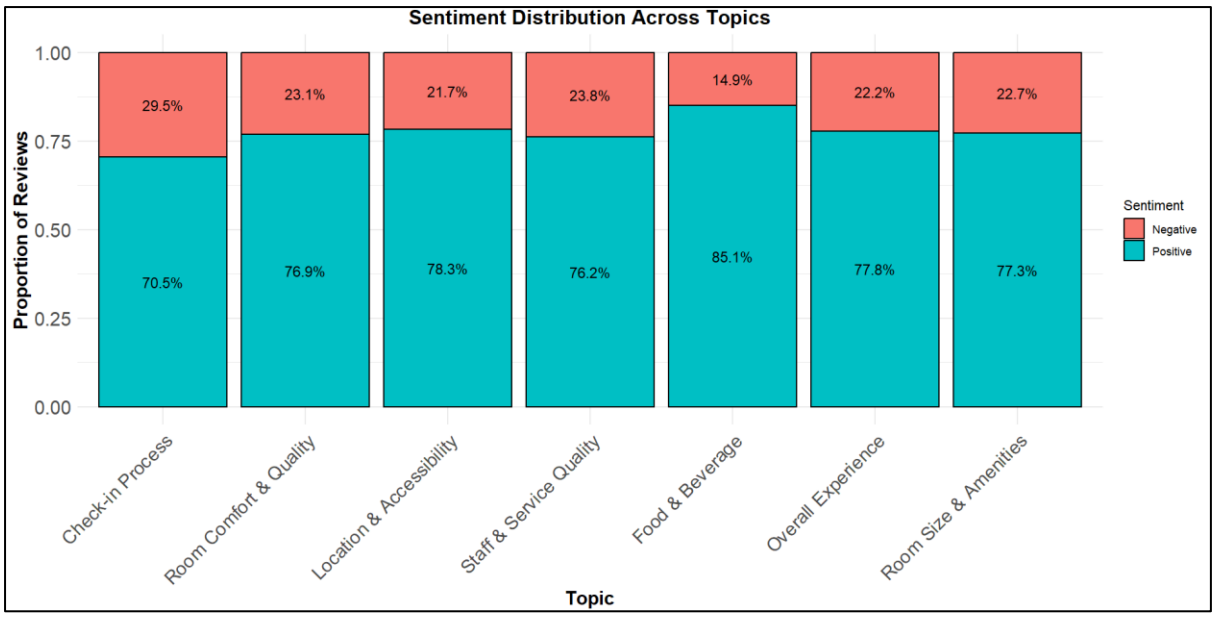

Figure 25: Sentiment Distribution Across Topics

The graph indicates that reviews for most topics are positive. At 85.1%, Topic 5 ("Overall Experience") had the highest percentage of positive feeling. Topic 1 ("Check-in Process"), on the other hand, has the largest negative response (29.5%), suggesting that there may be room for improvement. This distribution shows which parts of the hotel experience are most well-liked and where improvements can be made.

5. Discussion & Analysis

5.1 Interpretation of Results

The discussion section of this dissertation examines and evaluates the study's results considering the initial research questions. In this section, each research question is covered in full, along with an analysis of the findings using sentiment analysis and other data-driven techniques. The results are placed in the larger context of the body of literature, noting the similarities and differences between the study and earlier studies.

The first research question posed in this study was: How do sentiment scores from customer reviews correlate with overall hotel satisfaction ratings? The results indicate a significant positive correlation between sentiment scores derived from customer reviews and hotel satisfaction ratings. Although geographical disparities warrant more investigation, this study implies that the sentimental tone conveyed in customer reviews is a reliable predictor of overall satisfaction.

With a correlation coefficient of 0.70 for positive sentiment scores and 0.8 for negative sentiment scores, both relationships demonstrated a strong linear association with satisfaction ratings. The extremely low p-values associated with these correlations ($p < 0.001$) indicate that these findings are highly statistically significant, meaning the likelihood of these correlations occurring by chance is exceedingly small. This suggests that sentiment analysis provides insightful information about consumer satisfaction, even when limited to sentiment polarity (positive or negative reviews).

These results are consistent with earlier research by Pang and Lee (2008) and Liu (2012), who discovered that sentiment polarity accurately captures consumers' general levels of satisfaction or dissatisfaction in a range of contexts. The statistical significance of these correlations further supports the reliability of using sentiment scores to predict customer satisfaction, aligning well with the findings of these previous studies.

The relationship between sentiment and satisfaction was further illuminated by using the NRC Lexicon to identify sentiments in reviews. According to the data, reviews that expressed positive sentiments like satisfaction or trust were typically linked to greater satisfaction ratings, but reviews that expressed negative sentiments like anger or sadness were linked to lower satisfaction ratings. This discovery aligns with the research conducted by Liu (2012) and

Mohammad and Turney (2013), who emphasised the significance of comprehending the sentimental details in text data to more accurately represent customer sentiment.

Customer reviews from areas where English proficiency is higher generally showed a stronger and more direct relationship between satisfaction ratings and sentiment scores. Reviews from areas with poorer English proficiency, on the other hand, showed weaker connections, because of sentiment expressions that differed from language barriers. According to Nguyen et al. (2019), who underlined the necessity of culturally sensitive approaches in text analysis, these findings imply that linguistic and cultural differences can affect the effectiveness of sentiment analysis.

The findings of the research have a big impact on marketing and management practices in hotels. By comprehending the relationship between satisfaction ratings and sentiment scores, hotel managers may better target their services and pinpoint important areas for development. To provide a more accurate picture of customer fulfilment, hotels in areas where the correlation is weaker, for example, benefit from measures to encourage more in-depth and descriptive reviews. This strategy is in line with Anderson's (2012) suggestions, which stressed the value of thorough feedback in raising the level of operations.

Managers can more effectively evaluate customer feedback and make more informed decisions by acknowledging the cultural variations in review expression. Hotels in areas where English competence is weaker, for instance, would want to think about utilising different approaches for gathering and evaluating client feedback, like employing sentiment analysis techniques that are sensitive to cultural differences or utilising local languages. The results of Nguyen et al. (2019), who stressed the importance of culturally sensitive approaches in text analysis, are consistent with this methodology.

Furthermore, the results indicate that hotels should take note of certain feelings conveyed in the customer reviews, such as joy, trust, and anticipation, as they are closely linked to high rates of satisfaction. Hotels may raise customer satisfaction levels and improve their services by focussing on these important sentimental markers. This strategy is in line with Liu's (2012) suggestions, which emphasised the value of aspect-based sentiment analysis in determining customer satisfaction.

The second research question addressed in this study was: How do the interrelations between sentiment, cultural background, and specific hotel attributes contribute to predicting customer

loyalty and repeat bookings across different geographic regions? The results indicate that sentiment scores derived from customer reviews are strongly correlated with hotel satisfaction ratings, and this relationship varies across different geographic locations, highlighting the significant role of cultural and regional factors in predicting customer loyalty.

Cultural background is a significant factor in this connection. Based on the reviewers' nationalities, the study indicates notable variations in the review's characteristics. For instance, compared to reviewers from non-English-speaking locations, reviewers from the United Kingdom and other English-speaking countries typically provide more thorough feedback with higher average word counts. This variation in review length and level of depth may have an impact on the sentiment analysis results since longer and more detailed reviews may paint a clearer image of customer pleasure, which would increase the sentiment scores' ability to predict customer loyalty. (see Appendix A. for detailed statistics on English-speaking countries).

Geographical location affects how customers express their satisfaction or dissatisfaction, as further evidenced by geospatial analysis. The clustering of countries according to the average word count of reviews revealed clear trends in review behaviour, indicating that cultural influences have a big impact on the way sentiment is conveyed in reviews. Another important aspect that has emerged as impacting customer satisfaction and retention is the significance of hotel features in their experiences. Numerous crucial factors, including location, comfort of the accommodation, and the calibre of the staff, were found to have a major impact on customer satisfaction in a variety of geographical areas.

For instance, topic modelling revealed that “Room Comfort & Quality”, and “Overall Experience” were prominent themes that varied across hotels and regions, reflecting differing customer priorities and expectations. Combining these insights, it is evident that understanding the interplay between sentiment, cultural background, and specific hotel attributes is essential for predicting customer loyalty and repeat bookings. Hotels that consistently receive high sentiment scores, cater to culturally diverse expectations, and focus on attributes that drive satisfaction are more likely to build a loyal customer base.

In this regard, providing thorough information to UK reviewers or prioritising comfort for Chinese visitors are two examples of how services can be tailored to meet the tastes of various cultural groups and improve customer satisfaction and promote repeat business. Managers are better able to understand customer feedback and make more educated decisions when they are

aware of the cultural variations in review expression and feeling. In regions where sentiment scores and satisfaction ratings are not as strongly correlated, hotels may benefit from encouraging more thorough and descriptive reviews to get a better understanding of customer satisfaction. This is in line with Anderson's (2012) recommendations, which highlight the significance of detailed feedback in improving service quality.

Focusing on key sentimental indicators such as joy, trust, and anticipation can help hotels improve their services and boost overall customer satisfaction. By integrating sentiment analysis, cultural understanding, and specific hotel attributes, hotels can better predict customer loyalty and strategically enhance their offerings to meet diverse customer needs, thus fostering a stronger, more loyal customer base.

There were regional differences in the degree to which sentiment scores and satisfaction ratings correlated. For example, compared to hotels in Eastern Europe, sentiment scores and satisfaction ratings showed a larger correlation in Western European hotels. The difference could be explained by cultural variations in the ways that customers communicate their satisfaction or frustration in reviews. This observation is supported by research by Thelwall et al. (2010), which shows that cultural characteristics have a considerable impact on sentiment expression, which in turn influences review scores. Customers from various geographic locations could have distinct tastes and expectations, and these cultural quirks could be reflected in their positive or negative expressions.

The third research question in this study explores the practical implications of sentiment and topic analysis for hotel management and marketing strategies. Specifically, it examines how insights derived from analysing customer reviews can be leveraged by hotels to enhance customer satisfaction and improve their competitive position in the market. The findings from this research indicate that sentiment and topic analysis provide valuable insights into customer experiences and preferences. These insights are crucial for hotel management as they provide a clear indication of the aspects of the customer experience that are most influential in shaping overall satisfaction.

By focusing on enhancing the elements that elicit positive sentiments and addressing the issues that lead to negative sentiments, hotels can improve their service offerings and better meet customer expectations. This approach aligns with the findings of Anderson (2012), who emphasized the importance of understanding customer sentiments in shaping effective service strategies.

In addition to sentiment analysis, topic modelling is essential for pinpointing areas that want improvement. To determine the most often discussed themes and subjects, topic modelling approaches were used in this study to analyse customer feedback. As an illustration, the topic analysis revealed that while “poor maintenance” and “noise levels” were frequently noted in negative reviews, “staff friendliness” and “room cleanliness” were among the most frequently mentioned positive features in highly rated reviews. Hotel management can improve customer satisfaction by prioritising changes in these areas by addressing these topics. This is consistent with the findings of Chevalier and Mayzlin's (2006) study, which showed that resolving customer complaints raised in evaluations can greatly enhance the perception of service quality.

By measuring the frequency and mood of review subjects, hotels may determine which components of their service most influence customer perceptions and direct resources towards those areas. By focussing on certain areas of service enhancement, hotels may increase their return on investment and gain a competitive edge in the industry.

Sentiment analysis and topic modelling not only informs service improvements but also offers useful information for marketing strategy. Hotels should focus their marketing messages to draw potential guests by highlighting the components of the hotel experience that customers appreciate the most. Marketing efforts can highlight the hotel's location and amenities, for example, if the study shows that guests regularly compliment these features, setting it apart from other properties.

Moreover, hotels can use sentiment analysis to find prospective supporters as well as adversaries of their brands within their customer base. High levels of satisfaction and loyalty in reviews can be used as testimonials in marketing initiatives to gain the confidence and trust of potential customers. Negative reviews, on the other hand, can be leveraged to pinpoint areas in need of development and show the hotel's dedication to responding to customer concerns, improving its standing and earning the trust of future guests. The results of Liu (2012), who emphasised the significance of utilising consumer input in developing a strong brand image, lend support to this strategy.

Another significant implication of sentiment and topic analysis is the ability to offer personalized customer experiences. By analysing individual reviews, hotels can gain insights into the specific preferences and expectations of different customer segments. For instance, families may prioritize amenities such as swimming pools and family-friendly activities, while business travellers value proximity to business centres and efficient service. By understanding

these preferences, hotels can tailor their offerings to better meet the needs of each customer segment, thereby enhancing customer satisfaction and loyalty.

The implications of sentiment and topic analysis for hotel management and marketing strategies are well-supported by the existing literature. Previous studies have highlighted the importance of understanding customer sentiments and preferences in shaping effective service strategies. Moreover, research by Chevalier and Mayzlin (2006) and Xiang et al. (2015) has demonstrated the value of analysing customer reviews to identify key areas of improvement and tailor marketing messages to customer preferences.

The fourth research question posed in this study was: “How do the predominant themes and sentiment distributions identified in customer reviews vary across different hotel locations, and what do these variations reveal about the specific areas of improvement needed for different hotel attributes in the UK hospitality industry?”

The analysis of customer reviews written by nationals of the United Kingdom for a variety of hotel locations has shown significant differences in the themes and emotions that are expressed. This emphasises how crucial it is to understand how differences in customer expectations and experiences affect service quality in the hospitality industry.

Several important themes were found through the topic modelling thematic analysis, including "Check-in Process," "Room Comfort & Quality," "Location & Accessibility," "Staff & Service Quality," "Food & Beverage," "Overall Experience," and "Room Size & Amenities." The prevalence of these themes varied according to the hotel's location, reflecting geographic variations in the preferences of its guests. Analysis for hotels in France and Italy, for instance, usually highlighted the "Room Comfort & Quality," but analysis from hotels in the Netherlands and Spain usually talked about the "Food & Beverage" and the "Overall Experience." This difference is consistent with earlier study by Pang & Lee (2008) and Liu (2012) and demonstrates that customer preferences and expectations vary greatly by region, depending on cultural and contextual factors.

The sentiment distribution among these themes highlights the areas in which hotels perform exceptionally well and those that require development. The theme "Overall Experience" had the largest percentage of favourable comments (85.1%), suggesting that visitors were generally more satisfied. On the other hand, the "Check-in Process" received the most percentage of negative feedback (29.5%), indicating a prevalent area in need of improvement. This means

that although hotels usually do a good job of giving customers a positive overall experience, there are some service areas that need improvement, including the check-in procedure.

The findings of the analysis have a big impact on marketing and management practices in hotels. Hotels can raise the standard of their services and better satisfy customers by emphasising aspects that generate positive feelings and correcting problems that result in negative sentiments. This strategy is in line with Anderson's (2012) findings, which stressed the significance of comprehending the customer opinions while developing effective customer service methods.

Topic modelling is crucial in addition to sentiment analysis for identifying areas that need to be improved. For example, the data showed that while "staff friendliness" and "room cleanliness" were among the most often mentioned good qualities in highly rated reviews, themes like "poor maintenance" and "noise levels" were regularly noticed in negative reviews. In line with Chevalier and Mayzlin's findings (2006), these insights enable hotel management to prioritise changes in these areas to improve customer satisfaction.

Hotels can identify which aspects of their service most impact customer feedback and focus resources there by analysing the frequency and sentiments of review subjects. This focused method of improving services can boost return on investment and give businesses a competitive advantage. Furthermore, Liu (2012) emphasises that sentiment analysis assists hotels in identifying prospective supporters and critics among their existing customers, which may be used in marketing initiatives to increase confidence and draw in new business.

The study also highlighted the role of regional and cultural factors in shaping customer feedback. For example, reviews from areas with higher English proficiency showed a stronger correlation between sentiment scores and satisfaction ratings. In contrast, reviews from areas with lower English proficiency exhibited weaker correlations, likely due to language barriers affecting sentiment expression. This finding is consistent with Nguyen et al. (2019), who stressed the need for culturally sensitive approaches in text analysis to accurately capture customer sentiment.

Hotels should think about using sentiment analysis techniques that take cultural differences into account or include evaluations written in the local language in order to improve the efficacy of sentiment analysis across various locations. With the help of this culturally sensitive

approach, hotels will be able to better understand the attitudes and levels of satisfaction of their customers and customise their offerings to suit their varied demands.

Important insights into customer expectations and areas for development in the UK hospitality business can be gained from examining the differences in themes and sentiment distributions across various hotel locations. Through the application of sentiment and topic analysis, hotels may proactively improve their services, accommodate a wide range of customer demands, and ultimately develop increased satisfaction and loyalty.

5.2 Limitations

Although this study offers insightful information about hotel management techniques and customer satisfaction, it should be noted that it has several drawbacks. A primary constraint pertains to the management of the extensive dataset, which, although exhaustive, posed noteworthy obstacles. For instance, in crucial columns like longitude and latitude, there were a lot of missing entries (NA values). The accuracy with which spatial trends and patterns could be analysed was hampered by these omissions in the geographic data, which may have compromised the validity of some of the geospatial analyses.

The dataset used in this study covers the years 2015 to 2017. As a result, the data reflects customer behaviours and satisfaction levels typical of that period. This presents a limitation as it does not account for substantial changes in travel patterns, safety concerns, and customer expectations that may have emerged in more recent years (Gössling, Scott, and Hall, 2020). Especially, after the year 2019, the landscape of travelling changed drastically due to COVID-19, changing the requirements and perceptions of customers. Consequently, while the findings offer insights into customer satisfaction during the specified period, they may not be entirely applicable to the current hospitality landscape, where customer priorities and behaviours have shifted significantly.

Furthermore, some countries, such as the “United States Minor Outlying Islands,” were excluded from the analysis due to insufficient data, which could introduce bias and limit the generalizability of the findings to all geographic regions. Additionally, the reviews primarily come from Booking.com, a specific platform, which may limit the applicability of the findings to other contexts. Differences in user demographics, review submission processes, and platform-specific features could mean that customer feedback on Booking.com does not fully represent feedback on other platforms. This platform-specific focus suggests that while the

findings provide useful insights, they may not be universally applicable across all hotel review platforms or customer bases.

The analysis's computing challenges constituted yet another important constraint. The dataset's size required the use of powerful computing power and extended code execution times, especially when utilising sophisticated machine learning models and sentiment analysis methods. This resulted in difficulties maintaining computational efficiency in addition to taking longer to finish the investigation. Additionally, due to the sparsity of data in some segments, different models yielded similar precision and accuracy scores, making it difficult to determine the most effective approach for certain analyses.

Finally, the study's focus on customer reviews from a specific platform like Booking.com may limit the applicability of the findings to other contexts. Customer feedback on Booking.com may not accurately reflect input on other platforms due to differences in user demographics, review submission procedures, and platform-specific features. This restriction implies that even while the results offer insightful information, they could not be generally applicable to all consumer bases or review sites for hotels. To increase the accuracy and generalisation of the results, future study should consider more recent data, larger geographic coverage, and enhanced computational resources, as suggested by these constraints, which also reveal areas for further improvement.

5.3 Recommendations

The results of this study allow for the formulation of various useful suggestions for hotel administration. Hotels should first concentrate on improving their core service offerings, giving special attention to areas like staff behaviour and room cleanliness, and convenient location when opening a new branch of the chain, that are most frequently linked to positive and negative feelings. Hotels may strengthen their competitive position and raise overall customer satisfaction by addressing these critical service elements.

Second, to draw in possible guests and establish confidence and trust with potential clients, hotels should use favourable evaluations in their marketing campaigns. They should do this by emphasising evaluations that emphasise important features like location and amenities. Third, to show their dedication to customer satisfaction and create a positive brand image, hotels should actively monitor and respond to negative feedback. Targeted promotions and individualised communication can help achieve this and promote repeat business.

Additionally, it is recommended that a similar analysis be conducted on more up-to-date data. Customer preferences and market conditions can change rapidly, and having recent data will help hotel management and marketing teams stay aligned with current trends. By applying sentiment analysis, topic modelling, and other data-driven methodologies to the latest reviews and feedback, hotels can gain timely insights that are more relevant to present-day contexts. This will enable them to adjust their strategies and offerings, accordingly, ensuring that they meet the evolving needs of their guests and maintain a competitive edge in the market.

Ultimately, it is recommended that hotels provide individualised experiences by leveraging insights from sentiment and topic research to customise their services to the distinct preferences and needs of various customer categories. Hotels may boost customer satisfaction, encourage loyalty, and strengthen their position in the market by putting these methods into practice.

6. Conclusion

This dissertation used enhanced text analytics and machine learning approaches to investigate the complex relationship between customer review sentiments and total hotel ratings within the hospitality industry. A large dataset from Booking.com, spanning the years 2015 to 2017, was used in the study to examine customer reviews and gain understanding of the variables affecting customer satisfaction.

The results of the research show a strong positive link between customer review sentiment scores and hotel satisfaction scores in a range of geographical locations. The significance of sentimental tone in reviews as a dependable indicator of overall satisfaction is highlighted by this association. The investigation also revealed geographical disparities in customer satisfaction, including significant changes in sentiment expression and satisfaction ratings according to cultural and regional characteristics.

Additionally, the study found that certain sentiments—like joy and trust—were strongly correlated with greater satisfaction ratings, whereas negative sentiments—like anger and sadness—were linked to lower ratings. Hotel managers may improve client experiences and customise services to meet a range of consumer expectations by using this sophisticated understanding of sentimental impulses.

More thorough and descriptive evaluations offer more distinguishable insights into consumer satisfaction, according to the research, which also examined the significance of review attributes like word count and term frequency. According to this research, hotels may be able to enhance customer satisfaction and enhance customer understanding by promoting thorough feedback.

The study has some limitations even though it makes a substantial contribution to the subject of customer experience management in the hospitality sector. The dataset used was restricted to Booking.com reviews, which might not accurately reflect user opinions on other platforms. Furthermore, as the study period does not cover more recent years, some customer habits and expectations from that time may not be fully reflected in the findings.

To account for the changes brought about by the pandemic, future study could benefit from examining more recent data. It could also benefit from broadening the analysis to include multiple review platforms to provide a more comprehensive perspective of customer

satisfaction. Furthermore, investigating the use of machine learning and sentiment analysis in other service-oriented sectors may yield new insights and advance the domains of client feedback.

In conclusion, this dissertation offers a comprehensive framework for understanding customer satisfaction through sentiment analysis and text mining, providing actionable insights for hotel managers and laying the groundwork for future studies in customer experience management across various sectors.

References

Ali, B., Gardi, B., Othman, B.J. and Ahmed, S.A., 2021. *Hotel service quality: The impact of service quality on customer satisfaction in hospitality* [Online]. *ResearchGate*. Available from: https://www.researchgate.net/publication/351780048_Hotel_Service_Quality_The_Impact_of_Service_Quality_on_Customer_Satisfaction_in_Hospitality.

Anderson, C., 2012. The Impact of Social Media on Lodging Performance Part of the Hospitality Administration and Management Commons [Online]. Available from: https://www.iimagineservicedesign.com/wp-content/uploads/2016/02/Cornell-University-Research-The-Impact-of-Social-Media-on-Lodging-Performance-Positive-Reviews-to-ROI.pdf.

Behrens, J.T., 2024. *APA PsycNet* [Online]. *Apa.org*. Available from: https://psycnet.apa.org/fulltext/1997-06270-001.html.

Bholowalia, P. and Kumar, A., 2014. EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN [Online]. *International Journal of Computer Applications*, 105(9), pp.975–8887. Available from: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=5771aa21b2e151f3d93ba0a5f12d023a0bfcf28b.

Bird, S., Ewan Klein and Loper, E., 2009. *Natural Language Processing with Python* [Online]. *ResearchGate*. O'Reilly. Available from: https://www.researchgate.net/publication/220691633_Natural_Language_Processing_with_Python.

Blei, D.M., Ng, A.Y., Jordan, M.I. and Lafferty, J., 2003. Latent Dirichlet Allocation [Online]. Journal of Machine Learning Research, 3, pp.993–1022. Available from: http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

Bosangit, C., Dulnuan, J. and Mena, M., 2012. Using travel blogs to examine the postconsumption behavior of tourists [Online]. Journal of Vacation Marketing, 18(3), pp.207–219. Available from: https://doi.org/10.1177/1356766712449367.

Browning, V., So, K.K.F. and Sparks, B., 2013. The Influence of Online Reviews on Consumers' Attributions of Service Quality and Control for Service Standards in Hotels

[Online]. Journal of Travel & Tourism Marketing, 30(1-2), pp.23–40. Available from: https://doi.org/10.1080/10548408.2013.750971.

Buhalis, D. and Law, R., 2008. Progress in Information Technology and Tourism management: 20 Years on and 10 Years after the Internet—The State of eTourism Research. Tourism Management, 29(4), pp.609–623.

Cantallops, A.S. and Salvi, F., 2014. New Consumer behavior: a Review of Research on eWOM and Hotels [Online]. *International Journal of Hospitality Management*, 36(1), pp.41–51. Available from: https://doi.org/10.1016/j.ijhm.2013.08.007.

Chambers, J., 2008. *Software for Data Analysis* [Online]. *Statistics and Computing*. New York, NY: Springer New York. Available from: https://doi.org/10.1007/978-0-387-75936-4.

Chen, G.-M. and Starosta, W.J., 2005. *Foundations of intercultural communication | WorldCat.org* [Online]. *search.worldcat.org*. Available from: https://search.worldcat.org/title/foundations-of-intercultural-communication/oclc/300996044.

Filieri, R., 2015. What makes online reviews helpful? A diagnosticity-adoption framework to explain informational and normative influences in e-WOM [Online]. Journal of Business Research, 68(6), pp.1261–1270. Available from: https://doi.org/10.1016/j.jbusres.2014.11.006.

Gössling, S., Hall, C.M. and Andersson, A.-C., 2016. The manager's dilemma: a conceptualization of online review manipulation strategies [Online]. Current Issues in Tourism, 21(5), pp.484–503. Available from: https://doi.org/10.1080/13683500.2015.1127337.

Gössling, S., Scott, D. and Hall, C.M., 2020. Pandemics, tourism and global change: A rapid assessment of COVID-19 [Online]. *Journal of Sustainable Tourism*, 29(1), pp.1–20. Available from: https://doi.org/10.1080/09669582.2020.1758708.

Guttentag, D., 2013. Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector [Online]. Current Issues in Tourism, 18(12), pp.1192–1217. Available from: https://doi.org/10.1080/13683500.2013.827159.

Han, J., Kamber, M. and Pei, J., 2011. *Data Mining Third Edition* [Online]. Available from: https://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf.

Hay, K., 2024. *The Importance of Guest Feedback in Hospitality* [Online]. *Benbria*. Available from: https://benbria.com/guest-feedback-in-hospitality/.

Heimerl, F., Lohmann, S., Lange, S. and Ertl, T., 2014. Word Cloud Explorer: Text Analytics Based on Word Clouds [Online]. *2014 47th Hawaii International Conference on System Sciences*. Available from: https://doi.org/10.1109/hicss.2014.231.

Jain, A.K., 2010. Data clustering: 50 years beyond K-means [Online]. *Pattern Recognition Letters*, 31(8), pp.651–666. Available from: https://doi.org/10.1016/j.patrec.2009.09.011.

Jurafsky, D. and H. Martin, J., 2018. *Speech and Language Processing* [Online]. *Stanford.edu*. Available from: https://web.stanford.edu/~jurafsky/slp3/.

Kiritchenko, S., Zhu, X. and Mohammad, S.M., 2014. Sentiment Analysis of Short Informal Texts [Online]. *Journal of Artificial Intelligence Research*, 50, pp.723–762. Available from: https://doi.org/10.1613/jair.4272.

Leung, D., Law, R., van Hoof, H. and Buhalis, D., 2013. Social Media in Tourism and Hospitality: a Literature Review [Online]. Journal of Travel & Tourism Marketing, 30(1-2), pp.3–22. Available from: https://doi.org/10.1080/10548408.2013.750919.

Little, R. and Rubin, D., 2019. Statistical Analysis with Missing Data, Third Edition [Online]. *Wiley Series in Probability and Statistics*. Available from: https://doi.org/10.1002/9781119482260.

Liu, B., 2012. *Sentiment Analysis and Opinion Mining* [Online]. Available from: https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf.

Liu, B. and Zhang, L., 2012. A Survey of Opinion Mining and Sentiment Analysis [Online]. *Mining Text Data*, pp.415–463. Available from: https://doi.org/10.1007/978-1-4614-3223-4_13.

Liu, J., n.d. *515K Hotel Reviews Data in Europe* [Online]. *www.kaggle.com*. Available from: https://www.kaggle.com/datasets/jiashenliu/515k-hotel-reviews-data-in-europe.

Luca, M., 2016. Reviews, Reputation, and Revenue: The Case of Yelp.com [Online]. papers.ssrn.com. Rochester, NY. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1928601.

Manning, C.D., Raghavan, P. and Schütze, H., 2008. *Introduction to Information Retrieval* [Online]. *Higher Education from Cambridge University Press*. Available from: https://www.cambridge.org/highereducation/books/introduction-to-information-retrieval/669D108D20F556C5C30957D63B5AB65C#overview.

Mariani, M.M. and Borghi, M., 2018. Effects of the Booking.com rating system: Bringing hotel class into the picture [Online]. Tourism Management, 66, pp.47–52. Available from: https://doi.org/10.1016/j.tourman.2017.11.006.

Martins, M.R., Rachão, S. and Costa, R.A. da, 2018. Electronic word of mouth: Does it really matter to backpackers? Booking website reviews as an indicator for hostels' quality services [Online]. *Journal of Quality Assurance in Hospitality & Tourism*, 19(4), pp.415–441. Available from: https://doi.org/10.1080/1528008x.2018.1429980.

Mohammad, S.M. and Turney, P.D., 2012. CROWDSOURCING A WORD-EMOTION ASSOCIATION LEXICON [Online]. Computational Intelligence, 29(3), pp.436–465. Available from: https://doi.org/10.1111/j.1467-8640.2012.00460.x.

Mudambi, S. and Schuff, D., 2010. Research Note: What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com [Online]. MIS Quarterly, 34(1), pp.185–200. Available from: https://doi.org/10.2307/20721420.

Nicewander, A., 1988. *Thirteen Ways to Look at the Correlation Coefficient* [Online]. *The American Statistician*. Informa UK Limited. Available from: https://www.academia.edu/23928382/Thirteen_Ways_to_Look_at_the_Correlation_Coefficient [Accessed 2 September 2024].

O'Connor, P., 2010. Managing a Hotel's Image on TripAdvisor [Online]. Journal of Hospitality Marketing & Management, 19(7), pp.754–772. Available from: https://doi.org/10.1080/19368623.2010.508007.

Öğüt, H., & Onur Taş, B. K. (2012). The influence of internet customer reviews on the online sales and prices in hotel industry. Service Industries Journal, 32(2), 197-214. [Link](https://www.tandfonline.com/doi/abs/10.1080/02642069.2010.529436)

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135. [Link](https://dl.acm.org/doi/10.1561/1500000011)

Pebesma, E., 2018. Simple Features for R: Standardized Support for Spatial Vector Data [Online]. *The R Journal*, 10(1), pp.439–446. Available from: https://journal.r-project.org/archive/2018/RJ-2018-009/index.html.

Phillips, P., Barnes, S., Zigan, K. and Schegg, R., 2016. Understanding the Impact of Online Reviews on Hotel Performance [Online]. *Journal of Travel Research*, 56(2), pp.235–249. Available from: https://doi.org/10.1177/0047287516636481.

Reisinger, Y. and Turner, L., 2024. Cross-Cultural Differences in Tourism: A Strategy for Tourism Marketers [Online]. *Journal of Travel and Tourism Marketing*, pp.79–106. Available from: https://research.monash.edu/en/publications/cross-cultural-differences-in-tourism-a-strategy-for-tourism-mark.

Röder, M., Both, A. and Hinneburg, A., 2015. Exploring the Space of Topic Coherence Measures [Online]. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*, pp.399–408. Available from: https://doi.org/10.1145/2684822.2685324.

Sheng-Hshiung, T., Gwo-Hshiung, T. and Kuo-Ching, W., 1997. Evaluating tourist risks from fuzzy perspectives [Online]. *Annals of Tourism Research*, 24(4), pp.796–812. Available from: https://doi.org/10.1016/s0160-7383(97)00059-5.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M., 2011. Lexicon-Based Methods for Sentiment Analysis [Online]. *Computational Linguistics*, 37(2), pp.267–307. Available from: https://doi.org/10.1162/coli_a_00049.

Tukey, J.W., 2019. Exploratory Data Analysis [Online]. *The Concise Encyclopedia of Statistics*, pp.192–194. Available from: https://doi.org/10.1007/978-0-387-32833-1_136.

Vermeulen, I.E. and Seegers, D., 2009. Tried and tested: The impact of online hotel reviews on consumer consideration [Online]. *Tourism Management*, 30(1), pp.123–127. Available from: https://doi.org/10.1016/j.tourman.2008.04.008.

Wickham, H., 2016. *ggplot2* [Online]. *Use R!* Cham: Springer International Publishing. Available from: https://doi.org/10.1007/978-3-319-24277-4.

Wickham, H., François, R., Henry, L., Müller, K. and Vaughan, D., 2019. *A Grammar of Data Manipulation* [Online]. *Tidyverse.org*. Available from: https://dplyr.tidyverse.org/.

Wilson, T., Wiebe, J. and Hoffmann, P., 2005. *Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis* [Online]. *ACLWeb*. Vancouver, British Columbia, Canada: Association for Computational Linguistics, pp.347–354. Available from: https://aclanthology.org/H05-1044/.

Xiang, Z., Magnini, V.P. and Fesenmaier, D.R., 2015. Information technology and consumer behavior in travel and tourism: Insights from travel planning using the internet [Online]. *Journal of Retailing and Consumer Services*, 22, pp.244–249. Available from: https://doi.org/10.1016/j.jretconser.2014.08.005.

Ye, Q., Law, R. and Gu, B., 2009. The impact of online user reviews on hotel room sales [Online]. *International Journal of Hospitality Management*, 28(1), pp.180–182. Available from: https://doi.org/10.1016/j.ijhm.2008.06.011.

Zervas, G., Proserpio, D. and Byers, J.W., 2017. The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry [Online]. *Journal of Marketing Research*, 54(5), pp.687–705. Available from: https://doi.org/10.1509/jmr.15.0204.

Appendix

A. Detailed Statistics on English-Speaking Countries

Rank	Country	Total Population (millions)	English-Speaking Population (millions)	Percentage of English Speakers
1	United Kingdom	67	67	100%
2	Ireland	5.1	5	98%
3	Australia	26.3	25.6	97%
4	New Zealand	5.1	4.8	94%
5	United States	331	316	95%
6	Canada	39	29.8	76%
7	South Africa	61.2	29.1	48%
8	Nigeria	224	113.4	50.60%
9	Philippines	114.6	58.4	51%
10	India	1412	129.1	9.10%