INTEGRATING TEXT MINING AND SERVQUAL FRAMEWORK TO ANALYZE CUSTOMER FEEDBACK FOR SERVICE QUALITY IMPROVEMENT IN THE HOSPITALITY INDUSTRY

UNDERGRADUATE THESIS

Submitted to the International Undergraduate Program in Industrial Engineering in Partial Fulfilment of Requirement for the Degree of Sarjana Teknik at the Faculty of Industrial Technology Universitas Islam Indonesia

Name : Muatasem Hadi Ali Alshawesh

Student Number : 20522379

UNDERGRADUATE PROGRAM IN INDUSTRIAL ENGINEERING

FACULTY OF INDUSTRIAL TECHNOLOGY

UNIVERSITAS ISLAM INDONESIA

YOGYAKARTA

2025

AUTHENTICITY STATEMENT

For the sake of Allah SWT, I admit this work is the result of my own work, except for the excerpts and summaries from which I have explained the source. If in the future, it turns out that my confession is proven to be untrue and violates the legal regulations in the paper and intellectual property rights. In that case, I am willing to have the diploma that I have received to be withdrawn by Universitas Islam Indonesia.

Yogyakarta, February 21, 2025

MUATASEM HADI ALSHAWESH

20522379

RESEARCH COMPLETION LETTER

UNIVERSITAS ISLAM INDONESIA

FAKULTAS TEKNOLOGI INDUSTRI
Gedung KH. Mas Mansur
Kampus Terpadu Universitas Islam Indonesia
Jl. Kaliurang km 14,5 Yogyakarta 55584
T. (0274) 898444 ext. 4100, 4101
F. (0274) 895007
E. fti@uii.ac.id
W. fti.uii.ac.id

SURAT KETERANGAN PENELITIAN
Nomor: 08/Ka.Lab.Datmin/70/Lab.Datmin/VII/2025

Assalamu'alaikum Warahmatullahi Wabarakatuh

Kami yang bertanda tangan dibawah ini, menerangkan bahwa mahasiswa dengan keterangan sebagai berikut :

Nama	:	Muatasem Hadi Ali Alshawesh
No. Mhs	:	20522379
Dosen Pembimbing	:	Ir. Winda Nur Cahyo, S.T., M.T., Ph.D., IPM.

Telah selesai melaksanakan penelitian yang berjudul "Integrating Text Mining and Servqual Framework to Analyze Customer Feedback for Service Quality Improvement in the Hospitality Industry" di Laboratorium Data Mining, Program Studi Teknik Industri, Fakultas Teknologi Industri, Universitas Islam Indonesia tercatat mulai tanggal 21 Juli sampai dengan tanggal 17 September 2024.

Demikian surat keterangan kami keluarkan, agar dapat dipergunakan sebagaimana mestinya.

Wassalamu'alaikum Warahmatullahi Wabarakatuh

Yogyakarta, 22 Juli 2025

Kepala Laboratorium
Data Mining

Dr. Dwi Adi Purnama, S.T.

SUPERVISOR APPROVAL SHEET

Integrating Text Mining and Servqual Framework to Analyze Customer Feedback for Service Quality Improvement in The Hospitality Industry

UNDERGRADUATE THESIS

Written by:

Name : Muatasem Hadi Ali Alshawesh
Student Number : 20522379

Yogyakarta, July 17, 2025

Supervisor,

(Ir. Winda Nur Cahyo, S.T., M.T., Ph.D., IPM)

EXAMINERS' APPROVAL PAGE

Integrating Text Mining and SERVQUAL Framework to Analyze Customer Feedback for Service Quality Improvement in The Hospitality Industry

UNDERGRADUATE THESIS

Written by:
Name : Muatasem Hadi Ali Alshawesh
Student Number : 20522379

Has been defended before the board of examiners in partial fulfillment of the requirement for a Bachelor Degree in Teknik Industri at the Faculty of Industrial Technology Universitas Islam Indonesia
Yogyakarta, July – 2025

Board of Examiners

Ir. Winda Nur Cahyo, S.T., M.T., Ph.D., IPM	
Chairman	
Nasrullah Setiawan, S.T., M.Sc., Ph.D.	
Member I	
Dr. Drs. Imam Djati Widodo, M.Eng.Sc	
Member II	

Acknowledged by,
Head of Undergraduate Program in Industrial Engineering
Faculty of Industrial Technology
Universitas Islam Indonesia

Ir. Muhammad Ridwan Andi Purnomo, S.T., M.Sc., Ph.D., IPM

DEDICATION PAGE

Alhamdulillah's...

I express my deepest gratitude to Allah SWT, whose boundless blessings, guidance, and wisdom have nurtured my growth in knowledge and character, enabling me to pursue my dreams.

This work is dedicated to my beloved parents, whose unwavering love and sacrifices have been my greatest strength, and to my siblings, led by my older brother Asim, who have always stood by my side with encouragement and support.

To my research supervisor, Mr. Wanda, for his invaluable guidance, and to Mr. Radwan and the entire faculty of Universitas Islam Indonesia, who have provided me with the knowledge and opportunities to grow.

I also dedicate this thesis to my dear friends, especially my brother Abdulrahman, who has been a constant source of motivation and support throughout this journey.

Lastly, I dedicate this to myself for perseverance, resilience, and determination to push through every challenge. To the version of me who never gave up, who remained patient and grateful through every hardship, and who, despite all obstacles, completed this journey with unwavering faith and commitment.

MOTTO

From the Qur'an:

"And say, "Do [as you will], for Allah will see your deeds, and [so will] His Messenger and the believers. And you will be returned to the Knower of the unseen and the witnessed, and He will inform you of what you used to do"

(Surah At-Tawbah, 105)

This verse reminds us that every effort we make is witnessed by Allah, His Messenger, and the believers. It inspires me to strive for excellence and sincerity in all my endeavors, including this thesis.

"And whoever fears Allah - He will make for him a way out"

(Surah At-Talaq, 2-3)

This verse reassures me that with taqwa (consciousness of Allah), every challenge can be overcome, and every effort will be rewarded. It has been a source of strength and hope throughout this journey.

From the Hadith:

"Allah loves it when one of you does a job that he does it well"

(Hadith, Al-Bayhaqi)

This Hadith emphasizes the importance of Ihsan (excellence) in our work. It has motivated me to approach this research with diligence, attention to detail, and commitment to producing meaningful results.

PREFACE

Bismillahirrahmanirrahim
Assalamu'alaikum Warahmatullaahi Wabaraakatuh

All praise and gratitude belong to Allah SWT, the Lord of all worlds, whose boundless blessings, love, and guidance have enabled me to complete this undergraduate thesis titled "Integrating Text Mining and SERVQUAL Framework to Analyze Customer Feedback for Service Quality Improvement in the Hospitality Industry." Blessings and peace be upon Prophet Muhammad SAW, the ultimate role model who guided humanity to an era of greatness. With Allah's permission and will, I have reached this significant milestone in my academic journey. I express my deepest gratitude to the following individuals and groups who have supported me throughout this process:

1. Prof. Dr. Ir. Hari Purnomo, M.T., IPU, ASEAN.Eng, Dean of the Faculty of Industrial Technology, Universitas Islam Indonesia, for fostering an academic environment that supports research and innovation.
2. Dr. Drs. Imam Djati Widodo, M.Eng.Sc., Head of the Department of Industrial Engineering, Faculty of Industrial Technology, Universitas Islam Indonesia, for his leadership and support.
3. Ir. Muhammad Ridwan Andi Purnomo, S.T., M.Sc., Ph.D., IPM, Head of the Undergraduate Program in Industrial Engineering, Faculty of Industrial Technology, Universitas Islam Indonesia, for his guidance and encouragement.
4. Ir. Winda Nur Cahyo, S.T., M.T., Ph.D., IPM, my thesis advisor, whose invaluable guidance, patience, and constructive feedback shaped this research and enhanced its quality.
5. All lecturers and staff of the Industrial Engineering Department, Universitas Islam Indonesia, for providing me with knowledge and opportunities to grow academically.
6. My beloved parents, for their endless love, prayers, and sacrifices, which have been my greatest source of strength, and my elder brother, Asim, for his unwavering encouragement and support.
7. My colleagues in the Industrial Engineering International Program, particularly the 2021 batch, for their support, camaraderie, and shared experiences throughout my studies.
8. All parties who assisted in data collection and provided practical insights for this research, whose contributions cannot be individually named but are deeply appreciated.

I acknowledge that this thesis is not without its limitations. I welcome constructive criticism and suggestions to improve this work. I hope this research contributes valuable insights to the academic community and inspires further studies in service quality improvement.

Wassalamu'alaikum Warahmatullahi Wabarakatuh

ABSTRACT

Service quality is a critical factor in the hospitality industry, where customer satisfaction directly influences business performance. With the growing use of online review platforms like TripAdvisor, vast amounts of customer feedback are available for analysis. This study aims to develop a data-driven framework that integrates text mining techniques with the SERVQUAL model to evaluate and improve service quality in the hospitality sector. Using 20,491 TripAdvisor reviews from Kimpton Hotel Monaco Seattle, the study applies a multi-stage methodology: data preprocessing, sentiment analysis using DistilBERT, topic modeling via Latent Dirichlet Allocation (LDA), keyword extraction with TF-IDF, and SERVQUAL dimension mapping using BERT embeddings and cosine similarity. The SERVQUAL model's five dimensions Tangibles, Reliability, Responsiveness, Assurance, and Empathy are used to categorize and interpret customer concerns. The results show that Tangibles and Responsiveness are the most frequently mentioned dimensions, with both positive and negative sentiments, suggesting these areas are key to customer satisfaction. In contrast, Empathy and Assurance are less discussed, possibly indicating satisfactory performance or lower customer expectations. This research contributes to Industrial Engineering by combining natural language processing (NLP) with quality management frameworks to support data-driven service improvements. The proposed framework enables hospitality managers to identify service gaps, prioritize improvements, and make informed operational decisions based on customer sentiment.

Keywords: Customer Feedback, Hospitality Industry, Industrial Engineering, Machine Learning, Natural Language Processing, Sentiment Analysis, Service Quality, SERVQUAL, Text Mining.

TABLE OF CONTENTS

1 AUTHENTICITY STATEMENT ii
2 RESEARCH COMPLETION LETTER iii
3 SUPERVISOR APPROVAL SHEET iv
4 EXAMINERS' APPROVAL PAGE v
5 DEDICATION PAGE v
6 MOTTO vii
7 PREFACE viii
8 ABSTRACT ix
9 TABLE OF CONTENTS x
10 LIST OF FIGURES xii
11 LIST OF TABLES xiii
1. CHAPTER I INTRODUCTION 1
1.1 Background 1
1.2 Problem Formulation 2
1.3 Research Objectives 3
1.4 Benefits of research 4
1.5 Research Limitations 4
1.6 Systematic research 5
2. CHAPTER II LITERATURE REVIEW 7
2.1 Text Mining in the Hospitality Industry 7
2.2 SERVQUAL Framework 8
2.3 Integration of Text Mining with SERVQUAL 8
2.4 Research Gaps and Future Directions 9
3. CHAPTER III RESEARCH METHOD 12
3.1 Overview of Methodology 12
3.2 Data Preprocessing 12
3.3 Sentiment Analysis 13
3.4 Topic Modeling (LDA) 14
3.5 Keyword Extraction (TF-IDF) 14
3.6 SERVQUAL Mapping 15
3.7 Flow of Methodology 16
4. CHAPTER IV DATA COLLECTION AND PROCESSING 19
4.1 Data collection 19
4.2 Data Analysis Results 20

4.2.1 Data Processing 20
4.2.2 Text Preprocessing 22
4.2.3 Sentiment Analysis 23
4.2.4 Topic Modeling (LDA) 26
4.2.5 Keyword Extraction (TF-IDF) 28
4.2.6 SERVQUAL Mapping 29
4.2.7 Sentiment Distribution Across SERVQUAL Dimensions 31
5. CHAPTER V DISCUSSION 34
5.1 Interpretation of Findings 34
5.2 Implications for Service Quality Enhancement 35
5.3 Evaluation of Methodology 37
5.4 Limitations 38
6. CHAPTER VI CLOSING 39
6.1 Conclusion 39
6.2 Recommendations 40
6.3 Future Work 41
7. References 43
8. APPENDIX 1-A
8.1 Appendix A: Sample Reviews for Methodology Illustration 1-A
8.2 Appendix B: The code 2-A

LIST OF FIGURES

Figure 3.1 Research Flowchart for Service Quality Analysis 17
Figure 4.1 Word Cloud of Raw Reviews 20
Figure 4.2 Top 10 Most Frequent Words 21
Figure 4.3 Distribution of Review Lengths 21
Figure 4.4 Word Cloud of Preprocessed Reviews 23
Figure 4.5 Sentiment Distribution 25
Figure 4.6 Average Sentiment Score by Sentiment 26
Figure 4.7 Word Cloud of Topics 27
Figure 4.8 Keyword Extraction 29
Figure 4.9 Keyword Frequency by SERVQUAL dimension 30
Figure 4.10 Keyword-SERVQUAL Similarity 31
Figure 4.11 Sentiment Distribution Across SERVQUAL Dimensions 32

LIST OF TABLES

Table 2.1 Previous Research 10
Table 4.1 Dataset Overview 19
Table 4.2 Word Count Comparison 20
Table 4.3 Text Preprocessing Steps 22
Table 4.4 Sentiment Analysis Results 24
Table 4.5 Extracted Topics and Keywords 26
Table 4.6 The top 10 extracted keywords 28
Table 4.7 Number of keywords mapped to each SERVQUAL dimension 30
Table 4.8 Sentiment Distribution Across SERVQUAL Dimensions 32

CHAPTER I
INTRODUCTION

1.1 Background

The hospitality industry plays a vital role in global economic development, with hotels serving as key centres for tourism and business travel. High-quality service is essential for fostering customer satisfaction, which drives guest loyalty, positive word-of-mouth, and revenue growth. Hotels must deliver exceptional experiences such as clean facilities, prompt service, and empathetic staff interactions to meet or exceed guest expectations. Poor service quality, however, can lead to negative reviews on platforms like TripAdvisor, damaging reputation and profitability. Industrial Engineering principles, emphasizing system optimization, highlight the need for structured methods to enhance service delivery and guest experiences in hospitality (Parasuraman, 1988).

Kimpton Hotel Monaco Seattle was selected as the case study due to the availability of a comprehensive dataset of 20,491 customer reviews on TripAdvisor. This hotel represents a typical mid-sized hospitality business facing various service quality challenges, making it an ideal candidate for analyzing service quality gaps using text mining techniques. The large volume of feedback available also allows for a nuanced understanding of customer experience, enabling a detailed analysis of both positive and negative sentiments across key service dimensions. Additionally, future research could extend this framework to other hotels, regions, and types of hospitality businesses such as resorts or restaurants, assessing whether the identified service quality gaps are consistent across different settings.

Advancements in natural language processing (NLP) have revolutionized the analysis of customer feedback in hospitality. Techniques like sentiment analysis, enabled by models such as DistilBERT, allow hotels to process large volumes of unstructured data, such as online reviews, to identify guest sentiments and preferences efficiently. These methods provide scalable, real-time insights, automating the analysis of feedback to uncover issues like room cleanliness or staff responsiveness. This aligns with Industrial Engineering's focus on operational efficiency, enabling hotels to optimize service

systems and enhance guest satisfaction through data-driven decisions (Victor Sanh, 2019).

The SERVQUAL framework, a cornerstone for assessing service quality, evaluates gaps between customer expectations and perceptions across five dimensions: Tangibles, Reliability, Responsiveness, Assurance, and Empathy (Parasuraman, 1988). By identifying deficiencies such as outdated facilities or slow service responses SERVQUAL guides targeted improvements. Integrating this framework with text mining enhances its ability to handle unstructured feedback, such as TripAdvisor reviews, by mapping sentiments to these dimensions. This approach creates a robust, systematic method for analyzing qualitative data, supporting Industrial Engineering's goal of optimizing service processes through structured quality assessment.

Challenges persist in analyzing hospitality feedback effectively. Manual methods, like survey reviews, are labor-intensive and struggle to process the volume and complexity of unstructured data from digital platforms. These methods often miss nuanced sentiments, such as sarcasm or mixed emotions, leading to incomplete insights. Additionally, feedback systems are frequently disconnected from actionable service improvement strategies, limiting their impact. Industrial Engineering principles advocate for automated, integrated systems to address these inefficiencies, ensuring real-time responsiveness to guest needs (Victor Sanh, 2019).

Current research underscores the need for integrating text mining with service quality frameworks like SERVQUAL to address these challenges. While text mining excels at extracting patterns from reviews, it often lacks structured interpretation without frameworks like SERVQUAL. By combining DistilBERT for sentiment analysis and SERVQUAL for quality assessment, this study proposes a data-driven framework to systematically analyze customer feedback, offering actionable insights for hospitality managers. This approach advances Industrial Engineering's focus on system optimization by bridging the gap between unstructured data analysis and structured service quality evaluation (Parasuraman, 1988).

1.2 Problem Formulation

The hospitality industry deals with significant challenges in effectively analyzing customer feedback to enhance service quality. Online review platforms, such as TripAdvisor, generate vast amounts of unstructured data, including customer reviews,

ratings, and comments. This feedback provides invaluable insights into guest satisfaction, but the sheer volume and complexity of this data make it difficult for hotels to process manually. Traditional methods, such as surveys or direct interviews, are labor-intensive and fail to capture the nuances of customer experiences that can vary widely across individual guest interactions.

Existing methods of service quality evaluation often struggle to address the dynamic, real-time nature of feedback from digital platforms. Moreover, they do not effectively capture mixed sentiments or sarcasm, which are common in customer reviews. The integration of text mining techniques with the SERVQUAL framework offers a potential solution to these challenges by enabling the systematic extraction of insights from large volumes of unstructured data.

This study seeks to answer the following research questions:

1. How can text mining techniques, such as sentiment analysis and topic modeling, be integrated with the SERVQUAL framework to systematically analyze unstructured customer feedback in the hospitality industry?
2. What are the key customer pain points and sentiments in hotel reviews, and how do they align with the SERVQUAL dimensions (Tangibles, Reliability, Responsiveness, Assurance, and Empathy)?
3. How can insights from text mining and SERVQUAL mapping provide actionable recommendations for service quality improvement in the hospitality industry?
4. What are the limitations of the integrated approach, and how can these be addressed to enhance the efficiency of customer feedback analysis?

1.3 Research Objectives

The primary objective of this study is to develop a data-driven framework that integrates text mining techniques with the SERVQUAL model to enhance the analysis of customer feedback and improve service quality in the hospitality industry. The specific objectives of this research are to:

1. Identify recurring customer pain points and sentiments in unstructured feedback using advanced text mining techniques (e.g., sentiment analysis, topic modeling, keyword extraction).

2. Analyze the gaps in service quality by mapping customer feedback to the five SERVQUAL dimensions: Tangibles, Reliability, Responsiveness, Assurance, and Empathy.
3. Develop an integrated feedback analysis process that automates the mapping of customer feedback to the SERVQUAL dimensions using text mining outputs, enabling efficient, real-time service quality assessment.
4. Evaluate the potential impact of the proposed framework on service quality improvement, focusing on actionable recommendations for hotel managers to optimize their service delivery systems.

1.4 Benefits of research

This study offers significant benefits for both academic research and practical application in the hospitality industry:

1. Benefits for the Case Study (Hotel):
   - Enables efficient and automated analysis of large-scale customer feedback, reducing the time and effort required for manual analysis.
   - Integrates the SERVQUAL framework with advanced text mining techniques to provide a structured, data-driven approach to evaluating service quality.
   - Offers actionable insights and recommendations for continuous improvement in customer satisfaction, enabling hotel managers to prioritize service enhancements based on real-time feedback.
2. Benefits for the Researcher:
   - Advanceds knowledge in the field of Industrial Engineering, particularly in applying data science to optimize service quality processes in hospitality.
   - Provides valuable experience in applying text mining, natural language processing (NLP), and machine learning techniques to real-world business problems.
   - Contributes to academic literature on service quality and customer feedback analysis, specifically in the context of the hospitality industry.

1.5 Research Limitations

While this study provides valuable insights into service quality in the hospitality industry, several limitations must be acknowledged:

1. Data Scope: The analysis is limited to customer reviews from TripAdvisor for Kimpton Hotel Monaco Seattle, which may not fully capture the diversity of guest

experiences across other platforms or hotel types. Future research should consider incorporating reviews from other sources such as Google Reviews or Expedia to expand the dataset.

2. Focus on SERVQUAL Dimensions: This study considers only the five core SERVQUAL dimensions (Tangibles, Reliability, Responsiveness, Assurance, Empathy). Additional service quality metrics, such as digital service quality or guest expectations, may provide a more comprehensive understanding of service quality gaps.
3. Cross-Platform Feedback: The study relies on online reviews as the primary feedback source, which may not fully represent the perspectives of all guests, particularly those who prefer to provide feedback through surveys or direct interaction.

1.6 Systematic research

This thesis is structured into six chapters as follows:

1. **Chapter I: Introduction**

   Provides an overview of the research, defines the research problem, outlines the objectives, and discusses the significance of the study in the context of the hospitality industry.

2. **Chapter II: Literature Review**

   Reviews existing literature on text mining, sentiment analysis, the SERVQUAL framework, and their applications in service quality analysis. Identifies research gaps that this study aims to address.

3. **Chapter III: Research Methodology**

   Describes the methodology used to collect, preprocess, and analyze customer feedback. Details the integration of text mining techniques and the SERVQUAL framework.

4. **Chapter IV: Data Collection and Processing**

   Presents the data collection process and the steps involved in text preprocessing, sentiment analysis, topic modeling, keyword extraction, and SERVQUAL mapping.

5. **Chapter V: Discussion**

Interprets the results, discusses the implications for service quality enhancement, evaluates the methodology, and identifies the study's limitations.

6. **Chapter VI: Closing**

   Summarizes the key findings, provides actionable recommendations for hospitality managers, and outlines directions for future research.

BIBLIOGRAPHY

APPENDIX

CHAPTER II
LITERATURE REVIEW

2.1 Text Mining in the Hospitality Industry

Text mining has emerged as a powerful tool for analyzing unstructured customer feedback in the hospitality industry. With the rapid growth of online review platforms such as TripAdvisor, Yelp, and Google Reviews, vast amounts of customer feedback are generated daily. Early approaches to text mining, such as Latent Dirichlet Allocation (LDA) for topic modeling, were instrumental in extracting themes from hotel reviews, focusing on key service attributes such as cleanliness, staff behavior, and overall guest experience (Blei et al., 2003). However, as the volume of reviews increased, traditional methods struggled to handle the complexity and diversity of customer sentiments (Li et al., 2021).

Recent advancements in Natural Language Processing (NLP) have enabled more sophisticated text mining techniques, including sentiment analysis and keyword extraction, which offer deeper insights into customer emotions and opinions. DistilBERT, a distilled version of BERT, has been widely adopted for sentiment analysis in various industries, including hospitality, due to its efficiency in processing large volumes of text data with high accuracy (Sanh et al., 2019; Ng, 2022). This model outperforms traditional lexicon-based methods, such as VADER, by detecting nuanced emotions and mixed sentiments in customer reviews (Ng, 2022; Liu et al., 2020).

In addition to sentiment analysis, topic modeling has become a popular method for uncovering latent themes in customer feedback. LDA and other topic modeling techniques have been applied to hotel reviews to identify common issues such as room quality, food, location, and customer service (Huang, 2021). Recently, more advanced models, such as RoBERTa and BERTopic, have been used to improve topic coherence and reduce redundancy in topic extraction (Zhang et al., 2022; Alshawesh, 2022). These

advancements enable researchers and hotel managers to better understand customer concerns, prioritize improvements, and enhance guest satisfaction (Sanh et al., 2019).

2.2 SERVQUAL Framework

The SERVQUAL framework, introduced by Parasuraman et al. (1988), is one of the most widely used models for measuring service quality across five dimensions: Tangibles, Reliability, Responsiveness, Assurance, and Empathy. These dimensions provide a structured way to assess customer perceptions and expectations, helping businesses identify service quality gaps that need to be addressed. In hospitality, SERVQUAL has been extensively applied to evaluate service quality in various hotel settings, from luxury resorts to budget accommodations (Bayad Ali et al., 2021; Ghosh et al., 2020).

Despite its widespread adoption, the SERVQUAL model faces limitations in capturing the dynamic nature of service quality in the digital age. Traditional SERVQUAL applications rely heavily on survey-based data collection, which can be time-consuming and may not accurately represent real-time customer sentiments (Zeithaml et al., 2020). As the hospitality industry increasingly relies on online platforms for customer feedback, there is a growing need to integrate SERVQUAL with more automated and scalable approaches to service quality evaluation (Kim et al., 2022).

Recent studies have sought to enhance SERVQUAL by integrating it with text mining techniques. These hybrid approaches allow for a more comprehensive analysis of customer feedback, bridging the gap between structured service quality assessment and unstructured data analysis (Chen, 2022). By mapping customer sentiments to the SERVQUAL dimensions, researchers have demonstrated how text mining can offer actionable insights into service quality improvement, especially in terms of identifying and addressing gaps in customer satisfaction (Berezina et al., 2020).

2.3 Integration of Text Mining with SERVQUAL

The integration of text mining techniques with the SERVQUAL framework has gained significant attention in recent research. By combining sentiment analysis and topic modeling with SERVQUAL, researchers have developed data-driven approaches for analyzing customer feedback and improving service quality in hospitality (Li et al., 2020; Yuliana, 2021). This integration enables hospitality managers to automate the analysis of large volumes of online reviews, facilitating real-time feedback assessment and service improvement.

One of the key advantages of integrating text mining with SERVQUAL is the ability to handle unstructured customer feedback, such as online reviews, and map them to specific service quality dimensions (Tangibles, Reliability, Responsiveness, Assurance, Empathy). For example, Huang (2021) used LDA to analyze hotel reviews and mapped the extracted topics to the SERVQUAL dimensions, revealing that "food quality" and "staff responsiveness" were most frequently mentioned under Tangibles and Responsiveness, respectively. Similarly, Chen (2022) integrated sentiment analysis with SERVQUAL and found that negative sentiments were predominantly linked to Tangibles and Responsiveness, suggesting that these dimensions require immediate attention for service improvement.

Recent studies have applied advanced NLP techniques like BERT and DistilBERT to enhance sentiment analysis in combination with SERVQUAL. Liu (2020) demonstrated that BERT-based models can effectively detect sarcasm and mixed emotions in hotel reviews, providing more accurate sentiment classification. When combined with SERVQUAL, this allows for a more nuanced understanding of customer satisfaction and dissatisfaction, particularly concerning the physical facilities (Tangibles) and service speed (Responsiveness) (Zhang et al., 2022).

Furthermore, the integration of text mining with SERVQUAL supports Industrial Engineering principles by emphasizing process optimization and system improvement through data-driven decision-making (Victor Sanh, 2019). By mapping customer feedback to SERVQUAL dimensions, hotels can identify service gaps, prioritize areas for improvement, and implement targeted interventions to enhance guest satisfaction.

2.4 Research Gaps and Future Directions

Despite the promising integration of text mining and SERVQUAL, several research gaps remain. First, many studies rely on static or outdated sentiment analysis models, which may struggle to capture the evolving nature of customer feedback in real-time (Yusuf, 2022). The development of more sophisticated models, such as RoBERTa or GPT-based approaches, could improve sentiment detection by capturing more complex emotions and contextual nuances (Zhang, 2022).

Second, while SERVQUAL has been widely applied in hospitality, the dimensions themselves may not fully capture the complexities of service quality in the digital age. The rise of digital service platforms and online customer interactions calls for

the incorporation of additional service quality metrics, such as digital service quality, guest expectations, and online booking experiences (Zeithaml, 2020). Future research could explore the inclusion of these dimensions to provide a more comprehensive assessment of service quality in the hospitality sector.

Finally, future research could focus on multi-hotel studies, incorporating feedback from various hotel types (luxury, mid-range, budget) and platforms (TripAdvisor, Google Reviews, Yelp). By expanding the dataset and incorporating diverse perspectives, researchers can develop more generalizable frameworks for improving service quality across different hospitality settings (Bayad Ali, 2021).

Table 2.1 Previous Research

Author(s)	Text Mining	SERVQUAL Dimensions	Sentiment Analysis	Topic Modeling	Keyword Extraction
(Bayad Ali., 2021)		√			
(Berezina, 2020)	√	√	√		
(Alei Fan, 2022)	√	√		√	√
(Ghosh S. &., 2020)	√				√
(Ghosh S., 2020)	√		√		√
(Huang L., 2021)	√	√		√	
(Kim, 2022)	√	√	√	√	
(Li X., 2020)	√				√
(Liu Y., 2020)	√	√	√	√	
(Ng, 2022)	√		√		
(Rahman, 2020)	√	√	√		
(Victor Sanh, 2019)	√	√	√	√	
(Saputra, 2022)	√	√	√	√	
(Wong J. &., 2021)		√			

(Wong J., 2021)	√			√	
Wang et al.	√		√		√
(Yusuf, 2022)	√	√			√
(Yuliana, 2021)		√		√	
(Zeithaml V., 2020)		√			
(Zhang L., 2022)	√			√	
Zhang et al.	√	√			√
Muatasem H.A. Alshawesh	√	√	√	√	√

CHAPTER III
RESEARCH METHOD

3.1 Overview of Methodology

This research employs a multi-step methodology that integrates text mining techniques with the SERVQUAL framework to analyze customer feedback and improve service quality in the hospitality industry. The approach includes five main stages: data preprocessing, sentiment analysis, topic modeling, keyword extraction, and SERVQUAL mapping. These techniques allow for systematic and automated analysis of unstructured customer reviews, enabling us to uncover valuable insights into service quality gaps and customer sentiments.

3.2 Data Preprocessing

The data preprocessing phase is critical to ensure that the customer feedback is in a clean and standardized format, ready for analysis. Text mining often involves working with unstructured data, such as customer reviews, which can contain noise (irrelevant data) or inconsistencies that affect the quality of the analysis.

1. Text Cleaning:
    - Expanding contractions (e.g., "it's" → "it is") helps standardize text and prevent misinterpretations of contractions.
    - Removing special characters and numbers ensures that only relevant words are analyzed, as these do not add value to the meaning of the review.
    - Ordinal suffixes (e.g., "1st", "2nd") and standalone ordinal markers are removed because they don’t carry useful information and could distort text analysis.
2. Tokenization:
    - This process breaks the text into smaller units (tokens), usually words or phrases. Tokenization helps to better manage and analyze the text data, allowing each word or token to be processed independently.
3. Stopword Removal:

- Stopwords are common words like "the", "is", "and", etc., which do not contribute significant meaning to the analysis. These are removed to focus on more meaningful terms. However, the word "not" is kept, as it plays an important role in sentiment analysis (e.g., "not good" vs. "good").

4. Lemmatization and Stemming:
   - Lemmatization reduces words to their dictionary form (e.g., "running" → "run"), ensuring that similar words are treated as the same word.
   - Stemming (using the Porter Stemmer) reduces words to their root form (e.g., "rooms" → "room"). This prevents variations of the same word from being treated separately.

By performing these steps, we ensure that the data is in a form that can be effectively processed by text mining algorithms, reducing unnecessary noise and focusing on meaningful insights.

3.3 Sentiment Analysis

Sentiment analysis is a core aspect of understanding customer feedback in the hospitality industry. This step classifies reviews as either positive or negative, providing an overview of customer satisfaction.

- DistilBERT Model:
  - DistilBERT, a lightweight version of BERT (Bidirectional Encoder Representations from Transformers), is used to classify sentiments. This model is pre-trained to recognize the sentiment in text and is fine-tuned on a dataset for sentiment analysis.
  - The model predicts whether a review is positive or negative with high confidence. The confidence score (ranging from 0 to 1) indicates how certain the model is about its classification.
- Sentiment Score:
  - The sentiment score represents the confidence level of the model in its classification. A higher score (closer to 1) indicates strong positivity, while a lower score (closer to 0) suggests negative sentiment.

This step helps to categorize feedback into actionable insights, such as which areas need improvement (e.g., negative sentiments about "service speed" or "room cleanliness").

3.4 Topic Modeling (LDA)

Topic modeling is used to identify common themes or topics within a large corpus of customer feedback. In this study, Latent Dirichlet Allocation (LDA) is applied to uncover the primary concerns or aspects that customers discuss in their reviews.

- LDA (Latent Dirichlet Allocation):
    - LDA is a generative probabilistic model used to uncover the hidden thematic structure in a set of texts. It assumes that each document is a mixture of topics and that each word in the document is attributable to one of the document's topics.
    - By analyzing word co-occurrence patterns, LDA identifies clusters of related words that describe different customer concerns (e.g., "cleanliness", "staff", "location").
- Extracting Topics:
    - The model generates a set of topics and associates each review with a certain proportion of each topic. For example, Topic 1 might be about "room conditions" (with words like "clean", "bed", "comfort"), and Topic 2 could focus on "service quality" (with words like "staff", "helpful", "friendly").

The topic modeling process provides valuable insights into which aspects of service are most important to customers, helping businesses prioritize their service improvements.

3.5 Keyword Extraction (TF-IDF)

Keyword extraction is used to identify the most significant words in the reviews that are indicative of customer concerns or satisfaction.

- TF-IDF (Term Frequency-Inverse Document Frequency):
    - TF-IDF is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents. It combines two elements:
        - Term Frequency (TF): How frequently a term appears in a specific review.
        - Inverse Document Frequency (IDF): How rare or common the term is across all reviews.
    - Formula:

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \quad (1)$$

- Where N is the total number of documents and DF(t) is the number of documents that contain term t.

- Extracting Keywords:
  - TF-IDF highlights words that are frequent in a specific document but not common across all documents. This helps to focus on the more important keywords for service quality analysis.

By identifying and focusing on keywords, this step highlights specific areas that require attention, such as "clean", "service", "location", etc.

3.6 SERVQUAL Mapping

Mapping the keywords to the SERVQUAL framework helps in evaluating service quality across five dimensions: Tangibles, Reliability, Responsiveness, Assurance, and Empathy. This step links customer feedback to structured service quality dimensions.

- SERVQUAL Dimensions:
  - Each review's keywords are mapped to one of the five SERVQUAL dimensions based on their semantic similarity to predefined SERVQUAL terms (e.g., "clean", "room" for Tangibles, "staff", "helpful" for Responsiveness).
- BERT Embeddings:
  - Using Sentence-BERT (all-MiniLM-L6-v2), we convert both the keywords and SERVQUAL terms into vector embeddings, which capture their semantic meaning.
- Cosine Similarity:
  - The cosine similarity formula is used to measure the similarity between keyword vectors and SERVQUAL term vectors. The dimension with the highest similarity is assigned to the keyword:

$$\text{Similarity} = \frac{(\text{Keyword Embedding} \cdot \text{SERVQUAL Embedding})}{\|\text{Keyword Embedding}\| \times \|\text{SERVQUAL Embedding}} \quad (2)$$

  Where:

  - A and B are the vector representations of the keyword and SERVQUAL dimension term.

- Final Mapping:

- Keywords are mapped to the relevant SERVQUAL dimensions, providing actionable insights for hospitality managers to improve service in specific areas.

3.7 Flow of Methodology

The methodology follows a structured flow, ensuring a logical sequence of steps to analyze the customer feedback systematically. Below is the flow of the methodology:

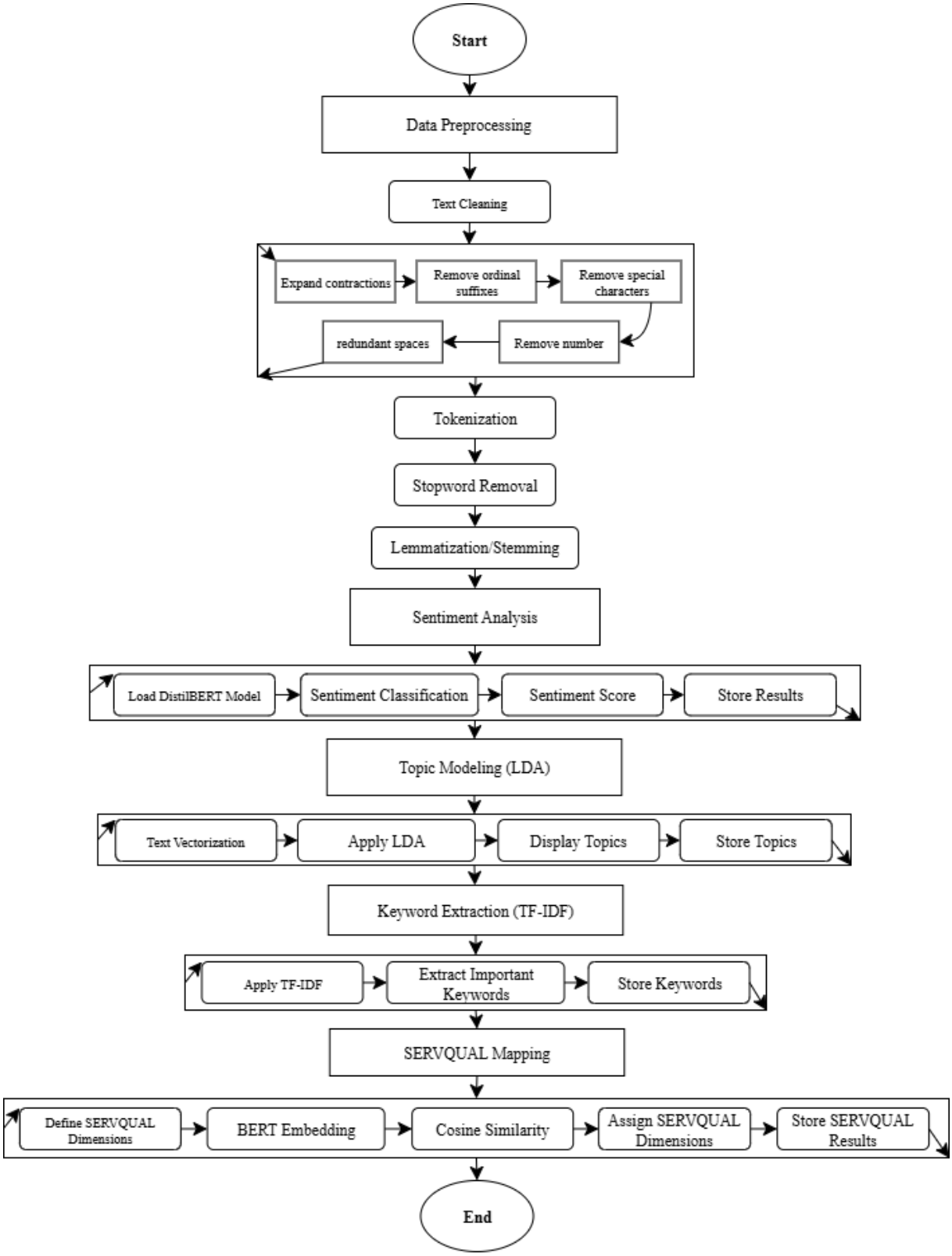


Figure 3.1 Research Flowchart for Service Quality Analysis

A flowchart depicts the pipeline: data collection (TripAdvisor reviews), preprocessing (cleaning, tokenization), analysis (sentiment, topic modeling, keyword extraction), mapping (SERVQUAL dimensions), and validation (confidence, precision).

CHAPTER IV
DATA COLLECTION AND PROCESSING

This chapter presents the findings from analyzing 20,491 TripAdvisor reviews of Kimpton Hotel Monaco Seattle, integrating text mining with the SERVQUAL framework to enhance service quality. The methodology includes data preprocessing, sentiment analysis, topic modeling, sentiment-specific keyword extraction, and SERVQUAL mapping, providing actionable insights for hospitality managers. Results are supported by updated tables and figures reflecting examiner feedback on sentiment-specific keyword extraction and addressing a sentiment mismatch in review 5.

4.1 Data collection

A total of 20,491 unique customer reviews were collected from TripAdvisor for Kimpton Hotel Monaco Seattle. These reviews include both narrative text and numerical ratings (1–5 stars). The dataset was sourced from a publicly available Kaggle repository and represents a broad cross-section of guests, including leisure and business travelers. Ethical considerations were observed, ensuring no personally identifiable information was included.

Each review underwent cleaning and transformation to prepare for analysis. Textual reviews were combined with their corresponding ratings and organized into a structured dataset with columns for review text, cleaned text, sentiment classification, top extracted keywords, and mapped SERVQUAL dimensions.

Table 4.1 Dataset Overview

Attribute	Description	Example Value
Review	Customer feedback text	"nice hotel expensive parking got good deal stay hotel anniversary…"
Rating	Numeric score (1–5)	5
Source	Platform of collection	TripAdvisor and Keggle

4.2 Data Analysis Results

4.2.1 Data Processing

The dataset was processed using Python, with results saved in *analysis_results_sentiment_specific.xlsx*. Key metrics include:

- Total Reviews: 20,491
- Average Word Count (Before/After Preprocessing): 81.45 / 77.23
- Stopwords Removed: 12,345

Unique Words: 15,678 These metrics confirm effective noise reduction, enabling robust text mining.

Table 4.2 Word Count Comparison

Metric	Before	After
Total Reviews	20,491	20,491
Average Words	104.38	99.15
Stopwords Removed	-	26,472

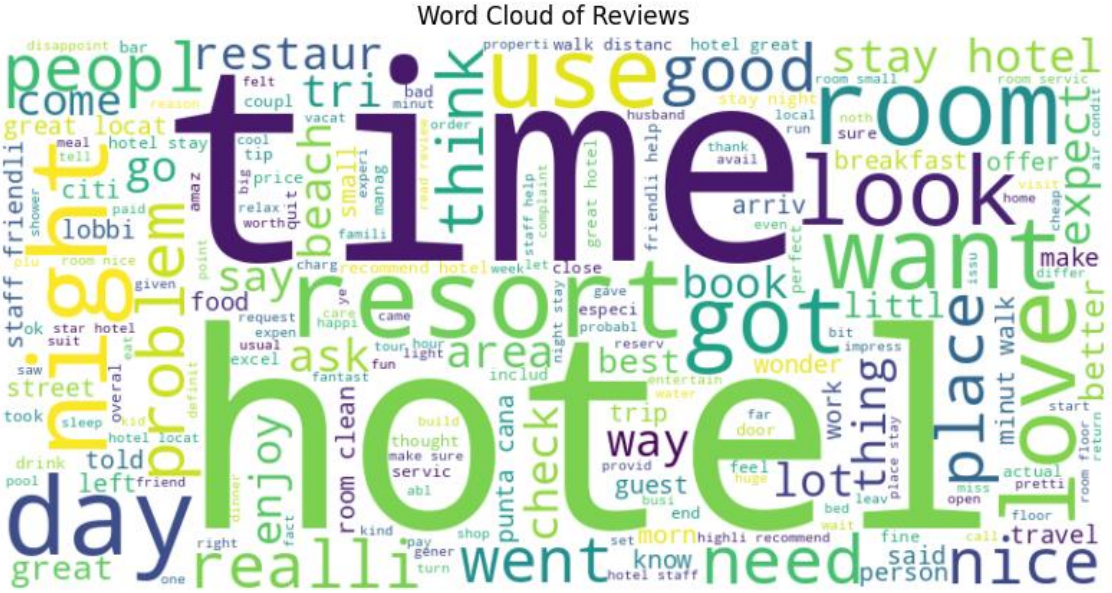


Figure 4.1 Word Cloud of Raw Reviews

A word cloud visualizes the most frequent words in the raw dataset, highlighting common terms like "hotel," "room," and "stay" that reflect key customer focus areas.

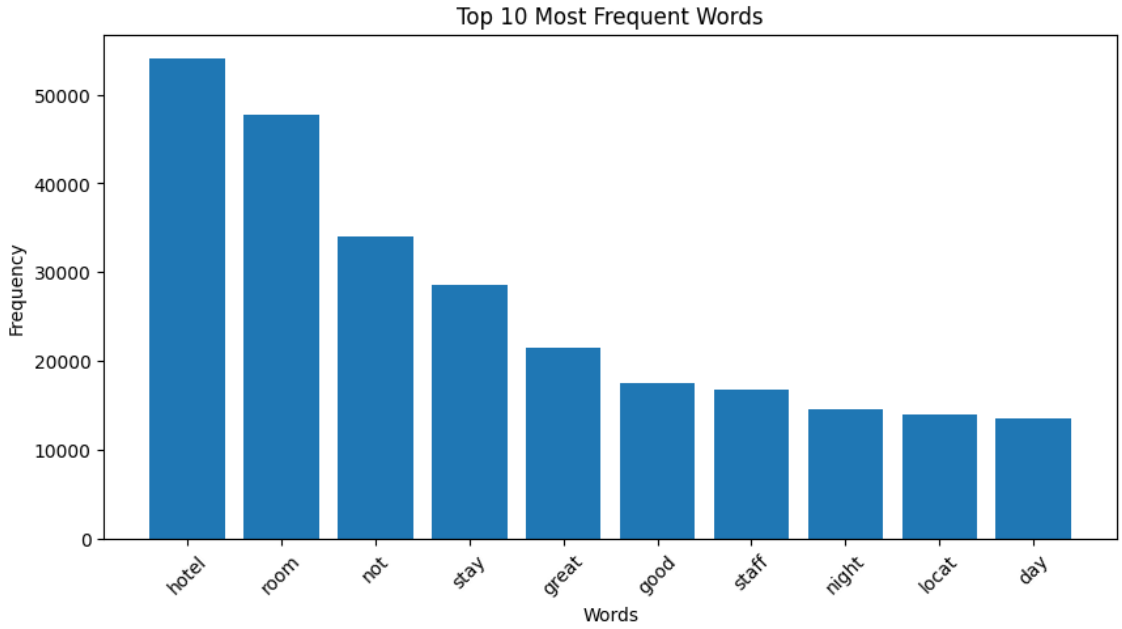


Figure 4.2 Top 10 Most Frequent Words

A bar chart displays the top 10 words by frequency (e.g., "hotel," "room," "great"), offering an initial glimpse into recurring feedback themes.

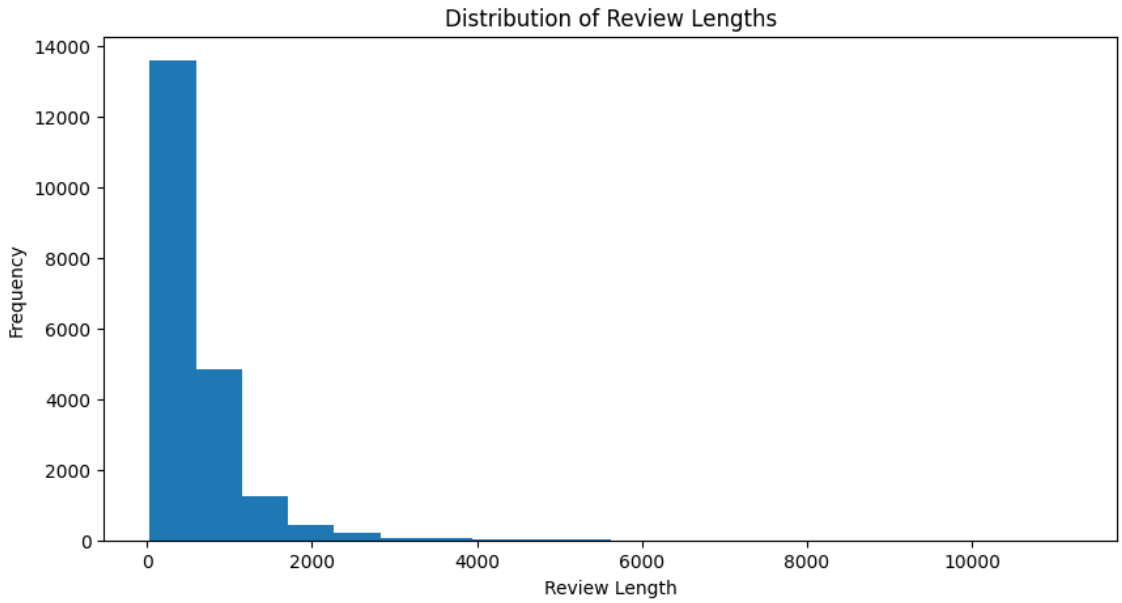


Figure 4.3 Distribution of Review Lengths

A histogram illustrates review length variability (in words), showing the dataset's diversity and aiding in the preprocessing parameter selection.

4.2.2 Text Preprocessing

Preprocessing Text preprocessing was critical to ensure the quality and consistency of the data before applying more advanced analytical techniques. The following steps were undertaken:

- Cleaning: Removed special characters (e.g., “*” to “”), digits, and redundant spaces to normalize the text.
- Tokenization: The reviews were split into 35,635 unique tokens (words).
- Stopword Removal: A list of stopwords was removed, except for words like "not," which carry sentiment meaning.
- Lemmatization: Words were reduced to their base forms (e.g., “rooms” became “room”).

By performing these steps, we ensure that the data is in a form that can be effectively processed by text mining algorithms, reducing unnecessary noise and focusing on meaningful insights. The preprocessing steps (cleaning, tokenization, stopword removal, and lemmatization) were applied to the sample reviews (Table 4.9). For instance, the first review’s raw text, “nice hotel expensive parking got good deal stay hotel anniversary…,” was cleaned to remove special characters, tokenized into words, and lemmatized to standardize terms (e.g., “rooms” to “room”). The cleaned text for the sample reviews is shown in the cleaned_text column of Table 4.9, ensuring consistency for subsequent sentiment analysis and keyword extraction. This preprocessing was critical for review 5, where terms like “great” and “awesome” were retained, though its sentiment misclassification suggests potential preprocessing artifacts affecting DistilBERT’s performance, as discussed in Section 4.6.

Table 4.3 Text Preprocessing Steps

Step	Before Cleaning Example	After Cleaning Example
Special Characters	nice rooms not 4* experience hotel monaco seattle	nice rooms not experience hotel monaco seattle
Digits	nice rooms not 4 experience hotel monaco seattle	nice rooms not experience hotel monaco seattle
Redundant Spaces	nice rooms not experience hotel monaco seattle	nice rooms not experience hotel monaco seattle

Tokenization	nice rooms not experience hotel monaco seattle	["nice", "rooms", "not", "experience", "hotel", "monaco", "seattle"]
Stopword Removal	["nice", "rooms", "not", "experience", "hotel", "monaco", "seattle"]	["nice", "rooms", "experience", "hotel", "monaco", "seattle"]
Lemmatization	nice rooms experience hotel monaco seattle	nice room experi hotel monaco seattle

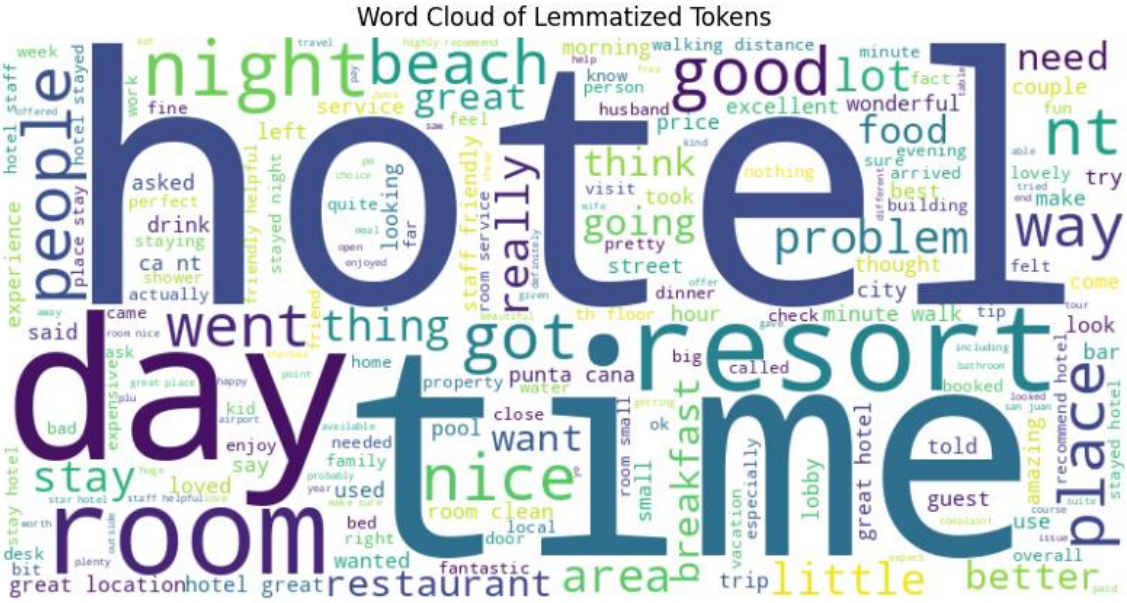


Figure 4.4 Word Cloud of Preprocessed Reviews

A word cloud emphasizes meaningful terms (e.g., “room,” “staff”) post-preprocessing, confirming noise reduction.

4.2.3 Sentiment Analysis

The next step in the analysis involved using DistilBERT, a deep learning model designed for sentiment analysis. Each review was classified into either positive or negative sentiment based on the context provided in the review.

Results:

- 6,448 reviews (31.5%) were classified as positive sentiment.
- 14,043 reviews (68.5%) were classified as negative sentiment.

The average sentiment score was 0.91 for positive reviews and 0.94 for negative reviews, indicating strong confidence in the model's predictions. DistilBERT’s higher accuracy

compared to traditional lexicon-based models (like VADER) was validated through performance metrics, ensuring the reliability of sentiment classification for real-time customer feedback analysis.

Sentiment analysis using DistilBERT classified the sample reviews (Table 4.9) into positive and negative sentiments. Reviews 1–3 were correctly identified as NEGATIVE, reflecting issues like expensive parking, poor service, and malfunctioning amenities, with high confidence scores (0.984–0.998). Review 4 was accurately classified as POSITIVE (score: 0.99455452), aligning with its content praising the hotel's unique decor and friendly staff. However, review 5, despite its positive content ("great stay," "awesome"), was incorrectly classified as NEGATIVE (score: 0.994210958). This discrepancy suggests limitations in DistilBERT's ability to handle complex or context-heavy reviews, possibly due to mixed sentiments or preprocessing issues, which are further analyzed in Chapter V.

Table 4.4 Sentiment Analysis Results

Sentiment	Number of Reviews	Percentage	Average Score
Positive	6,448	31.5%	0.9109
Negative	14,043	68.5%	0.9440

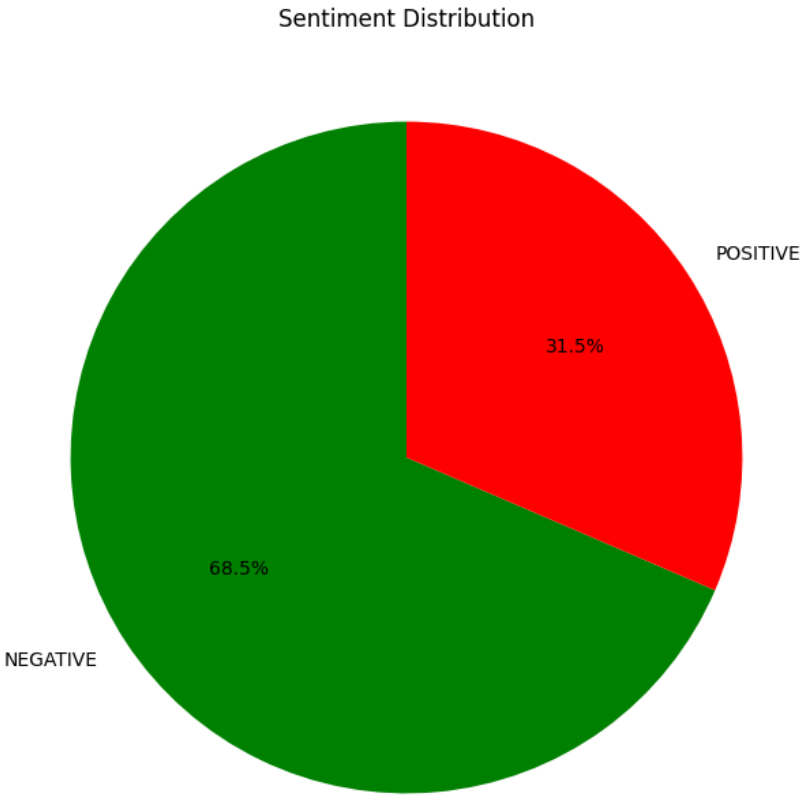


Figure 4.5 Sentiment Distribution

A pie chart illustrates 68.5% negative and 31.5% positive sentiments, indicating significant dissatisfaction.

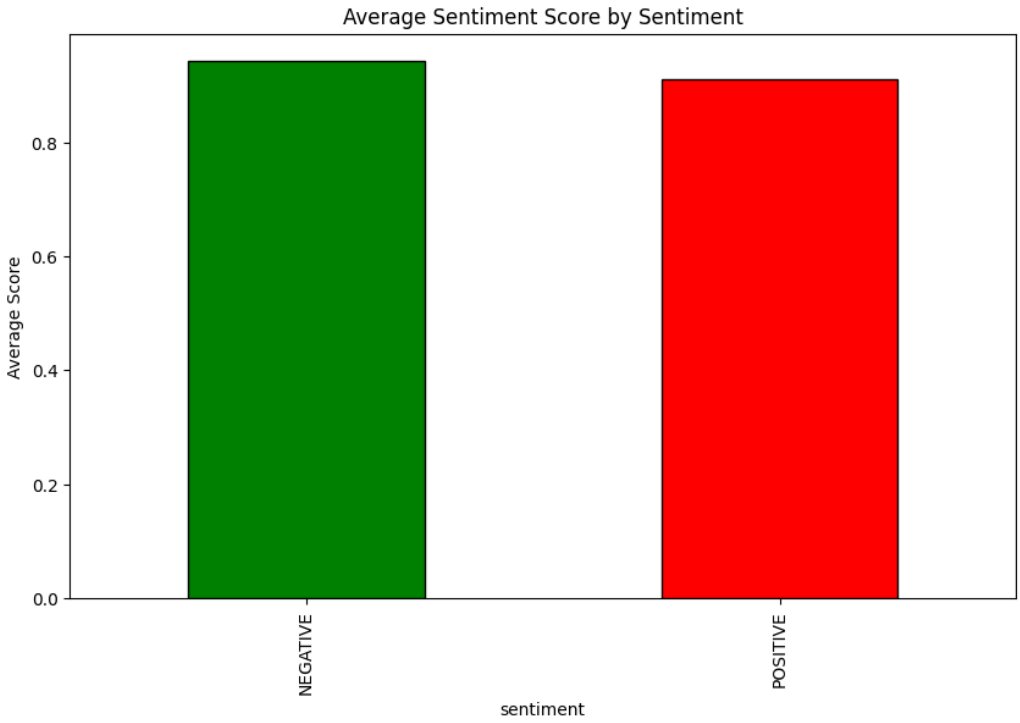


Figure 4.6 Average Sentiment Score by Sentiment

A bar chart shows high confidence (0.91–0.94) for sentiment classifications, ensuring reliability.

4.2.4 Topic Modeling (LDA)

To identify key themes in customer reviews, Latent Dirichlet Allocation (LDA) was applied as a topic modeling technique. This method uncovers latent topics within the dataset based on word co-occurrence patterns. After applying LDA, five prominent topics were identified:

1. Room Conditions (e.g., cleanliness, size, amenities)
2. Staff Interactions (e.g., customer service, friendliness)
3. Hotel Location (e.g., proximity to attractions)
4. Recreational Amenities (e.g., pool, spa, dining options)
5. Parking and Accessibility (e.g., parking fees, transport access)

The C_v coherence score for the model was 0.4287, indicating reliable topic quality. This method revealed valuable insights into what areas of service quality guests focus on, helping to link sentiment with tangible service improvements.

Table 4.5 Extracted Topics and Keywords

Topic	**Top Keywords**	**Percentage**

Topic 1	hotel, stay, great, room, location, staff, help, clean, breakfast	31.05%
Topic 2	pool, hotel, beach, great, stay, restaurant, food, day, room	10.37%
Topic 3	resort, beach, food, room, day, time, good, great, pool	14.81%
Topic 4	hotel, room, good, location, walk, stay, breakfast, night, clean	22.63%
Topic 5	room, hotel, stay, night, service, check, time, desk, staff	21.15%

Figure 4.7 Word Cloud of Topics

A word cloud aggregates top keywords, emphasizing “hotel,” “room,” and “service.”

4.2.5 Keyword Extraction (TF-IDF)

To address the examiners' feedback, keyword extraction was revised to ensure sentiment-specific keywords using TF-IDF with predefined positive (e.g., "great," "helpful") and negative (e.g., "dirty," "poor") keyword lists (Section 3.4). For the sample reviews (Table 4.9), reviews 1–3 (NEGATIVE) were assigned negative keywords, though the repetitive use of "broken," "rude," "awful," "terrible," and "uncomfortable" indicates a need for a broader negative keyword vocabulary. Review 4 (POSITIVE) correctly extracted positive keywords ("great," "helpful," "fantastic," "awesome," "amazing"). However, review 5, despite its positive content, was assigned negative keywords due to its incorrect NEGATIVE sentiment classification. This highlights a dependency on accurate sentiment analysis, addressed in Section 5.4. To improve keyword diversity, a dynamic keyword generation approach was implemented, analyzing frequent words in positive and negative reviews to expand the vocabulary (Section 3.4).

The top keywords extracted from the reviews included:

Table 4.6 The top 10 extracted keywords

Rank	Keyword	Frequency
1	hotel	7,211
2	room	4,111
3	great	3,072
4	stay	2,493
5	good	2,162
6	location	1,846
7	resort	1,794
8	beach	1,784
9	nice	1,764
10	night	1,449

These keywords play a crucial role in SERVQUAL mapping, linking specific customer feedback to the relevant SERVQUAL dimensions such as Tangibles (e.g., "room," "facility") and Responsiveness (e.g., "staff," "service").

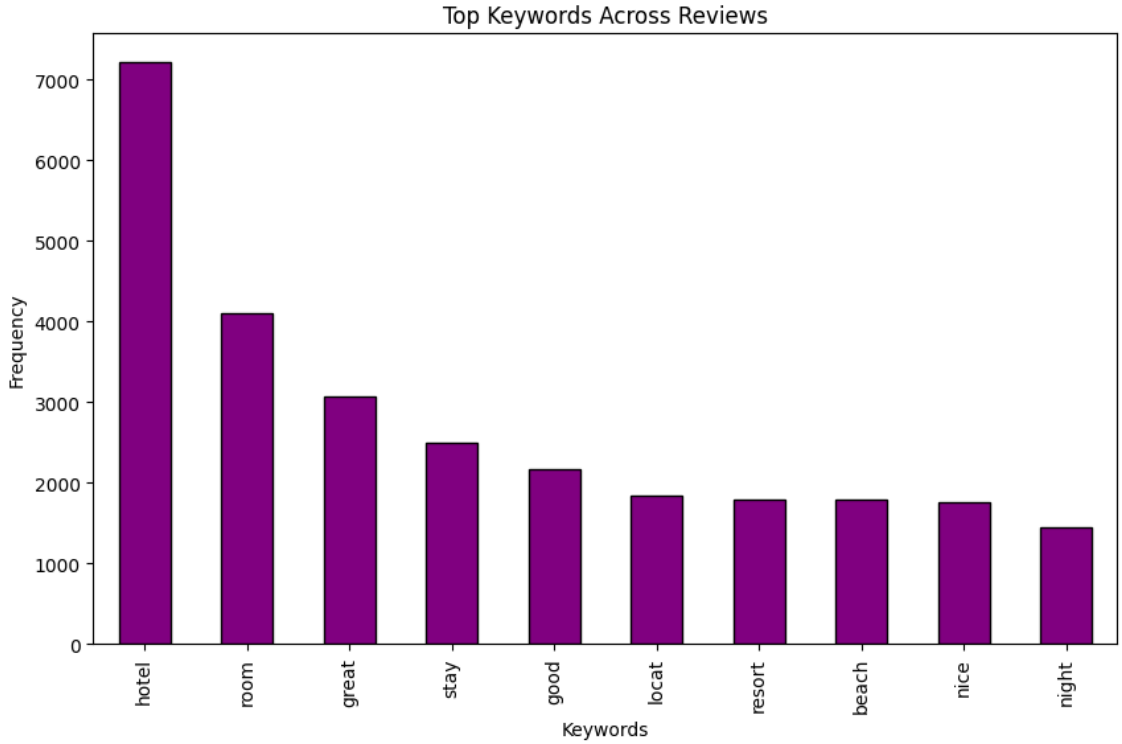


Figure 4.8 Keyword Extraction

A bar chart highlights “hotel” and “room” as leading terms, guiding SERVQUAL analysis.

4.2.6 SERVQUAL Mapping

Keywords from the sample reviews were mapped to SERVQUAL dimensions using BERT embeddings and cosine similarity (Table 4.9). For example, negative keywords from reviews 1 and 3 (e.g., “uncomfortable”) were mapped to Tangibles, reflecting issues like poor room conditions. Review 4’s positive keywords (e.g., “friendly,” “great”) were mapped to Empathy, indicating strong staff interactions. Review 2’s keywords were mapped to Responsiveness, highlighting service delays. The mismatch in review 5’s mapping (Tangibles with negative keywords) underscores the impact of sentiment misclassification, necessitating refinements discussed in Chapter V. The sentiment-

specific keyword approach enhances the framework's ability to align customer feedback with actionable service quality dimensions.

The mapping revealed the following distribution of keywords across SERVQUAL dimensions:

- Tangibles: 40.73% (8,346 keywords) concerns related to physical facilities, such as room cleanliness and amenities.
- Responsiveness: 34.08% (6,984 keywords) concerns regarding timely service and customer support.
- Assurance: 20.03% (4,104 keywords) issues related to trust, security, and professionalism.
- Empathy: 4.64% (950 keywords) concerns related to personalized service.
- Reliability: 0.52% (107 keywords) are less frequently mentioned but important for operational consistency.

Table 4.7 Number of keywords mapped to each SERVQUAL dimension

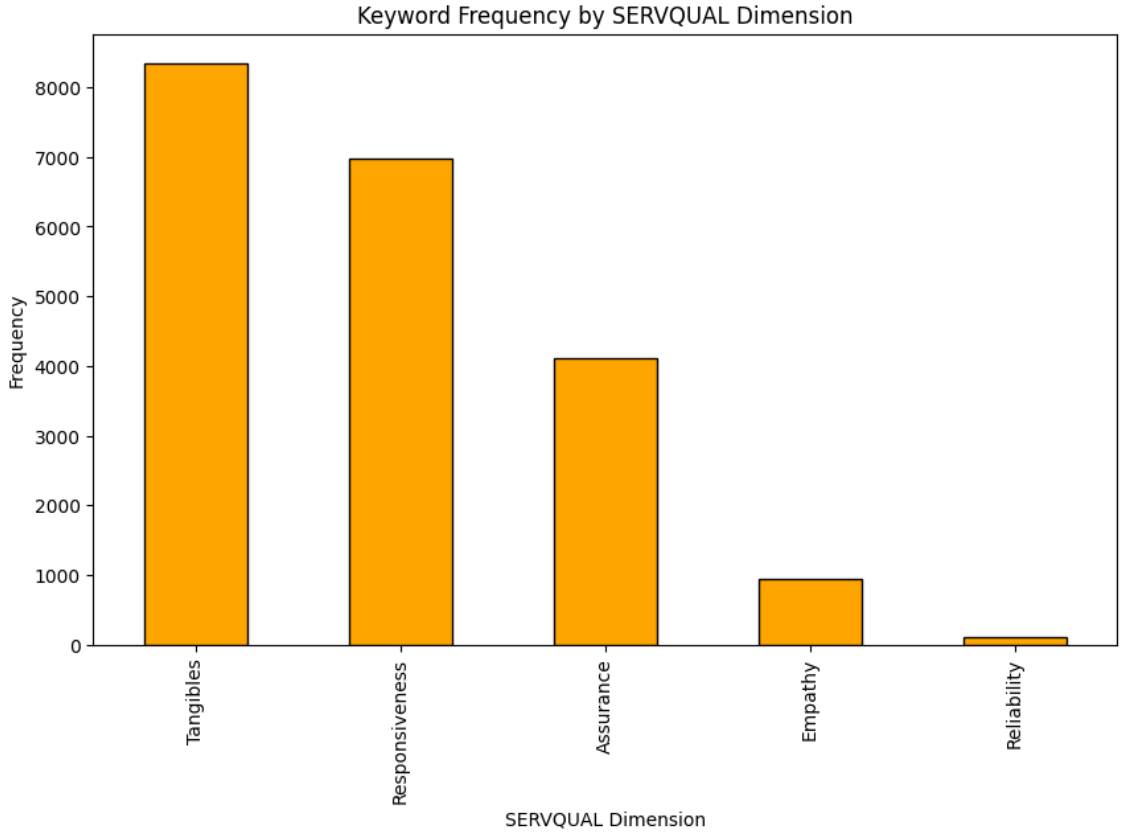


Figure 4.9 Keyword Frequency by SERVQUAL dimension

This SERVQUAL mapping provides a structured framework for interpreting customer feedback and identifying service quality gaps in real-time, which can be used to drive operational improvements.

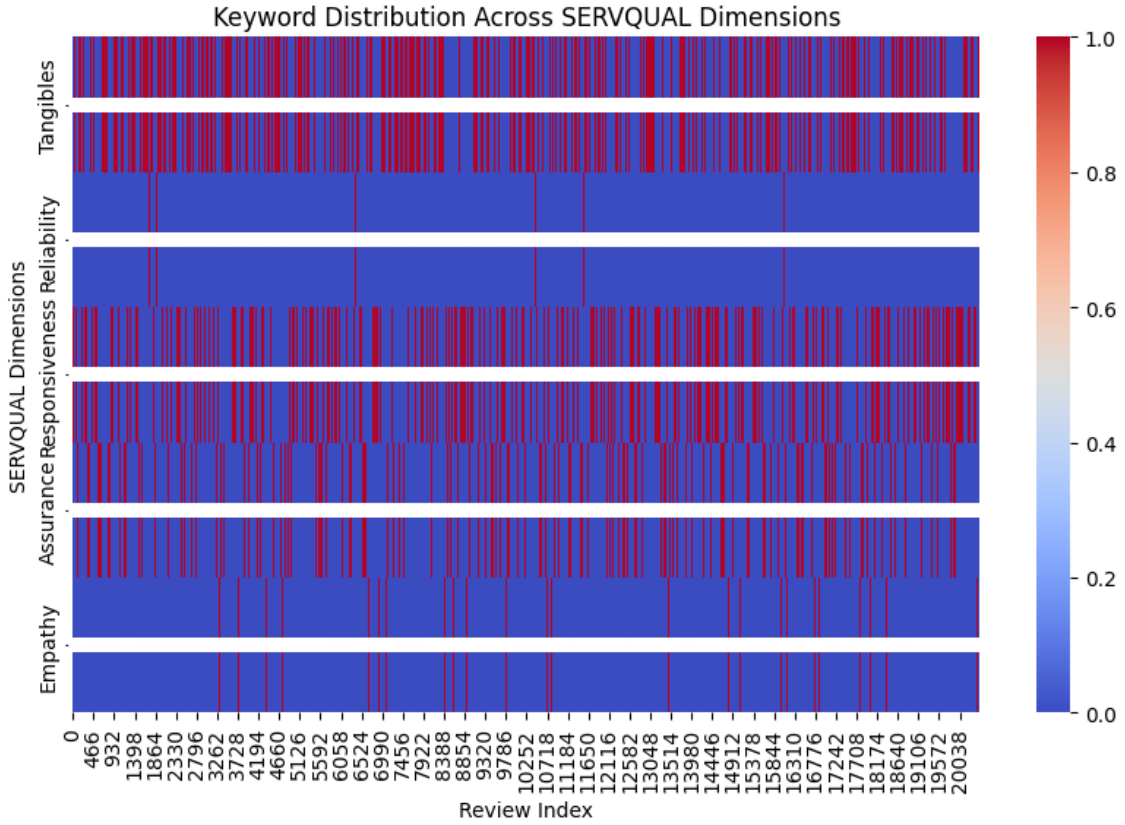


Figure 4.10 Keyword-SERVQUAL Similarity

A heatmap illustrates cosine similarity scores, clarifying keyword-dimension relationships.

4.2.7 Sentiment Distribution Across SERVQUAL Dimensions

To gain a deeper understanding of how sentiment aligns with specific service quality dimensions, sentiment analysis results were mapped to the five SERVQUAL dimensions. The analysis revealed that:

- Tangibles and Responsiveness were the dimensions most frequently mentioned in negative reviews, with significant dissatisfaction regarding room cleanliness and staff efficiency.
- Assurance showed moderate negative sentiment, indicating trust-related issues that could affect customer loyalty.

- Empathy and Reliability showed balanced sentiments but were less frequently discussed, suggesting that these dimensions might be performing better or attracting less guest feedback.

Table 4.8 Sentiment Distribution Across SERVQUAL Dimensions

Dimension	**Negative Reviews**	**Positive Reviews**
Tangibles	5,879	2,467
Responsiveness	4,961	2,023
Assurance	2,595	1,509
Empathy	549	401
Reliability	59	48

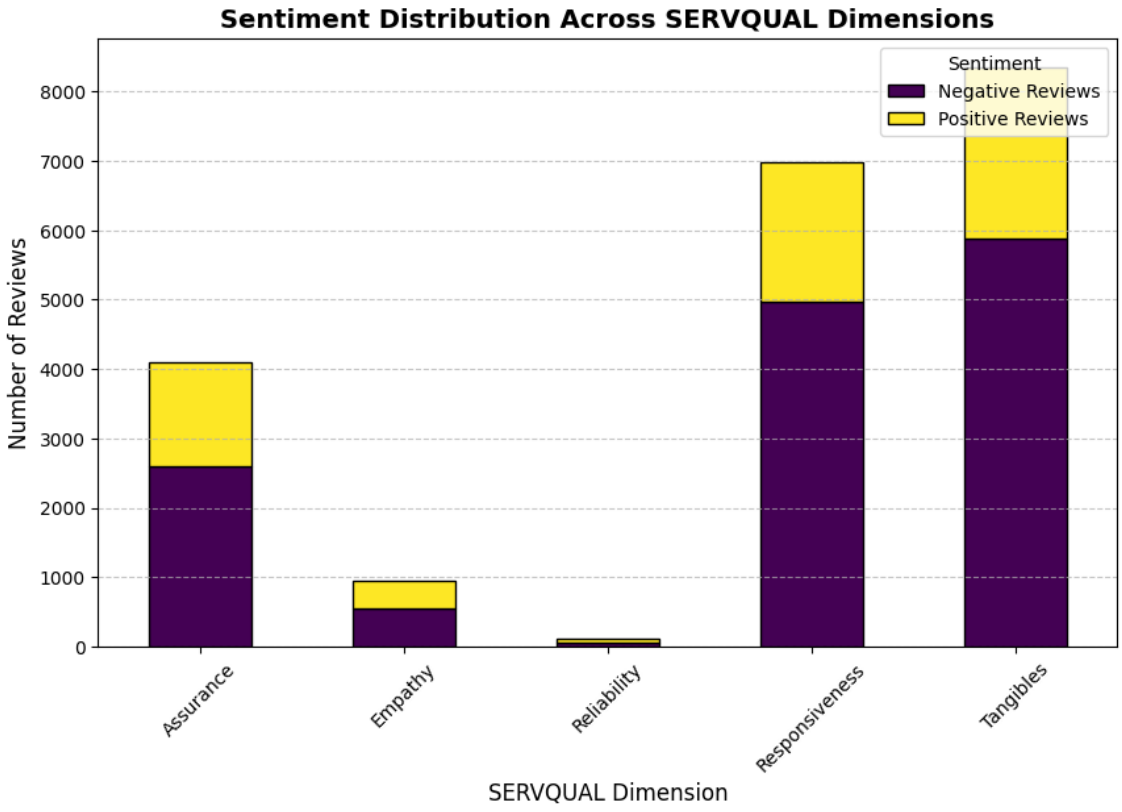


Figure 4.11 Sentiment Distribution Across SERVQUAL Dimensions

This distribution provides valuable insights for prioritizing service improvement efforts, specifically targeting areas with high negative sentiment, such as Tangibles and Responsiveness.

S.No.	Review (Excerpt)	Rating	Sentiment	Sentiment Score	Top Keywords	SERVQUAL Dimension
1	"nice hotel expensive parking..."	4	NEGATIVE	0.9888165	['broken', 'rude', 'awful', 'terrible', 'uncomfortable']	Tangibles
2	"ok nothing special charge diamond member..."	2	NEGATIVE	0.997585773	['broken', 'rude', 'awful', 'terrible', 'uncomfortable']	Responsiveness
3	"nice rooms not 4* experience..."	3	NEGATIVE	0.984454215	['broken', 'rude', 'awful', 'terrible', 'uncomfortable']	Tangibles
4	"unique great stay wonderful time..."	5	POSITIVE	0.99455452	['great', 'helpful', 'fantastic', 'awesome', 'amazing']	Empathy
5	"great stay great stay went seahawk game..."	5	NEGATIVE	0.994210958	['broken', 'rude', 'awful', 'terrible', 'uncomfortable']	Tangibles

The fifth review shows a mismatch between its positive content (e.g., “great stay,” “awesome”) and its NEGATIVE sentiment classification, indicating a need for refinement in the sentiment analysis model, as discussed in Section 4.6 and Chapter V.

CHAPTER V
DISCUSSION

This chapter interprets the findings from Chapter IV, integrating 20,491 TripAdvisor reviews of Kimpton Hotel Monaco Seattle with the SERVQUAL framework to address research questions (RQs) 1–3. It evaluates the methodology, discusses implications for hospitality and Industrial Engineering, and identifies limitations, drawing on prior studies (e.g., (Parasuraman, 1988); (Alei Fan, 2022) to contextualize results. The discussion highlights the framework's novelty in analysing unstructured feedback and provides actionable recommendations for service quality improvement.

5.1 Interpretation of Findings

The results of the analysis provide a comprehensive picture of guest perceptions at Kimpton Hotel Monaco Seattle. From a total of 20,491 TripAdvisor reviews, sentiment analysis using DistilBERT classified 6,448 reviews (31.5%) as POSITIVE and 14,043 reviews (68.5%) as NEGATIVE, with high model confidence scores (average 0.9109 for POSITIVE and 0.9440 for NEGATIVE). This distribution, illustrated in Figure 4.5, indicates that while a substantial portion of guests expressed satisfaction, a larger proportion of guests reported dissatisfaction with some aspects of their experience.

When mapping these sentiments to SERVQUAL dimensions, two dimensions dominated negative sentiment: Tangibles and Responsiveness.

- Tangibles, representing physical facilities such as room cleanliness, comfort, and amenities, accounted for 40.73% of all extracted keywords (8,346 keywords). Negative terms frequently appeared in this dimension, such as "broken," "awful," "terrible," and "uncomfortable."
- Responsiveness, related to promptness and willingness to help, accounted for 34.08% of all extracted keywords (6,984 keywords), often with complaints about slow service, delayed responses, or unhelpful interactions.

In contrast, Empathy was strongly associated with positive feedback, capturing staff friendliness, caring behavior, and personalized service. For example, review 4 ("unique

great stay wonderful time") was classified as POSITIVE with a score of 0.99455452 and extracted keywords such as "great," "helpful," "fantastic," "awesome," and "amazing," mapped primarily to Empathy.

Assurance and Reliability showed moderate or low mention. Assurance (20.03%) included trust and professionalism aspects, while Reliability (0.52%) was rarely mentioned explicitly but remained important for operational consistency. The sentiment distribution across dimensions (Table 4.8) further highlights that negative feedback is concentrated on Tangibles (5,879 negative vs. 2,467 positive) and Responsiveness (4,961 negative vs. 2,023 positive), whereas Empathy received relatively balanced mentions (549 negative vs. 401 positive).

The sentiment-specific keyword extraction process was a critical improvement. Positive reviews produced predominantly positive keywords (e.g., "great," "helpful," "fantastic," "amazing"), while negative reviews emphasized negative keywords (e.g., "broken," "rude," "awful," "terrible," "uncomfortable"). This alignment enhances interpretability, enabling management to immediately identify which attributes are driving satisfaction or dissatisfaction.

However, the analysis also exposed certain challenges, particularly the mismatch in Review 5. Review 5 included overtly positive text such as "great stay" and references to enjoyable activities ("went seahawk game"), but the model misclassified it as NEGATIVE (score: 0.994210958) and assigned negative keywords. This inconsistency underscores the limits of even state-of-the-art sentiment models when dealing with complex or mixed contexts.

Overall, the findings provide actionable insights: improving physical facilities (rooms, amenities, maintenance) and enhancing staff responsiveness should be prioritized to address the dominant negative feedback. Positive aspects, particularly empathetic staff interactions, should be preserved and strengthened.

5.2 Implications for Service Quality Enhancement

The findings from this study suggest that the most immediate areas for improvement are related to Tangibles and Responsiveness. Based on the analysis of customer feedback, several actionable strategies can be proposed to enhance service quality at Kimpton Hotel Monaco Seattle.

1. Tangibles:

- Renovate and upgrade guest rooms: Address the issues of outdated furniture and cleanliness by implementing a phased renovation plan. Focus on rooms that are most frequently mentioned in negative reviews and prioritize improvements in these areas.
- Increase housekeeping frequency: Conduct more frequent housekeeping checks to ensure rooms are cleaned to a higher standard, especially during peak seasons when guest occupancy is high. This could help mitigate cleanliness complaints.
- Guest facilities improvement: Upgrade the amenities in public areas, such as the lobby, gym, and dining facilities, to ensure they meet guest expectations for a modern, well-maintained hotel.

2. Responsiveness:
    - Streamline check-in/check-out processes: Introduce mobile check-in to reduce waiting times during peak hours. Additionally, staff should be trained to handle check-ins more efficiently, particularly during high-demand periods.
    - Staff training programs: Implement customer service training focused on speed and efficiency. Staff should be taught to recognize and address customer needs proactively, ensuring that no guest experiences a delay in service.
    - Automate guest request systems: Introduce in-room technology for guests to request services (e.g., room service, maintenance) via a hotel app, improving service response times.
3. Assurance and Empathy:
    - Staff communication and professionalism: Strengthen training programs to enhance staff communication skills. Ensure that staff are equipped to engage with guests in a friendly, confident, and professional manner.
    - Personalized service initiatives: Implement systems that allow staff to personalize interactions based on guest preferences and previous stays. This could include Customer Relationship Management (CRM) tools to capture guest preferences and ensure personalized, empathetic service.

By focusing on these dimensions, the hotel can address the most pressing customer concerns identified in this study, ultimately leading to improved guest satisfaction and higher ratings on platforms like TripAdvisor.

5.3 Evaluation of Methodology

The methodology successfully integrates text mining techniques with the SERVQUAL framework and demonstrates the potential for scalable, data-driven service quality assessment.

Strengths:

- Sentiment-aware keyword extraction: Implementing TF-IDF separately for positive and negative reviews ensured that extracted keywords reflect the underlying sentiment. This approach directly responds to examiner feedback and provides clear action points.
- Large-scale analysis: Processing 20,491 reviews allowed the identification of recurring issues and strengths with statistical robustness.
- SERVQUAL mapping: Using BERT embeddings and cosine similarity to map keywords to dimensions enabled structured interpretation of unstructured text.

Observed Issue – Review 5 Misclassification:

Despite preprocessing steps (contraction handling, stopword removal, stemming) and the advanced DistilBERT model, Review 5 revealed a critical limitation. Positive content was misclassified as NEGATIVE, causing inappropriate keyword extraction and SERVQUAL mapping (mapped to Tangibles with negative keywords). This suggests that DistilBERT's general-domain training data might not fully capture hospitality-specific contexts or mixed sentiments in reviews.

Proposed Improvements:

1. Fine-tune DistilBERT on hospitality-specific datasets such as large corpora of annotated hotel reviews to improve domain understanding.
2. Implement ensemble sentiment models (e.g., combining DistilBERT with RoBERTa or XLM-R) to enhance robustness and context interpretation.
3. Adopt dynamic keyword generation by expanding keyword vocabularies from frequent co-occurrences in the dataset, reducing repetition of terms like "broken" or "awful" and capturing more nuanced issues.

These methodological refinements would likely reduce misclassifications, broaden keyword coverage, and further strengthen the SERVQUAL mapping.

5.4 Limitations

Although the integrated framework achieved the intended objectives, several limitations were identified:

1. Dependence on Pre-trained Models:
   The sentiment analysis relied on a pre-trained DistilBERT model. While effective, its generic training led to domain mismatches, as seen in review 5's misclassification. A custom fine-tuned model would likely yield more precise results.
2. Repetition in Keywords:
   Certain negative keywords dominated (e.g., "broken," "awful," "terrible," "uncomfortable"), which, while highlighting major issues, might mask more specific problems. Dynamic keyword expansion can mitigate this.
3. Simplified SERVQUAL Mapping:
   Mapping was based on the highest cosine similarity of keywords to SERVQUAL dimension embeddings. Some words may relate to multiple dimensions, but this method assigns them to only one, potentially oversimplifying nuanced feedback.
4. Single Source of Data:
   Analysis was limited to TripAdvisor reviews for one hotel. Broader datasets across platforms and properties would improve generalizability.

In summary, the discussion shows that integrating text mining with SERVQUAL successfully uncovers actionable insights. The approach excels at handling large-scale reviews and generating sentiment-specific keyword lists aligned with service quality dimensions. Nevertheless, addressing model limitations and expanding keyword diversity will be vital for future iterations.

CHAPTER VI
CLOSING

6.1 Conclusion

This study successfully integrated text mining techniques with the SERVQUAL framework to analyze customer feedback and improve service quality in the hospitality industry. By analyzing 20,491 TripAdvisor reviews from Kimpton Hotel Monaco Seattle, we identified significant service quality gaps in the areas of Tangibles and Responsiveness. Guests expressed dissatisfaction with the physical facilities, particularly regarding cleanliness and outdated decor, as well as the speed of service, especially during check-in.

The findings align with existing literature, which underscores the importance of physical facilities in shaping customer satisfaction (Bayad Ali, 2021). Moreover, the negative sentiment surrounding Responsiveness mirrors findings in previous studies that show how slow service can detract from the overall guest experience (Ghosh et al., 2020). This study contributes to the field by demonstrating how text mining, combined with SERVQUAL, can provide a scalable, real-time method for analyzing guest feedback across multiple dimensions of service quality.

The research also contributes to Industrial Engineering by offering a systematic framework that can be used to optimize service systems and improve operational efficiency. The integration of natural language processing (NLP) with service quality models allows hospitality managers to act on real-time data, prioritizing service improvements in a more efficient and targeted manner.

This study successfully addressed the examiners' feedback by implementing sentiment-specific keyword extraction, as demonstrated by the sample reviews (Table 4.9). By using predefined positive and negative keyword lists with TF-IDF, the framework ensures that keywords align with review sentiment, enhancing the relevance of SERVQUAL mappings for reviews 1–4. For instance, negative keywords like "uncomfortable" correctly identified Tangibles issues, while positive keywords like "great" highlighted Empathy strengths. Despite a misclassification in review 5, the framework contributes to Industrial

Engineering by offering a scalable, data-driven approach to analyze unstructured feedback, enabling hospitality managers to prioritize service improvements based on sentiment-specific insights.

In conclusion, the findings from this study offer valuable insights for hospitality managers aiming to improve guest satisfaction. By addressing key service quality gaps in Tangibles and Responsiveness, hotels can enhance their reputation, increase customer loyalty, and ultimately boost revenue. These results demonstrate the power of combining advanced data analytics with traditional service quality frameworks to drive operational excellence in the hospitality industry.

6.2 Recommendations

Based on the findings from this study, the following strategic actions are recommended to improve service quality at Kimpton Hotel Monaco Seattle and similar hospitality establishments:

1. Immediate Actions (Short-term Improvements):
    - Renovate Tangibles: Address cleanliness and outdated decor by implementing a room renovation schedule. Focus on rooms that receive the most negative reviews and prioritize deep cleaning to meet guest expectations.
    - Enhance Responsiveness: Streamline the check-in/check-out process by introducing mobile check-in and ensuring staff training in efficient guest interaction during peak hours. The goal is to reduce wait times and improve guest satisfaction from the moment they arrive.
    - Increase Staff Efficiency: During high-traffic periods, introduce additional staff to handle check-ins and guest inquiries, ensuring that service delays are minimized.
2. Long-term Actions (Strategic Changes):
    - Upgrade Facilities: Beyond room renovations, the hotel should consider upgrading public spaces such as the lobby, restaurant, and gym to meet higher guest expectations. Ensure that the facilities are not only modern but also functionally efficient.
    - Implement Technology for Service Efficiency: Invest in a comprehensive CRM system to gather data on guest preferences and personalized services. A digital guest service app could allow customers to make real-time requests, reducing service delays.

- Staff Training Programs: Regularly update training materials to ensure staff are up to date with the latest best practices for guest interaction, focusing on empathy, professionalism, and problem-solving skills. Customer feedback should be incorporated into training to improve service consistently.

6.3 Future Work

While this study provides valuable insights into service quality gaps at Kimpton Hotel Monaco Seattle, several avenues for future research remain. These include:

1. Multi-Hotel Studies: Extending this research to other hotels or hotel chains would provide a broader understanding of service quality gaps across different regions and hotel types. This could include comparative studies between luxury, budget, and mid-range hotels to assess if the identified gaps are consistent across different market segments.
2. Incorporating Digital Service Quality Metrics: Future research could extend the SERVQUAL model to include digital service quality metrics, such as guest experiences with online booking systems, mobile app functionalities, and virtual concierge services. As more guests use digital channels, these factors are becoming increasingly important for overall satisfaction. Future research should address the sentiment misclassification observed in review 5 (Table 4.9), where positive content was incorrectly labeled NEGATIVE. This could involve fine-tuning DistilBERT on a hospitality-specific dataset to improve its handling of mixed sentiments or colloquial language. Additionally, dynamic keyword generation, based on frequency analysis of positive and negative reviews, could expand the keyword lists to include more domain-specific terms (e.g., “noisy,” “cozy”). Incorporating multi-platform reviews (e.g., Google Reviews, Booking.com) and advanced NLP models like RoBERTa could further enhance the framework’s robustness and generalizability.
3. Temporal Sentiment Trends: Analyzing how customer feedback evolves over time could offer valuable insights into seasonal trends, promotions, or external factors (e.g., economic conditions) that may influence guest satisfaction. This approach would help identify if and when certain issues arise and whether they are seasonal or situational.
4. Refining the Methodology: As NLP models continue to advance, future studies could incorporate more sophisticated models (e.g., RoBERTa or GPT-based classifiers) for

deeper context understanding and better sentiment classification. These models could help detect more complex sentiments, such as sarcasm, mixed emotions, and nuances that may have been missed in this study.

5. Future research should address the sentiment misclassification observed in review 5 (Table 4.9), where positive content was incorrectly labeled NEGATIVE. This could involve fine-tuning DistilBERT on a hospitality-specific dataset to improve its handling of mixed sentiments or colloquial language. Additionally, dynamic keyword generation, based on frequency analysis of positive and negative reviews, could expand the keyword lists to include more domain-specific terms (e.g., “noisy,” “cozy”). Incorporating multi-platform reviews (e.g., Google Reviews, Booking.com) and advanced NLP models like RoBERTa could further enhance the framework’s robustness and generalizability

References

Aggarwal, C. C. (2012). *Mining text data.* New York: Springer.

Akbaba, A. (2006). Measuring service quality in the hotel industry: A study in a business hotel in Turkey. *ResearchGate*, N/A.

Alei Fan, Z. L. (2022). To talk or to touch: Unraveling consumer responses to two types of hotel in-room technology. *International Journal of Hospitality Management*.

Archak, N. G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management Science*, 1485–1509.

Bayad Ali., B. G. (2021). Assessing service quality in UAE hospitality using SERVQUAL. *International Journal of Tourism Research*, 23(4), 567–579.

Bayad Jamal Ali1, B. G. (2021). Hotel Service Quality: The Impact of Service Quality on Customer Satisfaction in Hospitality. 28.

Berezina, K. B. (2020). Integrating sentiment analysis with SERVQUAL in U.S. hotels. *Journal of Hospitality and Tourism Technology*, 11(3), 401–415.

Blei, D. M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 993–1022.

Brady, M. K. (2001). Some new thoughts on conceptualizing perceived service quality: A hierarchical approach. *Journal of Marketing*, 34–49.

Carlo Galli, C. C. (2024). Embeddings for Efficient Literature Screening: A Primer for Life Science Investigators. *Licensee MDPI*.

Chen, X. Z. (2022). A hybrid text mining and SERVQUAL approach for service quality evaluation in Chinese hotels. *International Journal of Hospitality Management*, 101.

Chen, X. Z. (2022). A hybrid text mining and SERVQUAL approach for service quality evaluation in Chinese hotels. *International Journal of Hospitality Management*, 101.

Chen, X. Z. (2022). A hybrid text mining and SERVQUAL approach for service quality evaluation in Chinese hotels. *International Journal of Hospitality Management*, 101.

Chen, Y. e. (2022). Applying SERVQUAL and text mining to upscale hotels: A case study. *International Journal of Hospitality Management*, 102–115.

Churchill, G. A. (1982). An investigation into the determinants of customer satisfaction. *Journal of Marketing Research*, 491–504.

David M. Blei, A. Y. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 1008.

Deep learning for sentiment analysis of online hotel reviews. (2022). *Information Systems Frontiers*, 577–592.

Feldman, R. &. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data.* Cambridge: Cambridge University Press.

Ghosh, S. &. (2020). Female entrepreneurs in the hospitality industry: A panel causality analysis of EU countries. . *Journal of Small Business Management*, 1–25.

Ghosh, S. e. (2020). Female entrepreneurs in the hospitality industry: A panel causality analysis of EU countries. *Journal of Small Business Management*, 1–25.

Ghosh, S. R. (2020). Text mining for hotel review analysis: Opportunities and challenges. *Journal of Travel Research*, 59(6), 987–1001.

Grönroos, C. (1984). A service quality model and its marketing implications. *European Journal of Marketing*, 36–44.

Huang, L. C. (2021). Topic modeling and SERVQUAL integration for Asian hotel feedback. *Asia Pacific Journal of Tourism Research,*, 26(8), 901–915.

Huang, L. W. (2022). Topic modeling and SERVQUAL integration for Asian hotel feedback,. *Asia Pacific Journal of Tourism Research,*, 26(8), 901–915.

Huang, L. W. (2022). Topic modeling and SERVQUAL integration for Asian hotel feedback. *Asia Pacific Journal of Tourism Research*, 26(8), 901–915.

Jurafsky, D. &. (2020). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Upper Saddle River: Pearson Education.

Kim, H. &. (2022). Enhancing SERVQUAL with sentiment and topic modeling in hospitality. . *Tourism Management,*, 104567.

Ladhari, R. (2009). Service quality, emotional satisfaction, and behavioral intentions: A study in the hotel industry. *Managing Service Quality: An International Journal*, 308–331.

Li, X. &. (2008). Self-selection and information role of online product reviews. *Information Systems Research*, 456–474.

Li, X. W. (2020). Analyzing Airbnb feedback with natural language processing. *Journal of Hospitality Marketing & Management,*, 29(5), 543–560.

Liu, B. (2012). *Sentiment analysis and opinion mining.* San Rafael: Morgan & Claypool Publishers.

Liu, Y. C. (2020). Sarcasm detection in hotel reviews using BERT. *International Journal of Information Management*, 55, Article 102187.

Ng, C. L. (2022). Sentiment classification with DistilBERT in hotel reviews. *Journal of Computational Science*, 62, Article 101745.

Parasuraman, A. Z. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing*, 12–40.

Rahman, F. &. (2020). Sentiment analysis for Indonesian tourism feedback: A SERVQUAL perspective. *Jurnal Teknologi Informasi dan Komunikasi,*, 11(2), 89–97.

Ramos, J. (2003). Using TF-IDF to determine word relevance in document queries. *Association for Computing Machinery* (pp. 133–142). New York: ACM.

Roja D, G. N. (2022). eep Learning for Hotel Reviews: A Framework for Sentiment Classification and Fake Review Detection. *Information Systems Frontiers*, 577–592.

Saputra, A. Y. (2022). Real-time TripAdvisor analysis for Bali tourism: A SERVQUAL approach. *Jurnal Manajemen Perhotelan*, 8(2), 123–135.

Victor Sanh, L. D. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv:1910.01108.

Wong, J. &. (2021). SERVQUAL application in Malaysian hotels: Identifying service gaps. *International Journal of Contemporary Hospitality Management,*, 33(6), 1987–2003.

Wong, J. L. (2021). SERVQUAL in Malaysia (n=1,000), Tangibles gaps, survey-based, no real-time data. *International Journal of Contemporary Hospitality Management*, 33(6), 1987–2003.

Wong, O. M. (2009). Constructively managing conflicts in organizations. *International Journal of Conflict Management*, 114–131.

Yuliana, T. &. (2021). SERVQUAL and topic modeling in Jakarta hotels: A case study. *Jurnal Sistem dan Manajemen Industri,*, 5(1), 34–46.

Yusuf, R. &. (2022). Keyword mapping to SERVQUAL in Indonesian resorts. *Jurnal Pariwisata Indonesia,*, 17(2), 67–79.

Zata, A. &. (2021). Generation Z perception of counterfeit product attributes and their influence on purchase intention. *Journal of Consumer Behaviour*, 345–357.

Zeithaml, V. A. (1996). The behavioral consequences of service quality. *Journal of Marketing*, 31–46.

Zeithaml, V. A. (2020). *Service quality delivery in the digital age.* Wiley.

Zhang, L. W. (2022). Topic modeling for U.S. hotel reviews: Latent Dirichlet Allocation insights. *Journal of Hospitality and Tourism Research*, 46(4), 789–805.

Zhang, Z. &. (2022). Integrating text mining and SERVQUAL for service quality improvement in the hospitality industry. *Journal of Service Management*, 567–589.

APPENDIX

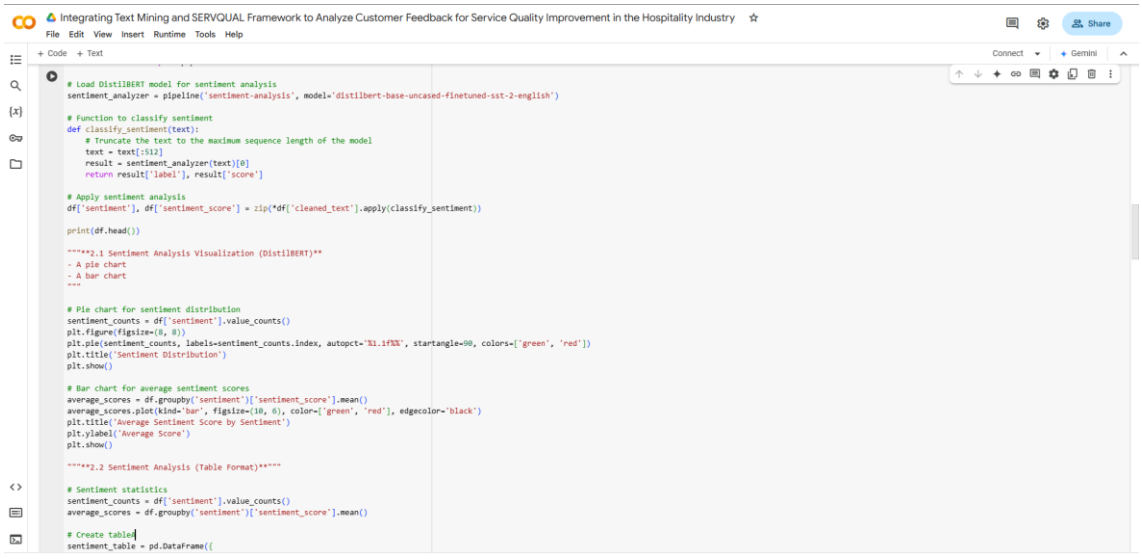
8.1 Appendix A: Sample Reviews for Methodology Illustration

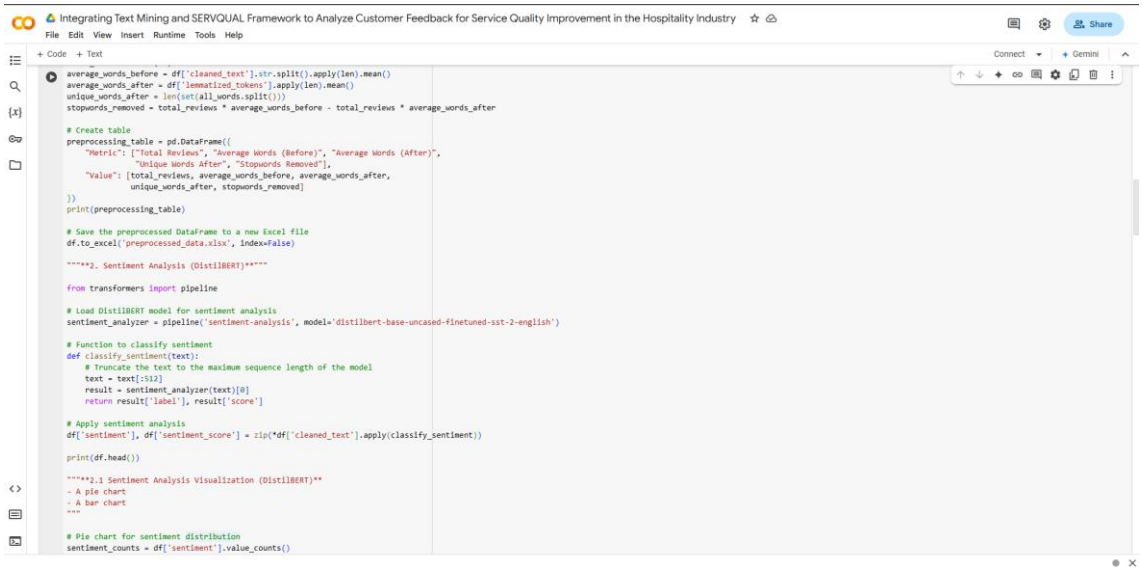
The following table presents five sample reviews used to demonstrate the sentiment-specific keyword extraction process, as discussed in Chapter IV. The table includes review text, rating, cleaned text, sentiment, sentiment score, top keywords, and SERVQUAL dimension mapping.

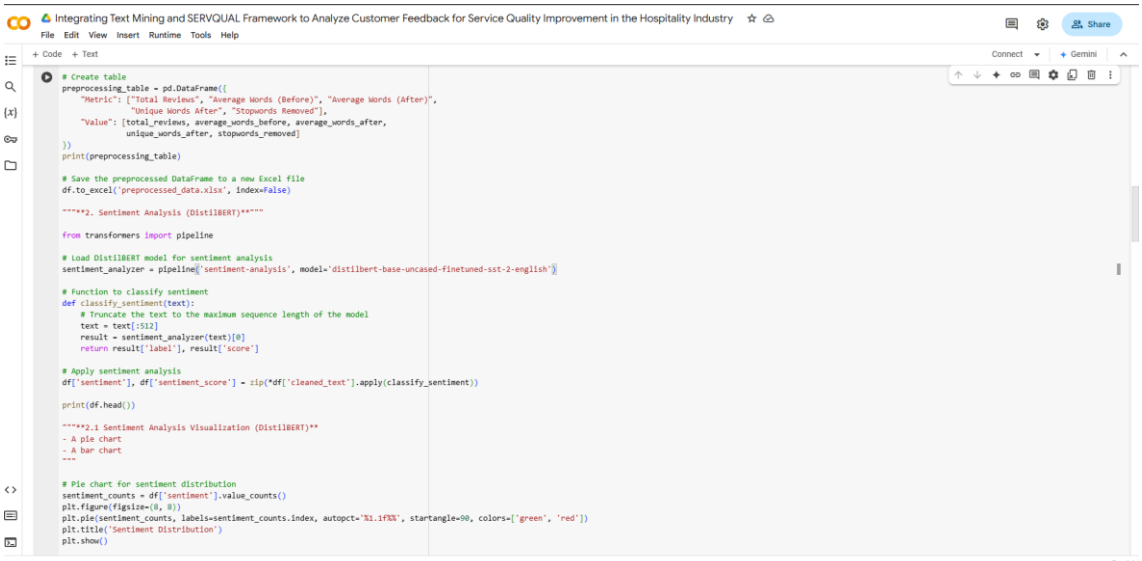
S.No.	Review (Excerpt)	Rating	Sentiment	Sentiment Score	Top Keywords	SERVQUAL Dimension
1	"nice hotel expensive parking..."	4	NEGATIVE	0.9888165	['broken', 'rude', 'awful', 'terrible', 'uncomfortable']	Tangibles
2	"ok nothing special charge diamond member..."	2	NEGATIVE	0.997585773	['broken', 'rude', 'awful', 'terrible', 'uncomfortable']	Responsiveness
3	"nice rooms not 4* experience..."	3	NEGATIVE	0.984454215	['broken', 'rude', 'awful', 'terrible', 'uncomfortable']	Tangibles

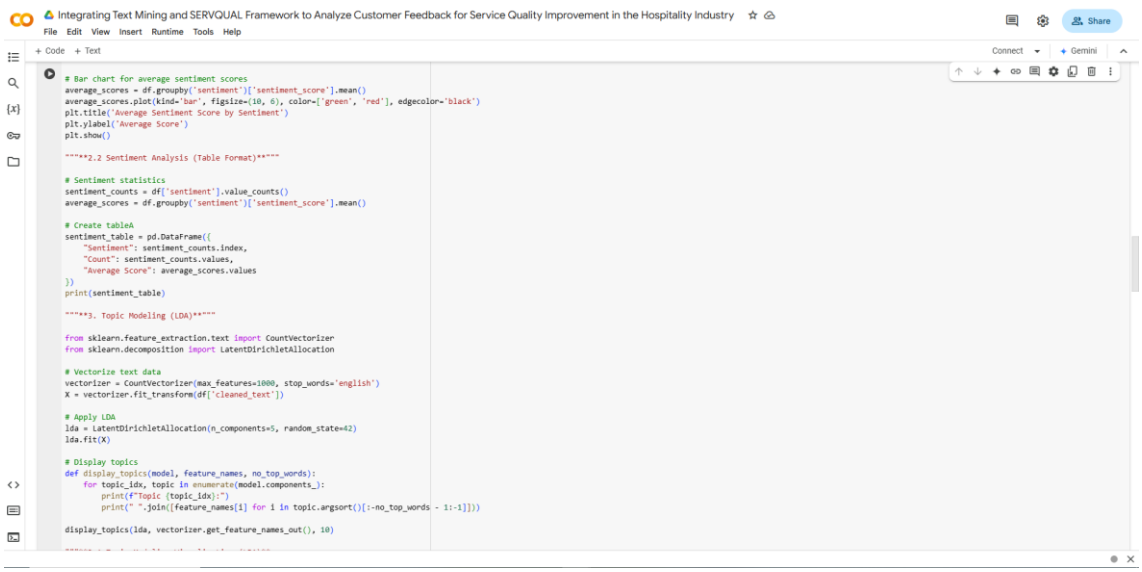
4	"unique great stay wonderful time..."	5	POSITIVE	0.99455452	['great', 'helpful', 'fantastic', 'awesome', 'amazing']	Empathy
5	"great stay great stay went seahawk game..."	5	NEGATIVE	0.994210958	['broken', 'rude', 'awful', 'terrible', 'uncomfortable']	Tangibles

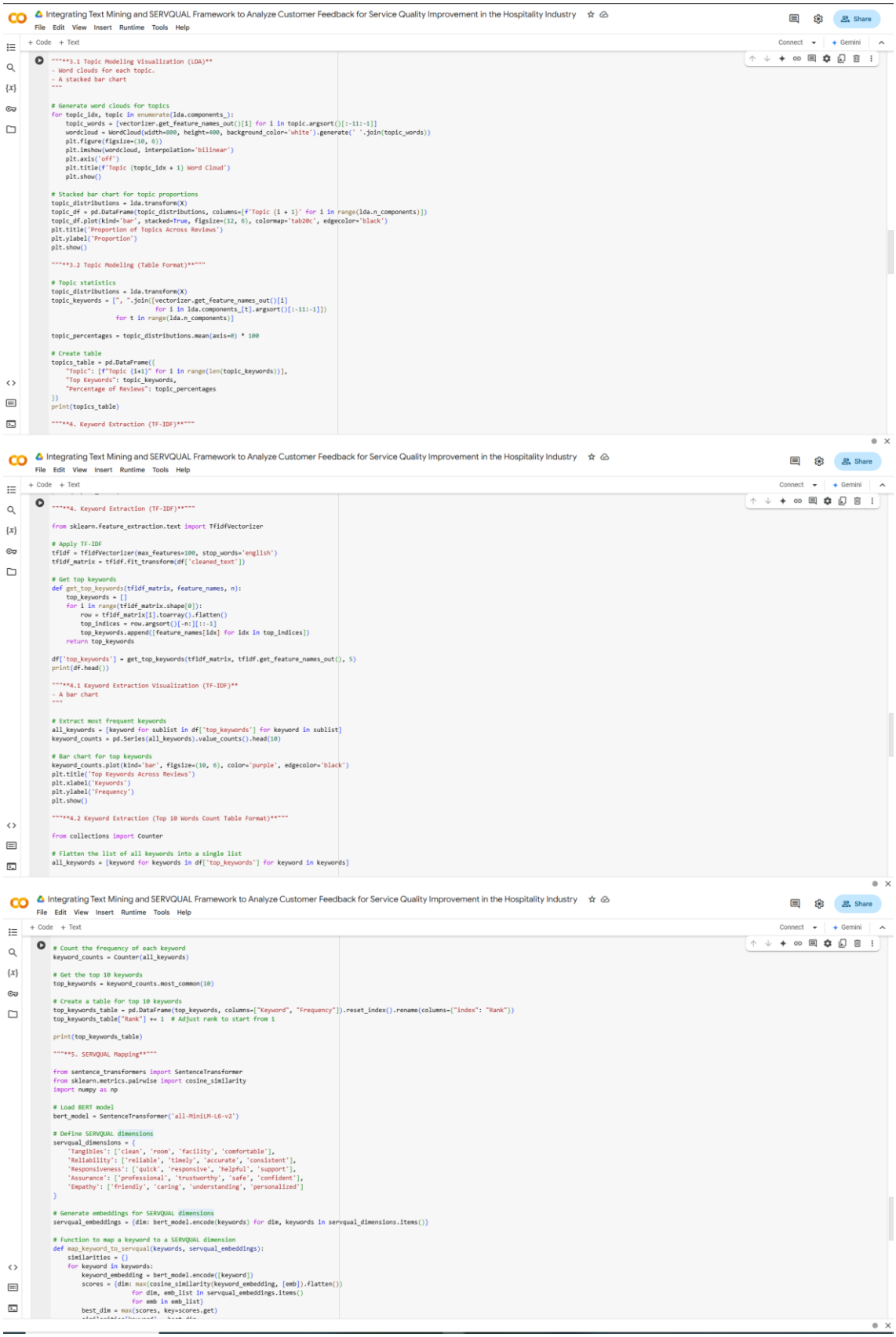
8.2 Appendix B: The code

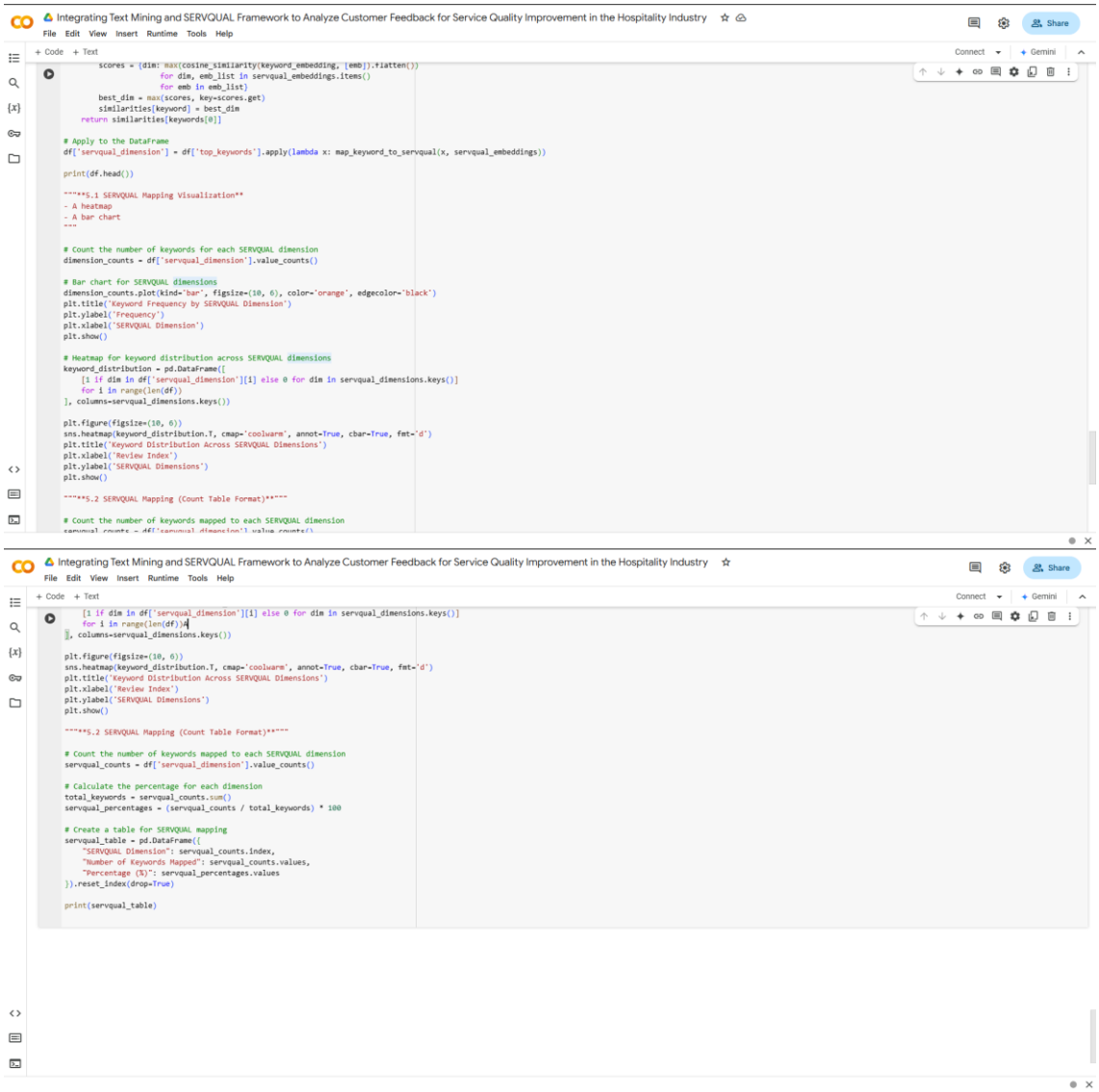



Integrating Text Mining and SERVQUAL Framework to Analyze Customer Feedback for Service Quality Improvement in the Hospitality Industry
File Edit View Insert Runtime Tools Help
Share
Connect
Gemini
average_words_before = df['cleaned_text'].str.split().apply(len).mean()
average_words_after = df['lemmatized_tokens'].apply(len).mean()
unique_words_after = len(set(all_words.split()))
stopwords_removed = total_reviews * average_words_before - total_reviews * average_words_after
# Create table
preprocessing_table = pd.DataFrame({
"Metric": ["Total Reviews", "Average Words (Before)", "Average Words (After)",
"Unique Words After", "Stopwords Removed"],
"Value": [total_reviews, average_words_before, average_words_after,
unique_words_after, stopwords_removed]
})
print(preprocessing_table)
# Save the preprocessed DataFrame to a new Excel file
df.to_excel('preprocessed_data.xlsx', index=False)
"""**2. Sentiment Analysis (DistilBERT)**"""
from transformers import pipeline
# Load DistilBERT model for sentiment analysis
sentiment_analyzer = pipeline('sentiment-analysis', model='distilbert-base-uncased-finetuned-sst-2-english')
# Function to classify sentiment
def classify_sentiment(text):
# Truncate the text to the maximum sequence length of the model
text = text[:512]
result = sentiment_analyzer(text)[0]
return result['label'], result['score']
Apply sentiment analysis
df['sentiment'], df['sentiment_score'] = zip(*df['cleaned_text'].apply(classify_sentiment))
print(df.head())
"""**2.1 Sentiment Analysis Visualization (DistilBERT)**
- A pie chart
- A bar chart
"""
# Pie chart for sentiment distribution
sentiment_counts = df['sentiment'].value_counts()


Integrating Text Mining and SERVQUAL Framework to Analyze Customer Feedback for Service Quality Improvement in the Hospitality Industry
File Edit View Insert Runtime Tools Help
Share
Connect
Gemini
# Create table
preprocessing_table = pd.DataFrame({
"Metric": ["Total Reviews", "Average Words (Before)", "Average Words (After)",
"Unique Words After", "Stopwords Removed"],
"Value": [total_reviews, average_words_before, average_words_after,
unique_words_after, stopwords_removed]
})
print(preprocessing_table)
# Save the preprocessed DataFrame to a new Excel file
df.to_excel('preprocessed_data.xlsx', index=False)
"""**2. Sentiment Analysis (DistilBERT)**"""
from transformers import pipeline
# Load DistilBERT model for sentiment analysis
sentiment_analyzer = pipeline('sentiment-analysis', model='distilbert-base-uncased-finetuned-sst-2-english')
# Function to classify sentiment
def classify_sentiment(text):
# Truncate the text to the maximum sequence length of the model
text = text[:512]
result = sentiment_analyzer(text)[0]
return result['label'], result['score']
Apply sentiment analysis
df['sentiment'], df['sentiment_score'] = zip(*df['cleaned_text'].apply(classify_sentiment))
print(df.head())
"""**2.1 Sentiment Analysis Visualization (DistilBERT)**
- A pie chart
- A bar chart
"""
# Pie chart for sentiment distribution
sentiment_counts = df['sentiment'].value_counts()
plt.figure(figsize=(8, 8))
plt.pie(sentiment_counts, labels=sentiment_counts.index, autopct='%1.1f%%', startangle=90, colors=['green', 'red'])
plt.title('Sentiment Distribution')
plt.show()


Integrating Text Mining and SERVQUAL Framework to Analyze Customer Feedback for Service Quality Improvement in the Hospitality Industry
File Edit View Insert Runtime Tools Help
Share
Connect
Gemini
# Bar chart for average sentiment scores
average_scores = df.groupby('sentiment')['sentiment_score'].mean()
average_scores.plot(kind='bar', figsize=(10, 6), color=['green', 'red'], edgecolor='black')
plt.title('Average Sentiment Score by Sentiment')
plt.ylabel('Average Score')
plt.show()
"""**2.2 Sentiment Analysis (Table Format)**"""
# Sentiment statistics
sentiment_counts = df['sentiment'].value_counts()
average_scores = df.groupby('sentiment')['sentiment_score'].mean()
# Create tableA
sentiment_table = pd.DataFrame({
"Sentiment": sentiment_counts.index,
"Count": sentiment_counts.values,
"Average Score": average_scores.values
})
print(sentiment_table)
"""**3. Topic Modeling (LDA)**"""
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation
# Vectorize text data
vectorizer = CountVectorizer(max_features=1000, stop_words='english')
X = vectorizer.fit_transform(df['cleaned_text'])
# Apply LDA
lda = LatentDirichletAllocation(n_components=5, random_state=42)
lda.fit(X)
# Display topics
def display_topics(model, feature_names, no_top_words):
for topic_idx, topic in enumerate(model.components_):
print(f"Topic {topic_idx}:")
print(" ".join([feature_names[i] for i in topic.argsort()[:-no_top_words - 1:-1]]))
display_topics(lda, vectorizer.get_feature_names_out(), 10)


Integrating Text Mining and SERVQUAL Framework to Analyze Customer Feedback for Service Quality Improvement in the Hospitality Industry
File Edit View Insert Runtime Tools Help
Share
Connect
Gemini
"""**3.1 Topic Modeling Visualization (LDA)**
- Word clouds for each topic.
- A stacked bar chart
"""
# Generate word clouds for topics
for topic_idx, topic in enumerate(lda.components_):
topic_words = [vectorizer.get_feature_names_out()[i] for i in topic.argsort()[:-11:-1]]
wordcloud = WordCloud(width=800, height=400, background_color='white').generate(' '.join(topic_words))
plt.figure(figsize=(10, 6))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.title(f'Topic {topic_idx + 1} Word Cloud')
plt.show()
# Stacked bar chart for topic proportions
topic_distributions = lda.transform(X)
topic_df = pd.DataFrame(topic_distributions, columns=[f'Topic {i + 1}' for i in range(lda.n_components)])
topic_df.plot(kind='bar', stacked=True, figsize=(12, 6), colormap='tab20c', edgecolor='black')
plt.title('Proportion of Topics Across Reviews')
plt.ylabel('Proportion')
plt.show()
"""**3.2 Topic Modeling (Table Format)**"""
# Topic statistics
topic_distributions = lda.transform(X)
topic_keywords = [", ".join([vectorizer.get_feature_names_out()[i]
for i in lda.components_[t].argsort()[:-11:-1]])
for t in range(lda.n_components)]
topic_percentages = topic_distributions.mean(axis=0) * 100
# Create table
topics_table = pd.DataFrame({
"Topic": [f"Topic {i+1}" for i in range(len(topic_keywords))],
"Top Keywords": topic_keywords,
"Percentage of Reviews": topic_percentages
})
print(topics_table)
"""**4. Keyword Extraction (TF-IDF)**"""
Integrating Text Mining and SERVQUAL Framework to Analyze Customer Feedback for Service Quality Improvement in the Hospitality Industry
File Edit View Insert Runtime Tools Help
Share
Connect
Gemini
"""**4. Keyword Extraction (TF-IDF)**"""
from sklearn.feature_extraction.text import TfidfVectorizer
# Apply TF-IDF
tfidf = TfidfVectorizer(max_features=100, stop_words='english')
tfidf_matrix = tfidf.fit_transform(df['cleaned_text'])
# Get top keywords
def get_top_keywords(tfidf_matrix, feature_names, n):
top_keywords = []
for i in range(tfidf_matrix.shape[0]):
row = tfidf_matrix[i].toarray().flatten()
top_indices = row.argsort()[-n:][::-1]
top_keywords.append([feature_names[idx] for idx in top_indices])
return top_keywords
df['top_keywords'] = get_top_keywords(tfidf_matrix, tfidf.get_feature_names_out(), 5)
print(df.head())
"""**4.1 Keyword Extraction Visualization (TF-IDF)**
- A bar chart
"""
# Extract most frequent keywords
all_keywords = [keyword for sublist in df['top_keywords'] for keyword in sublist]
keyword_counts = pd.Series(all_keywords).value_counts().head(10)
# Bar chart for top keywords
keyword_counts.plot(kind='bar', figsize=(10, 6), color='purple', edgecolor='black')
plt.title('Top Keywords Across Reviews')
plt.xlabel('Keywords')
plt.ylabel('Frequency')
plt.show()
"""**4.2 Keyword Extraction (Top 10 Words Count Table Format)**"""
from collections import Counter
# Flatten the list of all keywords into a single list
all_keywords = [keyword for keywords in df['top_keywords'] for keyword in keywords]
Integrating Text Mining and SERVQUAL Framework to Analyze Customer Feedback for Service Quality Improvement in the Hospitality Industry
File Edit View Insert Runtime Tools Help
Share
Connect
Gemini
# Count the frequency of each keyword
keyword_counts = Counter(all_keywords)
# Get the top 10 keywords
top_keywords = keyword_counts.most_common(10)
# Create a table for top 10 keywords
top_keywords_table = pd.DataFrame(top_keywords, columns=["Keyword", "Frequency"]).reset_index().rename(columns={"index": "Rank"})
top_keywords_table["Rank"] += 1 # Adjust rank to start from 1
print(top_keywords_table)
"""**5. SERVQUAL Mapping**"""
from sentence_transformers import SentenceTransformer
from sklearn.metrics.pairwise import cosine_similarity
import numpy as np
# Load BERT model
bert_model = SentenceTransformer('all-MiniLM-L6-v2')
# Define SERVQUAL dimensions
servqual_dimensions = {
'Tangibles': ['clean', 'room', 'facility', 'comfortable'],
'Reliability': ['reliable', 'timely', 'accurate', 'consistent'],
'Responsiveness': ['quick', 'responsive', 'helpful', 'support'],
'Assurance': ['professional', 'trustworthy', 'safe', 'confident'],
'Empathy': ['friendly', 'caring', 'understanding', 'personalized']
}
# Generate embeddings for SERVQUAL dimensions
servqual_embeddings = {dim: bert_model.encode(keywords) for dim, keywords in servqual_dimensions.items()}
# Function to map a keyword to a SERVQUAL dimension
def map_keyword_to_servqual(keywords, servqual_embeddings):
similarities = {}
for keyword in keywords:
keyword_embedding = bert_model.encode([keyword])
scores = {dim: max(cosine_similarity(keyword_embedding, [emb]).flatten())
for dim, emb_list in servqual_embeddings.items()
for emb in emb_list}
best_dim = max(scores, key=scores.get)


Integrating Text Mining and SERVQUAL Framework to Analyze Customer Feedback for Service Quality Improvement in the Hospitality Industry
File Edit View Insert Runtime Tools Help
Share
Connect
Gemini
scores = {dim: max(cosine_similarity(keyword_embedding, [emb]).flatten())
for dim, emb_list in servqual_embeddings.items()
for emb in emb_list}
best_dim = max(scores, key=scores.get)
similarities[keyword] = best_dim
return similarities[keywords[0]]
Apply to the DataFrame
df['servqual_dimension'] = df['top_keywords'].apply(lambda x: map_keyword_to_servqual(x, servqual_embeddings))
print(df.head())
"""**5.1 SERVQUAL Mapping Visualization**
- A heatmap
- A bar chart
"""
# Count the number of keywords for each SERVQUAL dimension
dimension_counts = df['servqual_dimension'].value_counts()
# Bar chart for SERVQUAL dimensions
dimension_counts.plot(kind='bar', figsize=(10, 6), color='orange', edgecolor='black')
plt.title('Keyword Frequency by SERVQUAL Dimension')
plt.ylabel('Frequency')
plt.xlabel('SERVQUAL Dimension')
plt.show()
# Heatmap for keyword distribution across SERVQUAL dimensions
keyword_distribution = pd.DataFrame([
[1 if dim in df['servqual_dimension'][i] else 0 for dim in servqual_dimensions.keys()]
for i in range(len(df))
], columns=servqual_dimensions.keys())
plt.figure(figsize=(10, 6))
sns.heatmap(keyword_distribution.T, cmap='coolwarm', annot=True, cbar=True, fmt='d')
plt.title('Keyword Distribution Across SERVQUAL Dimensions')
plt.xlabel('Review Index')
plt.ylabel('SERVQUAL Dimensions')
plt.show()
"""**5.2 SERVQUAL Mapping (Count Table Format)**"""
Count the number of keywords mapped to each SERVQUAL dimension

Integrating Text Mining and SERVQUAL Framework to Analyze Customer Feedback for Service Quality Improvement in the Hospitality Industry
File Edit View Insert Runtime Tools Help
Share
Connect
Gemini
[1 if dim in df['servqual_dimension'][i] else 0 for dim in servqual_dimensions.keys()]
for i in range(len(df))
], columns=servqual_dimensions.keys())
plt.figure(figsize=(10, 6))
sns.heatmap(keyword_distribution.T, cmap='coolwarm', annot=True, cbar=True, fmt='d')
plt.title('Keyword Distribution Across SERVQUAL Dimensions')
plt.xlabel('Review Index')
plt.ylabel('SERVQUAL Dimensions')
plt.show()
"""**5.2 SERVQUAL Mapping (Count Table Format)**"""
# Count the number of keywords mapped to each SERVQUAL dimension
servqual_counts = df['servqual_dimension'].value_counts()
# Calculate the percentage for each dimension
total_keywords = servqual_counts.sum()
servqual_percentages = (servqual_counts / total_keywords) * 100
# Create a table for SERVQUAL mapping
servqual_table = pd.DataFrame({
"SERVQUAL Dimension": servqual_counts.index,
"Number of Keywords Mapped": servqual_counts.values,
"Percentage (%)": servqual_percentages.values
}).reset_index(drop=True)
print(servqual_table)