

**Final Report**

**Level 4**

# **Intelligent Analytics System for Promotion of Tourism and Hospitality**

**Group Name – Team AnalytIQ**

<b>Index Number</b>	<b>Name</b>
185004F	G.Abirame
185050P	Nikalya.N
185079L	Sinthuja. S
185028G	Gobinthiran.K

**Faculty of Information Technology**

**University of Moratuwa**

**2023**

**Final Report**

**Level 4**

# **Intelligent Analytics System for Promotion of Tourism and Hospitality**

**Group Name – Team AnalytIQ**

<b>Index Number</b>	<b>Name</b>
185004F	G.Abirame
185050P	Nikalya.N
185079L	Sinhuja. S
185028G	Gobinthiran.K

Supervised by

Dr.(Ms) U Ganegoda

Mrs MB Mufidha

**Faculty of Information Technology**

**University of Moratuwa**

**2023**

## **Abstract**

This paper proposes the development of an interactive chatbot to address the marketing challenges faced by small and medium tourism businesses in Sri Lanka. The country's tourism industry has suffered from an economic crisis and political turmoil, hindering its growth potential. The chatbot aims to empower these businesses by providing insightful analytics and decision-making support for effective tourism promotion. By analyzing search queries and social media reviews, the chatbot can forecast tourism demand and identify tourists' preferences. Additionally, it offers personalized assistance to business owners, recommends Forecasting ADR prices as Optimized value strategies, and facilitates informed decision-making. Through the adoption of this chatbot, small and medium tourism businesses can overcome resource and expertise limitations, attract foreign funds, and contribute to the revival and sustainable growth of Sri Lanka's tourism industry.

# Table of Contents

Abstract.....	2
Chapter 1.....	6
Introduction.....	6
1.1 Introduction.....	6
1.2 Background & Motivation.....	7
1.3 Problem in Brief.....	8
1.4 Aim & Objectives.....	9
1.5 Proposed Solution .....	9
Chapter 2.....	11
Literature Review .....	11
2.1 Chapter Overview.....	11
2.2 Problem Justification.....	11
2.3 Related work for Search Query based Tourism Forecasting and Analysis .....	12
2.4 Related work for Identify customer’s preferences through analyzing the small medium business hotel’s reviews on social media.....	14
2.5 Related work for Price Optimization.....	15
2.6 Related work for Building a question-answer based chatbot for helping SME businesses in tourism.....	18
2.7 Summary .....	19
Chapter 3.....	20
Technology Adapted.....	20
3.1 Chapter Overview.....	20
3.2 NLP and Large Language Models .....	20
3.3 Python .....	21
3.4 Colab .....	22
3.5 Llamandex.....	22
Chapter 4.....	23
Our Approach .....	23
4.1 Chapter Overview.....	23
4.2 Search Query Analysis .....	23
4.3 Sentiment Analysis.....	25
4.4 Price Optimization.....	27

4.5 Implement Chatbot.....	29
Chapter 5.....	31
Analysis and Design.....	31
5.1 Chapter Overview.....	31
5.2 Top level architectural diagram for Intelligent Analytics System for Promotion of Tourism and Hospitality .....	31
5.3 Search Query Analysis .....	33
5.4 Sentiment Analysis.....	354
5.5 Price Optimization.....	36
5.6 Building conversational chatbot for helping SME businesses in Tourism.....	37
Chapter 6.....	39
6.1 Chapter Overview.....	39
6.2 Search Query Analysis .....	40
6.3 Aspect based Sentiment analysis .....	46
6.4 Price optimization.....	552
6.5 Implement Chatbot.....	60
Conclusion and Further Work.....	68
Appendix A - References.....	70
Appendix B -Individual Contributions.....	75
Appendix C – Survey Form and Responses for Search Query Data .....	79

## Table of Figures

Figure 1: Top Level Diagram .....	32
Figure 2 : Search Query Analysis Flow .....	32
Figure 3 : Sentiment Analysis Flow Diagram.....	33
Figure 4 : Price Optimization Flow .....	34
Figure 5 : Chatbot Analysis diagram .....	35
Figure 6 : Query Preprocessing.....	38
Figure 7 : Genetic Feature Algorithm .....	40
Figure 8 : Feature Selection .....	40
Figure 9 : Feature Selection Evaluation.....	42
Figure 10 : Data categorization of search query .....	42
Figure 11 : Data collection of sentiment analysis .....	45
Figure 12 : Preprocessing of sentiment analysis .....	45
Figure 13 : Keyword Extraction .....	46
Figure 14 : Keyword Extraction .....	47
Figure 15 ; Rule based Approach for auto aspect categorization Model .....	47
Figure 16 : Sentiment Score Prediction .....	48
Figure 17 : Logistic Regression model Train and test .....	48
Figure 18: SVM model Train and test... ..	49
Figure 19: Gradient Boosting Classifier model Train and test.....	49
Figure 20: Random Forest model Train and test.....	50
Figure 21: Hyperparameter Tuning model Train and test .....	50
Figure 22: Predict the aspect and sentiment. ....	51
Figure 23: Evaluation of sentiment analysis .....	51
Figure 24 : Implementation of price optimization .....	53
Figure 25 : Result Interpretation .....	54
Figure 26 : Chatgpt 3 based price optmization.....	56
Figure 27 : Price optimization evaluation .....	58
Figure 28 : Performance evaluation metrices .....	59
Figure 29 : Chatbot work flow .....	62
Figure 30 : Chatbot data collection form .....	63
Figure 31: Response data .....	63
Figure 32 : Data analysis based on categories.....	64
Figure 33 : Overview of the categories .....	64
Figure 34 : Chatbot implementation source code .....	65
Figure 35 : Application View.....	65
Figure 36 : Integrate other modules data .....	66
Figure 37 : Output overview .....	66
Figure 38 : Evaluation.....	65

# **Chapter 1**

## **Introduction**

### **1.1 Introduction**

Sri Lanka has long been recognized as a highly sought-after tourist destination, with its captivating landscapes and rich cultural heritage. However, the country's recent economic crisis and political turmoil have significantly hindered the growth potential of its tourism industry. Despite the gradual reopening of the country following the pandemic outbreak, it is essential to address the challenges faced by small and medium tourism businesses in effectively marketing themselves, as they often lack access to resources and expertise available to larger enterprises.

To mitigate these limitations, this paper presents an innovative solution in the form of an interactive chatbot. The chatbot aims to empower informed decision-making for effective tourism promotion by providing insightful analytics on the preferences of tourists. This is accomplished through several key functionalities. Increasing accuracy in forecasting tourism demand through search query analysis: By analyzing search queries made by potential tourists, the chatbot can identify emerging trends, predict demand patterns, and enable businesses to make data-driven decisions to align their offerings with the evolving market demands.

Identifying tourists' preferences through analyzing reviews on social media: Leveraging natural language processing techniques, the chatbot extracts and analyzes reviews and feedback from social media platforms. This analysis provides valuable insights into tourists' preferences, allowing businesses to tailor their services to better meet customer expectations.

Building a conversational chatbot for assisting SME businesses in tourism: The chatbot serves as a virtual assistant, engaging in interactive conversations with business owners, offering guidance, and providing personalized recommendations based on the specific needs and goals of each enterprise. This empowers SME businesses with the tools and knowledge to navigate the challenges of tourism marketing effectively.

Forecasting ADR prices as Optimed value and recommendation: The chatbot utilizes pricing optimization algorithms to suggest competitive and profitable pricing strategies, taking into account market dynamics, customer behavior, and industry benchmarks. This enables SME businesses to optimize their pricing decisions and enhance their competitiveness in the tourism market.

By leveraging the capabilities of this interactive chatbot, small and medium tourism businesses in Sri Lanka can bridge the gap in resources and expertise. Through enhanced decision-making, accurate demand forecasting, preference analysis, and Forecasting ADR prices as Optimed value, these businesses can effectively promote themselves, attract foreign funds, and contribute to the revival and sustainable growth of Sri Lanka's tourism industry.

## **1.2 Background & Motivation**

Sri Lanka, being a demandingly famous destination for vacationers, should be usually swarming with tourists at this time of the year. Instead, the unprecedented economic crisis and the prevailing political turmoil have all but wreaked the potential growth of the tourism[2]. However, this situation is regrettable, as tourism is one of the few remaining means through which Sri Lanka can attract foreign funds into the country and whereas it has also been on the rebound during the past few months after the country gradually opened following the pandemic outbreak. Thus, today a country of 22 million people is facing a critical economic instability, the worst in seven decades, leaving thousands of lives striving to pass each day, battling to buy food, medicine, fuel, and other essentials. So given the circumstances, tourism can be identified to play a decisive role in resurrecting the country's distressed economy, by attracting foreign exchange earnings and complementing the growth rate in the following years. With the country desperately in need of a solution in the economic front, the tourism sector can be guaranteed to revive its fortunes, which in the past had been a source of much glory for the Sri Lankan economy. Thus, it is obvious that a consistent influx of tourists would undoubtedly aid in the recovery regarding the foreign exchange woes of the country. Moreover, it's high time for high level authorities to vigorously engage in taking necessary promotional actions to attract the interest of tourists and provide them with the required facilities to continue the long standing tourism sector of Sri Lanka.



### **1.3 Problem in Brief**

As Sri Lanka now grapples with the worst financial crisis the country has ever faced, the tourism industry which was a significant source of foreign income is rapidly going off the edge. According to the Sri Lanka Tourism Development Authority (SLTDA), the number of tourist arrivals decreased by 60 per cent in June. Thus, the country is in a critical state and is desperately in need of finding effective solutions to increase their tourist arrivals and promote tourism to help uplift the economy. Taking this critical context into notice, we understand that the Sri Lankan Tourism Development Board and other related government bodies such as the Municipal Councils, must proactively engage in identifying the key aspects that need to be considered for improvement and promotion, regarding variety of tourism destinations. Accordingly, availability of macro level data and deep insights of tourists preferences and interests for attraction of Sri Lankan destinations can empower them to make decisions and act wisely to increase the arrival of tourists and promote tourism. For despite the crisis, fortunately there are still tourists visiting in small amounts which still shows the green light to rebuild the sector[3].

Incorporating a large group of search query data into searching may reveal relevant information, but it may also introduce irrelevant noise and cause problems such as spurious correlation and data overfitting. Thus the need to determine which search keywords should be retained to maintain the accuracy.

## **1.4 Aim & Objectives**

### **1.4.1 Aim**

Empower informed decision making for effective tourism promotion with insightful analytics on preferences of tourists.

### **1.4.2 Objectives**

- ⊄ Critical Analysis of search query for insightful decision making.
- ⊄ Identify Customer's preferences through aspect based sentiment analysis of reviews on social media.
- ⊄ Building a question-answering based chatbot for helping SME businesses in tourism
- ⊄ Price forecasting of hotel prices based on seasonal trends as Optimized forecasted price

## **1.5 Proposed Solution**

An intelligent analysis approach to empower the concerned bodies in the tourist industry to help promote tourism and attract more tourists. The solution will focus on analysing various contributing features that can help marketers to find effective avenues to promote tourism. The proposed system enables in forecasting demand to various locations with increased accuracy and identifying various points of interest that can help them capture the attention of tourists. The major focus point of this solution is to engage in an in-depth of analysis from the tourists perspectives to get a profound understanding of their preferences, point of attractions.

Sentiment analysis approach is used to improve the small medium business like hotel, restaurants based on the analysis of the reviews of the hotels which are posted by the customer in the social media like websites (tripadvisor.com). Analyze the reviews based on some of the aspects of the hotels (Services, Rooms, Food, Facility, Accommodation, Staff and Atmosphere) to categorize the positive, negative and neutral reviews with help of review text.

Chatbot implementation approach proposed to improve SME's businesses' strategies. There are different kind of small medium businesses like hotel, surfing, and restaurants and so on. When the marketers find their solutions through internet, they couldn't find all details on time. Because, they can't find particular data on time. There are lot of information in websites. In such case, Chatbot can summarize the data and provide appropriate response on time. Further chatbot information are trustable than the website because, chatbots are managed by marketers.

An optimized price by forecasting ADR prices is a key aspect of empowering the concerned bodies in the tourist industry to promote tourism and attract more tourists. The proposed system focuses on analyzing various contributing features to help marketers find effective avenues for promoting tourism. By accurately forecasting demand to various locations, the system identifies points of interest that can capture the attention of tourists. A profound understanding of tourists' preferences and attractions is gained through in-depth analysis from their perspectives.

The sentiment analysis approach improves small and medium businesses, such as hotels and restaurants, by analyzing customer reviews posted on social media platforms like TripAdvisor. Reviews are categorized as positive, negative, or neutral based on various aspects of the hotels, including services, rooms, food, facilities, accommodation, staff, and atmosphere. This analysis helps businesses understand customer feedback and make informed decisions to enhance their offerings.

To improve SMEs' business strategies, a chatbot implementation approach is proposed. With a vast amount of information available on websites, marketers often struggle to find specific data in a timely manner. Chatbots provide a solution by summarizing information and delivering appropriate responses promptly. As chatbots are managed by marketers, their information is considered trustworthy, enabling businesses to access reliable and up-to-date information efficiently.

In the research project "An intelligent analysis approach to empower the concerned bodies in the tourist industry to help promote tourism and attract more tourists," the Recommendation and Visualization Based on Markdown Optimization module aims to develop a system that provides

personalized recommendations and insights to marketers in the tourist industry. By analyzing data from multiple modules, including search query analysis, review analysis using social media, and chatbot interactions, the system leverages machine learning and optimization techniques to identify trends and patterns. These insights are used to optimize marketing efforts and attract more visitors. Visualizations of the recommendations are presented through a high-performance dashboard, enabling marketers to easily understand and utilize the information.

### **1.5.1 Summary**

This chapter provides an overview of the research project, focusing on the introduction, background, motivation, problem statement, aim, and objectives. The primary goal of this chapter is to establish the context and rationale for the study, outlining the key factors that led to the identification of the problem and the subsequent aims and objectives. The chapter begins by introducing the research project and its significance in the field of tourism marketing. Next, the background and motivation behind the research project are presented. This includes a discussion of the current state of the tourism industry, emphasizing the evolving trends, increasing competition, and the need for effective marketing strategies to attract tourists. Following the background and motivation, the problem statement is succinctly described. Subsequently, the aim and objectives of the research project are presented.

## **Chapter 2**

### **Literature Review**

#### **2.1 Chapter Overview**

This chapter will discuss the existing work of each of the 4 modules of the proposed system. It will present the summary of the approaches, findings and limitation of the prior studies relevant to each module

#### **2.2 Problem Justification**

Currently, there are a lot of tourism destinations around the world, but not each can boast of its difference and creativity in the promotion of its tourist product. Given the Sri Lankan context,

comparatively different tourist destinations are becoming interchangeable and exhaustible, resulting in a reduction in tourist arrivals. To prevent that, it is necessary to apply marketing techniques as a tool for attracting tourism flows. However, there is an intense lack of specifically customized marketing system and tools personalized for the tourism industry base SME businesses.

### **2.3 Related work for Search Query based Tourism Forecasting and Analysis**

Forecasting tourist demand is an established research field. Noncausal time series, econometric, and AI-based techniques are the three primary categories of modeling methodologies. Researchers have discovered that big data from the Internet, such as data from search queries, can be utilized to properly forecast travel demand. In particular, search query data can enhance conventional data sources and reflect tourist behavior to forecast tourism demand (Choi and Varian 2012; X. Yang et al. 2015). Y. Search query data from Google and Baidu are most widely used for different predicted contexts. X. Yang et al. (2015) indicated that Baidu search data perform better when forecasting tourist arrivals in China, while Google search data are found to be suitable for the forecasting of countries and cities that mainly speak English.

Search engine data, which consists of real-time daily, weekly, and monthly search queries entered by users on platforms like Google and Baidu, have become valuable sources of information for tourism forecasting (Song et al., 2019). Among the 25 reviewed papers on forecasting with search engine data, Google Trends and Baidu Index were the primary sources for generating search query indexes or volumes (Yang et al., 2015a). Out of these papers, 36% utilized Baidu data, while 64% relied on Google data for their forecasting purposes. Baidu data have found extensive use in forecasting tourist arrivals to destinations, scenic areas, and hotels in China, including Chinese tourists' travel to other countries such as Thailand (Tang, 2018; Li et al., 2017, 2018; Huang et al., 2017; Li et al., 2019; Zhang, Pu, & Wang, 2019).

On the other hand, Google data have mainly been applied in forecasting tourism arrivals and hotel occupancy for various countries such as the US, Spain, Italy (Florence, Milan, and Catania), Germany, the United Kingdom, South Korea, China, Hong Kong, and Macau, particularly from

Western countries (Choi & Varian, 2012; Emili, Figini, & Guizzardi, 2019; Law et al., 2019; Park et al., 2017; Rivera, 2016). The selection of search engine data at different frequencies, based on specified keywords, varied depending on the context being predicted. Most articles utilized monthly search query data for modeling and forecasting, while only a few directly used weekly search engine data without transformation (Pan et al., 2012; Bangwayo-Skeete & Skeete, 2015). Daily search data proved suitable for forecasting tourist attractions like the Forbidden City in China and hotel occupancy (Huang et al., 2017; Pan et al., 2012). High-frequency search engine data can also be utilized directly to forecast low-frequency tourism demand by employing mixed data-sampling frameworks, enhancing the efficiency of modeling and forecasting (Bangwayo-Skeete & Skeete, 2015; Song & Liu, 2017).

Regarding the selection of keywords for search engine data, the reviewed articles primarily relied on domain knowledge related to tourism demand and the search query index within specific categories. Initial works, such as Pan et al. (2012) and Choi and Varian (2012), used a limited number of tourist-related keywords from Google Trends to predict hotel room demand or forecast specific vacation destinations. Subsequent research expanded the number of keywords to encompass comprehensive aspects of tourist activities. For instance, Li et al. (2017) collected Baidu search data using 46 keywords to predict tourism demand in Beijing, China. Law et al. (2019) utilized 211 Google and 45 Baidu keywords to obtain search engine data for predicting tourism demand in Macau, aiming to cover a wide range of keywords related to tourists' interests in the destination (Law et al., 2019).

Performance of tourism forecasting is strongly influenced by keyword selection from search queries. In the existing literature, many studies have defined keywords on the basis of prior domain knowledge and then collected search query data from Baidu or Google to represent tourists' interests. However, the question of how to strike a compromise between forecasting performance accuracy and search keyword coverage, has not yet been answered. In other words, an effective method for choosing search query data has not yet been suggested.

## **2.4 Related work for Identify Customer's preferences through analyzing the small medium business hotel's reviews on social media**

S. Asharaf and V.S. Anoop used a dataset of 194,439 Amazon reviews from May 1996 to July 2014 for their research. For topic modeling, they used the LDA (Latent Dirichlet Allocation) algorithm. At first, they used collections of uni-gram words from customer reviews to represent topics. They extracted elements based on likelihood from these themes. Then, Naive Bayes was used to analyze sentiment. Prior to using LDA to extract topics from the data, the data underwent pre-processing. The algorithm scanned the pre-processed corpus to extract aspect-specific phrases by matching them with subject terms. Each topic was given to a particular aspect. For each factor, the method had accuracy ranging from 74% to 81%. The automatic extraction of aspects and the creation of an algorithm to map subjects to the extracted aspects are among the tasks to be completed in the future.

Sentiment analysis is performed on the basis of user reviews using three different classifiers as Naive Bayes, Random Forest and Support Vector Machine (Saman Zahid,2020). Text mining using Natural Language Processing (NLP) techniques is often used to analyze one's responses and reviews to perform sentiment analysis. In this project they used 1000 hotels data to analyze the three polarity levels(positive, negative and average). In the Text classification process data preprocessing (tokenization,data cleaning,feature selection) happens after collecting the data from kaggle. performance of each algorithm with different settings is evaluated on test data. SVM and multinomial naive bayes perform really well as compared to the most complex of these three that is Random forest. The small amount of data could be the reason for it. English language Reviews are mostly used in the existing research rather than other languages.

Researchers are investigating the possibilities of deep learning algorithms in sentiment classification and aspect extraction as they have surpassed more conventional machine learning techniques in popularity. These deep learning techniques provide speed and error correction that is automatic. Deep neural networks frequently use convolutional neural networks (CNN) because they are efficient with little datasets and produce better results. However, using CNN by itself to do aspect-based sentiment analysis frequently fails to increase accuracy. To combat this, researchers have coupled deep learning methods with CNN, LSTM, and GA (Genetic Algorithm)

to provide better outcomes. In the research of Adnan Ishaq, Sohail Asghar, and Saira Andleeb Gillani, aspect-based sentiment analysis is performed using CNN, while data cleaning is accomplished using the Panda library. The precision, accuracy, recall, and F1 metrics obtained using the CNN technique are 91.6%, 93.4%, 92.2%, and 89.23%, respectively .

Document, phrase, and entity/aspect levels of sentiment analysis are all used. The sentiment analysis classifies the entire document (such as a review or a collection of tweets) as positive or negative at the document level. On the other hand, sentence-level classification classifies specific sentences inside the document as positive, negative, or neutral. The feature levels, sometimes referred to as the entity and aspect levels, concentrate on assessing attitudes toward certain entities or features stated in the text. The university setting was split into six sectors for the study by Zhiwen Song and Jianhong (Cecilia) Xia: science and engineering buildings, social science buildings, libraries, lecture halls, dorms, amusement, and parking spaces. The study also took four time periods into account: before the exam period, during the exam period, in the middle of the semester, and at the end of the semester. This made it possible to analyze sentiment across several time zones and zones of emotion. The researchers' data set consisted of almost 5000 tweets published between May 12, 2014, and January 5, 2015 .

## **2.5 Related work for Forecasting ADR prices as Optimed value**

### **2.5.1 Introduction**

This chapter reviews the existing literature concerning Forecasting ADR prices as Optimed value in the hotel industry. The aim is to understand the methodologies and techniques currently used in this area, their benefits, drawbacks, and how our proposed approach fits into the current knowledge landscape.

### **2.5.2 Forecasting ADR prices as Optimed value in the Hotel Industry**

Forecasting ADR prices as Optimed value is a crucial aspect of the hotel industry, aimed at maximizing revenue (Kimes, 1989 [1]). Various techniques have been developed, ranging from



simple pricing strategies to complex dynamic pricing models, to achieve effective revenue management in this industry (Ivanov & Zhechev, 2012 [2]).

### **2.5.3 Time Series Forecasting**

The advent of machine learning and AI has paved the way for their increased utilization in forecasting. Hybrid models that combine traditional time-series techniques with machine learning algorithms have shown improvements in forecast accuracy (Tsai et al., 2015 [46]).

### **2.5.4 Machine Learning and AI in Forecasting**

With the advent of machine learning and AI, these technologies have increasingly been used for forecasting. Hybrid models that combine traditional time-series techniques with machine learning algorithms have shown to improve the accuracy of forecasts (Tsai et al., 2015).

### **2.5.5 Use of Prophet for Forecasting**

One recent development in time series forecasting is Facebook's Prophet model, which has demonstrated promising results in various fields, including tourism (Taylor & Letham, 2017 [47]). Prophet's strengths lie in its ability to effectively handle seasonality, trends, and holidays, making it particularly relevant in the hotel industry.

### **2.5.6 Natural Language Processing with GPT-3 LLM**

OpenAI's GPT-3, a powerful language prediction model, has revolutionized the field of Natural Language Processing. GPT-3 has been employed for tasks such as text generation, language translation, and context-based question answering (Brown et al., 2020 [48]). In a business context, GPT-3 holds potential for applications in chatbots, report generation, and data analysis.

### **2.5.7 The Intersection of Forecasting and AI**

Our approach falls at the intersection of traditional forecasting methods and AI techniques. By utilizing the Prophet model and integrating GPT-3, we aim to enhance user interaction and comprehension of complex forecasting models. This novel application of GPT-3 for interacting with tabular data builds upon recent research on transforming structured data into natural language for AI processing (Chen et al., 2019 [49]). This approach offers more interactive and user-friendly ways of exploring data and interpreting results from forecasting models.

### **2.5.8 Antonio and Taylor's Work in Forecasting ADR prices as Optimized value**

Antonio, Almeida, and Nunes (2019) [42] have contributed significantly to the field of Forecasting ADR prices as Optimized value through their research. They conducted a comprehensive study focusing on the application of forecasting models in the hotel industry. The study aimed to enhance pricing strategies and revenue management by leveraging advanced analytical techniques.

In their work, Antonio, Almeida, and Nunes explored the use of time series forecasting models, including traditional statistical approaches and machine learning algorithms. They analyzed the historical data of ADR rates in the hotel industry to develop accurate forecasting models. These models were trained to capture seasonality, trends, and other significant factors affecting ADR rates.

The study emphasized the importance of accurate forecasting in the hotel industry, as it enables revenue managers and decision-makers to make informed pricing decisions. By utilizing the insights derived from their forecasting models, Antonio, Almeida, and Nunes provided valuable guidance for revenue optimization, cost control, and competitive positioning.

Furthermore, their work highlighted the potential of incorporating advanced forecasting models into the Forecasting ADR prices as Optimized value process. By integrating data-driven forecasting techniques, businesses in the hotel industry can enhance their pricing strategies and gain a competitive edge.

Antonio and Taylor's research, along with their proposed forecasting models, have laid the foundation for further advancements in Forecasting ADR prices as Optimed value. Their contributions have not only improved revenue management practices but have also inspired further research and innovation in the field.

### **2.5.9 Conclusion**

In conclusion, our approach presents a novel application of the Prophet model and GPT-3 for Forecasting ADR prices as Optimed value and forecasting in the hotel industry. The literature suggests potential for improved user interaction and comprehension of complex forecasting models through this approach. Further research and practical application are necessary to evaluate its efficacy and potential implications for the hotel industry.

## **2.6 Related work for Building a conversational chatbot for helping SME businesses in tourism**

“AI-Driven Conversational Agents for SME Tourism Businesses” by Lee et al. (2022) research paper investigated the use of AI-driven conversational agents to support small- and medium-sized tourism businesses. Research has highlighted the benefits of chatbots in providing personalized recommendations, improving customer engagement, and handling customer inquiries. It presents a case study of a chatbot implemented for a group of SME travel agencies, demonstrating a positive impact on customer satisfaction and business growth.

However, in general, the methodology section of a research paper on AI-driven conversational agents for SME tourism businesses may include elements such as research design, data collection methods (surveys, interviews, chatbot interaction logs), case selection, data analysis techniques, and findings/discussion. Specific details of the methodology are outlined in the paper itself, including sample size, interview protocols, survey questions, data analysis procedures, and any specific tools or software used.

“Chatbots Opportunities and Challenges for Tourism Destination Marketing” Kim et al. (2018) research paper examines the use of chatbots for tourism destination marketing and the associated opportunities and challenges with a focus on SMEs. Interactive experiences for tourists, real-time information and the authors examine the potential benefits of chatbots in providing personalized recommendations. They also discuss challenges related to data integration, privacy concerns, and technology adoption. These findings provide insights into considerations and strategies for implementing chatbots in the SME tourism sector.

Here interview and survey methods have been used to collect data. Then qualitative analysis and quantitative analysis have been done for data analysis. This analysis helped measure relationships between user perceptions, satisfaction levels, and variables related to chatbot usage.

Also, this study describes the technical aspects of developing chatbots for tourism destination marketing. It involves discussing programming languages, frameworks and tools used in the development process. The study also describes the architecture and components of the chatbot system, such as natural language processing (NLP) algorithms, conversation management techniques, and backend integration. It also includes discussing methods for connecting the chatbot to relevant data sources such as databases, APIs, or content management systems. The paper may also explore techniques for obtaining real-time information, maintaining data accuracy, and updating destination-related data in a chatbot system.

## **2.7 Summary**

This chapter provides a comprehensive review based on extensive research, focusing on existing solutions pertaining to the individual models. It has diligently examined the shortcomings associated with existing solutions, which have served as valuable insights for the development of our application. By identifying and addressing these drawbacks, we aim to enhance the functionality and effectiveness of the proposed application, ultimately providing an improved solution to the identified problem.

# Chapter 3

## Technology Adapted

### 3.1 Chapter Overview

In this chapter, we explore the technology that has been adapted to solve the problem at hand. We delve into the techniques and tools that have been specifically chosen and customized to address the challenges faced in our problem domain. The goal is to highlight the appropriateness of these techniques and demonstrate how they effectively tackle the problem, leading to desired outcomes and results.

### 3.2 NLP and Large Language Models

Natural language processing (NLP) is a branch of artificial intelligence (AI) that enables computers to comprehend, generate, and manipulate human language. Natural language processing has the ability to interrogate the data with natural language text or voice.

Natural Language Processing (NLP) techniques, combined with large language models and vector stores, play a crucial role in developing the proposed interactive chatbot for tourism promotion. Here is a description of how NLP can be employed in different aspects of the application:

**Search Query Analysis:** Large language models, such as GPT-3.5, can be utilized to understand and analyze the search queries made by potential tourists. These models have been trained on vast

amounts of text data and possess a deep understanding of language, enabling them to accurately interpret and extract meaningful insights from search queries. By analyzing these queries, the chatbot can identify emerging trends, predict tourism demand patterns, and assist businesses in making informed decisions.

**Sentiment Analysis and Preference Extraction:** Vector stores, such as word embeddings, enable the representation of words or phrases as dense vectors in a high-dimensional space. These vectors capture semantic relationships between words, allowing for effective sentiment analysis and preference extraction from social media reviews. NLP techniques, like sentiment analysis algorithms and word embedding models (e.g., Word2Vec or GloVe), can be employed to analyze and categorize the sentiment and preferences expressed in the reviews. This analysis provides valuable insights into tourists' opinions and expectations, enabling businesses to align their offerings accordingly.

**Conversational Chatbot Development:** Large language models, such as GPT-3.5, can be utilized to build the conversational capabilities of the chatbot. These models have been trained on a wide range of conversational data, enabling them to generate coherent and contextually relevant responses. By leveraging the conversational abilities of such models, the chatbot can engage in interactive conversations with business owners, providing guidance, answering queries, and offering personalized recommendations.

**Forecasting ADR prices as Optimized value and Recommendation:** NLP techniques can be employed to analyze market dynamics, customer behavior, and industry benchmarks. Large language models can assist in understanding pricing patterns and suggesting optimized pricing strategies. By training these models on pricing data and customer preferences, businesses can leverage them to recommend competitive and profitable pricing decisions. This helps SME businesses enhance their competitiveness and maximize revenue.

### **3.3 Python**

Python is a versatile programming language that can be used effectively in developing the mentioned tourism promotion application. Python has excellent libraries like BeautifulSoup and Scrapy that facilitate web scraping. These libraries allowed us to extract relevant data from google based website results, reviews, ratings, financial pricing info and other information from social media platforms or tourism-related websites. Python's simplicity and rich ecosystem made it an ideal choice for gathering data from various online sources.

The scikit-learn library (also known as sklearn) is a widely used machine learning library in Python. It provides a comprehensive set of tools for various machine learning tasks, including classification, regression, clustering, dimensionality reduction, and model selection. It has been used in the keyword extraction process for a fitness function used to evaluate the quality or relevance of selected keywords and further enabled to encode them into numerical representation.

### **3.4 Colab**

It is a cloud based jupyter notebook environment for writing and executing arbitrary python code through the browser to machine learning. This contains text processing libraries for tokenization, parsing, classification, and semantic. Open source library. It lets users run code on remote servers, obviating the need for local hardware resources, and seamlessly integrates with other Google services such as Google Drive for storing and accessing data files. Overall, Google Colab offers a robust and accessible tool for Python project collaboration, experimentation, and coding. It is a cloud based jupyter notebook environment for writing and executing arbitrary python code through the browser to machine learning. This contains text processing libraries for tokenization, parsing, classification, and semantic. Open source library.

### **3.5 Llamandex**

It's a text-davinci-003 based framework that is used for implementing large language model applications. That means it provides some tools for building applications: provides data connectors to ingest existing data sources and data formats (APIs, documents, PDFs, SQL, etc.). Provides ways to structure data (graphs, symbols) so that this data can be easily used during development. The Llama Index, a project that links LLMs to external data, switched from using GPT to using Llama due to its open source nature.

# **Chapter 4**

## **Our Approach**

### **4.1 Chapter Overview**

The systematic and theoretical analysis of the methodologies used in a given field of study is known as methodology. It entails a critical assessment of the methods and guiding ideas used in research, including its conception, constraints, and potential biases. Methodology essentially describes how a study is approached and carried out.

In this chapter, we present the approach adopted to solve the problem at hand by leveraging advanced technology. We describe how the chosen technology is utilized to address the specific challenges and requirements associated with the problem statement. The discussion is framed in terms of the users, inputs, outputs, process, and the underlying technology that implements the solution. The primary objective of this chapter is to provide a comprehensive understanding of how we have harnessed technology to tackle the identified problem effectively. By outlining the various components of the solution and their interdependencies, we aim to demonstrate the feasibility and efficacy of our approach.

### **4.2 Search Query Analysis**

This module focuses on utilizing google search query data to analyze and understand the market demand in the tourism industry. This module provides a powerful tool for tourism forecasting,



enabling users to tap into real-time search engine data and leverage the insights gained to make informed decisions in various tourism-related domains, such as destination planning, hotel management, and marketing strategies.

#### **4.2.1 User**

Small Medium Business owners can use this module output to forecast the demand for their services beforehand, thus enabling them to allocate their resources and plan ahead proactively.

#### **4.2.2 Process**

The development of the search query analysis module consists of two distinct phases. In the first phase, the module focuses on keyword selection, taking into consideration various aspects of tourism such as destination, food, activities, accommodation, travel, and dining. To accomplish this, the search query data is preprocessed, and sophisticated machine learning-based feature selection algorithms are employed to identify and extract keywords that are highly relevant to the specific tourism context under investigation.

Once the keywords are extracted, they serve as input for querying Google Trends, a powerful tool that provides insights into search interest patterns over time. By analyzing the search interest data obtained from Google Trends, the module gains valuable information regarding the popularity and trends associated with the chosen keywords. This data is then utilized to make predictions about the potential demand related to the selected tourism aspects.

By leveraging the search interest data obtained from Google Trends, the module offers the capability to forecast and estimate the level of demand for specific tourism products, services, or destinations. This information can be immensely valuable for decision-making processes, allowing stakeholders in the tourism industry to understand and anticipate the preferences and interests of potential tourists.

Overall, the search query analysis module effectively combines machine learning techniques with the wealth of data available through Google Trends. Through the selection of relevant keywords

and the analysis of search interest patterns, the module provides actionable insights that can significantly contribute to forecasting and predicting the potential demand in various tourism domains.

#### **4.2.3 Input**

Search Query dataset. Google trends data

#### **4.2.4 Output**

Forecasts on different tourism aspect relative to Galle.

### **4.3 Sentiment Analysis**

This module's objective is to analyze hotels reviews based on auto categorize the aspects using the rules-based approach and furthermore analyze the sentiment based on reviews as positive, negative, neutral with help of different aspects and sentiment polarity to understanding the customer preference of the hotels. For that using different approaches to compare the accuracy of the sentiment and aspect accuracy.

#### **4.3.1 User**

Small medium business (hotels) owners are using this module to improve their business in furthermore based on the customer preference related their hotels.

#### **4.3.2 Process**

**Data collection:** Small medium business-like hotel's reviews are scraped and collected from the hotel's own websites, Trip advisor.com and booking.com using the parse hub application in this project and store them as csv format. Tools like pandas, NumPy, and nltk are used to preprocess the data for our projects. Before preprocessing the review text need to convert emojis as text. In this preprocess stage cleaning the text, removing the irrelevant information, removing stop words, removing punctuation, removing new lines, tokenizing, etc. After that based on the preprocess text do the keyword extraction using TFIDF to collect the set of keywords to predefine the dictionary to the rule based approach. Based on keyword's frequency easy to find the high amount of

frequency word as aspects. For that select seven aspects like service, food, accommodation, atmosphere, staff, facilities and room. Based on the rule-based approach, able categorized the aspect in automatically when give the new reviews. After that analyze sentiment for review text with help of sentiment polarity score. This sentiment polarity score will check the review sentence and predict the score related to aspect whether positive, negative and neutral related values are available if available then provide the score for that. Based on this analyze then provide overall polarity score to each review sentence. Based on that overall score if the score  $>0.5$  then predict the review as positive. If the score  $<-0.5$  then predict the review as negative. if the score between 0.5 and -0.5 then it will predict as neutral. Based on sentiment prediction and autocategorization for aspect train and test the dataset for different algorithms like logistic regression, SVM, Random Forest, Hyperparameter tuning and Gradient Boost Classifier. Based on the fine tune model to predict the accuracy, F1-Score, Recall, Precision matrix performance for aspect and sentiment prediction.

#### **4.3.3 Input**

Social media reviews like Hotel's own websites, Trip advisor.com and booking.com

#### **4.3.4 Output**

Based on the auto categorized the aspect predict the reviews as positive, negative and neutral reviews, small medium Business(hotel's owners) can enhance their business than their competitors by understanding the customer preferences.

## **4.4 Forecasting ADR prices as Optimed value**

### **4.4.1 User**

The Forecasting ADR prices as Optimed value process involves two primary approaches. The first approach utilizes the forecasting capabilities of the Facebook Prophet model, a robust tool for time-series forecasting. This model leverages historical hotel pricing data to predict future prices, taking into account trend components and seasonality. By adopting this approach, users gain valuable insights into potential future trends, enabling strategic planning and informed decision-making in setting hotel prices.

The second approach integrates tabular data with GPT-3, a powerful language model developed by OpenAI. This unique method enables users to interact with the data through a chatbot interface, engaging in natural language conversations to extract valuable insights. GPT-3's text generation capabilities are employed to provide additional context, qualitative insights, and address user queries related to the forecasted data. The aim is to enhance the user experience by leveraging GPT-3's ability to generate human-like responses and deliver meaningful information.

### **4.4.2 Process**

The Forecasting ADR prices as Optimed value process involves two primary approaches. The first approach utilizes the forecasting capabilities of the Facebook Prophet model, a robust tool for time-series forecasting. This model leverages historical hotel pricing data to predict future prices,

taking into account trend components and seasonality. By adopting this approach, users gain valuable insights into potential future trends, enabling strategic planning and informed decision-making in setting hotel prices.

The second approach involves implementing an LSTM (Long Short-Term Memory) model, a type of recurrent neural network known for its ability to capture long-term dependencies in sequential data. This model is employed as part of the system implementation for the Forecasting ADR prices as Optimized value application. It contributes to the forecasting process by providing accurate predictions of future ADR rates based on historical data patterns.

In addition to the LSTM model, GPT-3, a powerful language model developed by OpenAI, is incorporated into the system for the application. The GPT-3 model is utilized to enhance the user experience and facilitate natural language interactions through a chatbot interface. Users can input queries or commands in natural language, and GPT-3 generates text-based responses to provide qualitative insights, clarifications, or contextual information based on the input data. These responses aim to assist users in understanding the forecasted data and making more informed decisions related to pricing strategies.

#### **4.4.3 Input**

The input data for the Forecasting ADR prices as Optimized value process consists of historical hotel pricing data organized in a time-series format. Additional variables such as service type (e.g., full board, half board) may also be included to provide further context. This data is utilized for training both the Facebook Prophet model and the LSTM model, enabling accurate forecasting of future ADR rates.

#### **4.4.4 Output**

The output from the Facebook Prophet model includes forecasts of future hotel prices, visualized through easily interpretable plots that highlight trends and seasonality patterns. These forecasts serve as valuable guidance for decision-making processes related to hotel pricing, assisting in the optimization of profits and customer satisfaction.

The output from the system's chatbot interface, powered by GPT-3, comprises generated text responses to user queries. These responses provide qualitative insights, clarifications, or contextual information based on the input data. Users can gain additional understanding of the forecasted data, receive recommendations for pricing strategies, or obtain answers to various questions they may have regarding the data.

Together, these two outputs—the forecasts from the Facebook Prophet model and the text responses from GPT-3—provide a comprehensive toolset for hotel Forecasting ADR prices as Optimized value. By combining quantitative forecasting with qualitative insights, the system empowers users to make data-informed decisions, optimizing pricing strategies, and maximizing revenue in the hospitality industry.

## **4.5 Implement Chatbot**

In the module described as to provide users with personalized and tailored experiences. Unlike generic chatbots that provide standard responses, a personalized chatbot is designed to understand and adapt to the specific needs, preferences and context of individual users. It also focuses on innovating and solving some existing problems (for example legal issues, some limitations ....) with new technologies unlike earlier developed chatbots.

Moreover, when marketers find their problems and solutions through the Internet, they cannot find the required details at the right time. By using chatbot they can avoid unnecessary information and clarify their personal doubts.

### **4.5.1 User**

Owners of small and medium enterprises (hotels) use this module to learn foremost ways and ideas to further improve the accuracy and efficiency of their enterprises by chatting through personal questions quickly and summarized.

### **4.5.2 Process**

Data was used from survey forms and other modules' data to train the chatbot. The data collected from the survey form was using data analysis for categorize the challenges and validate the datas

with other three modules data. After that, the chatbot was implemented with these. A Framework called LlamaIndex is used for that. It can be used to develop Large Language Model (LLM) applications. Provides data connectors to ingest existing data sources and data formats (APIs, documents, PDFs, SQL, CSV....). Provides ways to index the data (vector store Index) so that this data can be easily used. It easily and quickly understands the user's natural language and increases system responsiveness using textblob library.

Initially those modules data was saved in csv format in google drive and then upload to the application as LLM was used in Open AI. After that pandas AI loader (Llma hub loader) was used to access the document in csv format. In this Pandas Library the data is converted to a frame. Then by putting it in the loader it will execute the index method. After that the system works by querying the LLM.

#### **4.5.3 Input**

Collected data from questionnaires. Also additional data that has been get from the other three modules (Search Query Analysis, Review Analysis using Social Media, and Recommendation & visualization).

#### **4.5.4 Output**

Based on the chatbot responses SME's marketers can promote their businesses. Also can be improved their accuracy and efficiency.

# **Chapter 5**

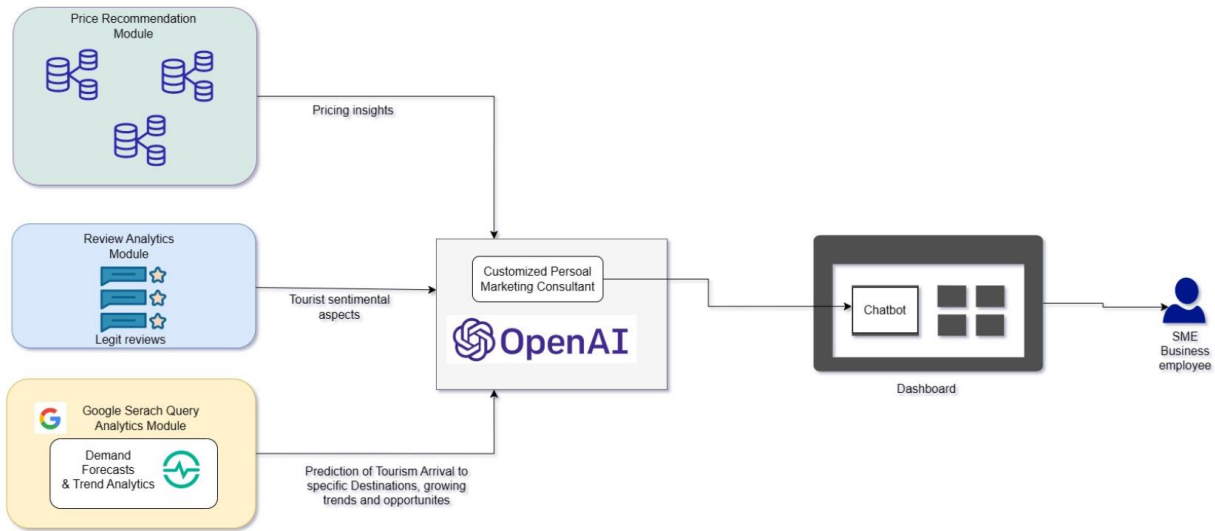
## **Analysis and Design**

### **5.1 Chapter Overview**

This chapter presents an overview of the design and analysis of the developed system. It focuses on providing a top-level understanding of the system's design by outlining the various modules and their interactions. Additionally, this chapter includes separate diagrams for each module to illustrate their functionality and contribution to the overall system.

### **5.2 Top level architectural diagram for Intelligent Analytics System for Promotion of Tourism and Hospitality**





*Figure 1: Top Level Diagram*

The architecture diagram showcases four key modules within the system. The first module is the "Search Query Analysis" module, responsible for processing Google search queries and extracting relevant keywords. These keywords serve as input for the "Google Trends" module, which analyzes search interest patterns over time.

The output of the Search Query Analysis module, along with the insights gained from Google Trends, is then passed as input to the "Chatbot" module, which is the fourth module in the architecture. The Chatbot module utilizes the information obtained from both the Search Query Analysis and Google Trends modules to provide interactive and informative responses to user queries.

The second module in the architecture is the "Sentiment Analysis" module. This module analyzes hotel review data to gain insights into customer's preferences and sentiments. The output of the Sentiment Analysis module is also directed to the Chatbot module, enhancing its ability to understand and respond to user inquiries accurately.

The third module, known as the "Forecasting ADR prices as Optimed value" module, scrapes financial data from various hotel websites and performs an analysis to forecast potential prices for different services. The output of this module is then fed into the Chatbot module, enabling it to provide up-to-date and optimized pricing information to users.

Overall, the architecture diagram illustrates a cohesive system where the Search Query Analysis, Sentiment Analysis, and Forecasting ADR prices as Optimed value modules work synergistically to gather relevant data, extract valuable insights, and provide personalized responses through the Chatbot module. This architecture ensures a comprehensive and user-centric approach, enhancing the system's capabilities in addressing user queries related to tourism, sentiment analysis, and Forecasting ADR prices as Optimed value.

### **5.3 Search Query Analysis**

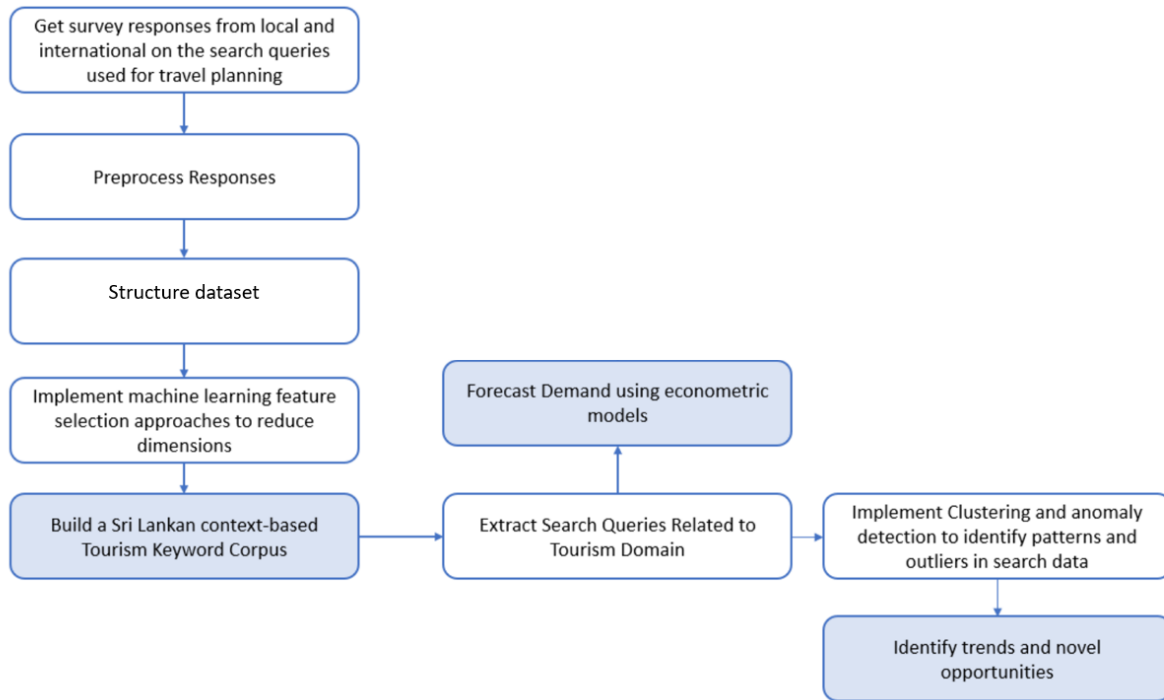


Figure 2 : Search Query Analysis Flow

This module processes Google search queries and extracts relevant keywords. It employs techniques such as natural language processing and machine learning algorithms to analyze the search queries. The module performs data preprocessing tasks to clean and normalize the search query data, ensuring its suitability for further analysis.

Using feature selection algorithms, the module identifies keywords that are highly relevant to the specific tourism context under investigation, such as destination, food, activities, accommodation, travel, and dining. The extracted keywords serve as input for the Google Trends module, enabling the analysis of search interest patterns over time.

The Google Trends api leverages the selected keywords from the Search Query Analysis module to query the Google Trends platform. By querying Google Trends, the module retrieves search interest data associated with the chosen keywords. This data provides insights into the popularity, trends, and fluctuations in search interest over time.

The module analyzes the search interest data, identifying patterns, seasonality, and emerging trends related to the keywords. The output of the Google Trends module, which includes the search interest data and derived insights, is then passed as input to the Chatbot module.

## 5.4 Sentiment Analysis

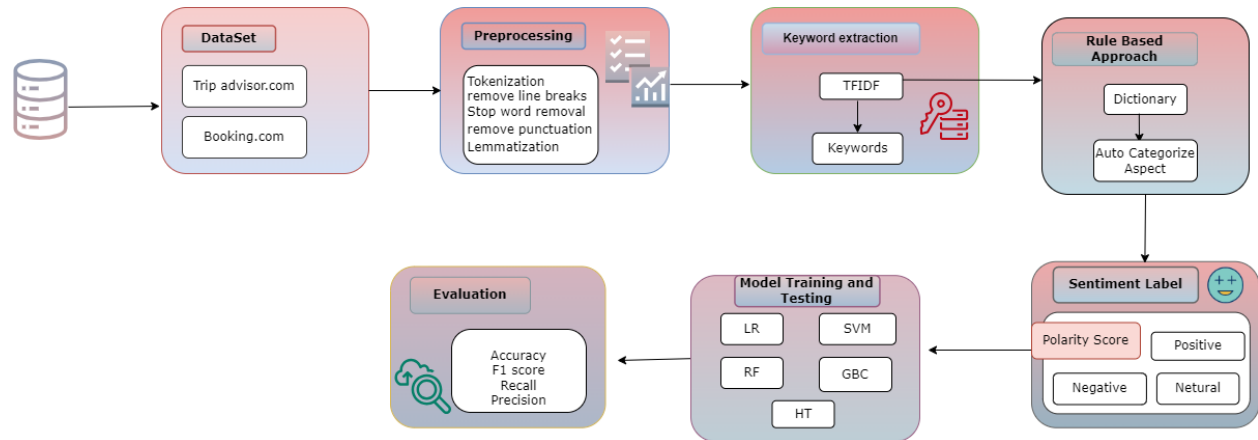


Figure 3 : Sentiment Analysis Flow Diagram

The Aspect based Sentiment Analysis module mainly focuses on analyzing hotel reviews' data to understand the customer preferences and sentiments based on aspects. For that review data scraped from the social media. After that do the preprocessing for the review text in this stage remove the punctuation, tokenization etc.

After that did the key word extraction using TFIDF to get the predefined set of keywords to do the rules-based approach to automatically categorized the aspects from the reviews. For that set the rules based on the aspects as key and assign the values to those keys in the pattern format. When provide the new review that will automatically predict which kind of aspects are available review sentences with help of mapping the key-value pairs which are defined by the rule-based approach. After that predict the sentiment prediction for reviews based on these aspects related positive, negative and neutral using sentiment polarity score. After that based on auto categorized

Rule based approach and sentiment prediction to train the dataset and test based on different algorithms to compare the accuracy, F1-score, Recall and Precision. Based these values can

compare this module is better than exiting researchers. By analyzing the sentiment of the reviews, the module gains insights into customers' opinions, satisfaction levels, and specific preferences regarding their hotel experiences.

The output of the Aspect based Sentiment Analysis module, which includes sentiment prediction and aspects and derived sentiment insights, is sent as input to the Chatbot module.

## 5.5 Forecasting ADR prices as Optimed value

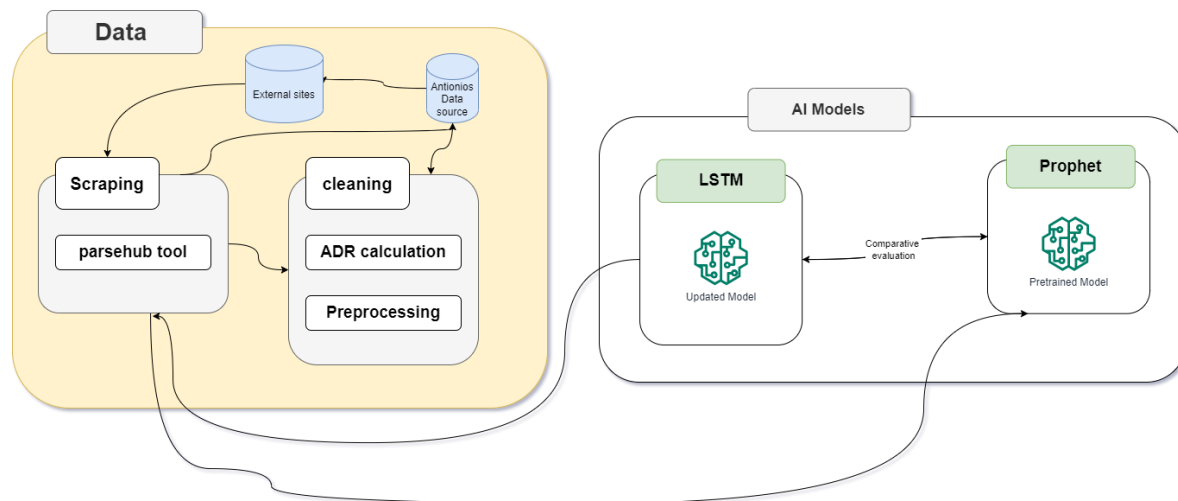


Figure 4 : Chatbot Flow

The analysis phase began with the acquisition of ADR pricing data, which was sourced from Antonio, Almeida, and Nunes in 2019 [42]. This dataset served as the foundation for training the predictive models. Preprocessing techniques were applied to clean and transform the data, ensuring its suitability for the forecasting models.

The first model employed in this study was Facebook Prophet, a time series forecasting tool [43]. The Prophet model offers flexibility and robustness in handling various data patterns and

seasonality. By leveraging its capabilities, we aimed to capture the underlying patterns and trends in the ADR rates over time.

The second model utilized was LSTM (Long Short-Term Memory), a type of recurrent neural network known for its ability to capture long-term dependencies in sequential data [44]. LSTM models are well-suited for time series forecasting tasks, and we sought to leverage their power in predicting future ADR rates accurately.

During the design phase, careful consideration was given to the implementation of these models in a chatbot framework. The chatbot was developed to provide real-time access to ADR rate predictions based on the trained models. To achieve this, data indexing techniques were employed using Pandas data loader, enabling efficient retrieval and manipulation of the ADR pricing data.

Additionally, a custom indexing technique called "llama index" was introduced to enhance the efficiency of data retrieval and processing within the chatbot. The llama index leverages advanced indexing algorithms to optimize the search and retrieval of specific data points from the ADR pricing dataset, further improving the responsiveness of the chatbot.

The analysis and design section of this report provides a comprehensive overview of the approach taken to forecast ADR rates and implement Forecasting ADR prices as Optimized value in the hospitality industry. The combination of Facebook Prophet and LSTM models, along with the integration of these models into a chatbot framework using pandas data loader and the llama index, offers a robust and user-friendly solution for informed decision-making and revenue maximization in the industry.

## **5.6 Building Question-answering based chatbot for helping SME businesses in Tourism**

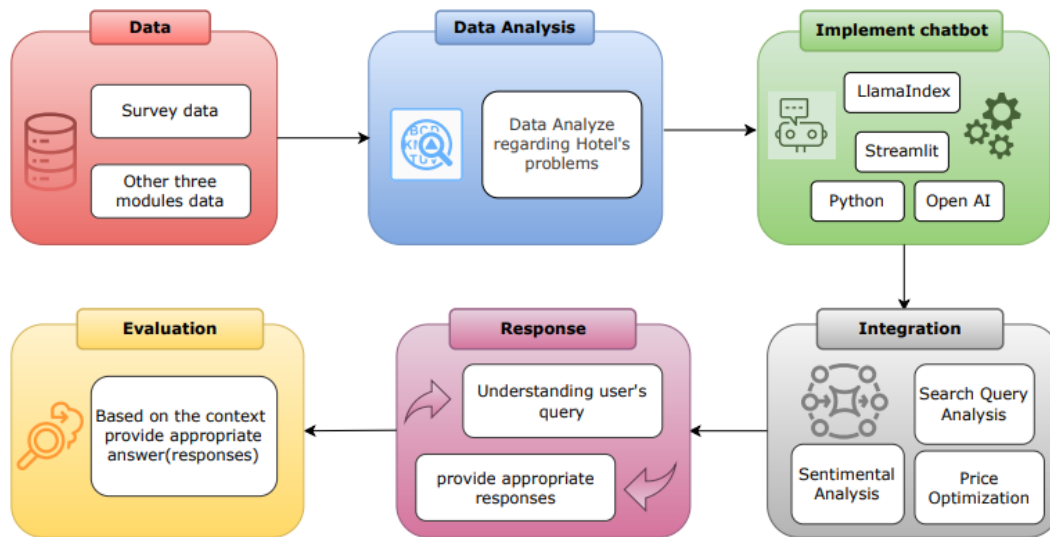


Figure 5 : Chatbot Analysis diagram

The Chatbot module plays a crucial role in facilitating interactive communication with users and providing relevant and informative responses. To enhance its capabilities, the module leverages a Large Language Model (LLM) to integrate and process data from various sources and generate context-aware outputs.

The LLM within the Chatbot module serves as a powerful tool for natural language understanding and generation. It has been trained on vast amounts of textual data, enabling it to comprehend and generate human-like responses. The LLM employs advanced techniques such as deep learning and neural networks to understand the semantics, context, and nuances of user queries.

When a user interacts with the Chatbot module, their query is received and processed. The module applies natural language processing techniques to understand the intent and meaning behind the query. It analyzes the user's input, considering factors such as keywords, sentiment, and contextual information derived from the Search Query Analysis, Google Trends, and Sentiment Analysis modules.

Once the user's query has been processed and understood, the Chatbot module leverages the LLM to generate a response. The LLM utilizes its vast knowledge base and language understanding

capabilities to provide informative and contextually appropriate answers. It takes into account the user's query, the insights obtained from the analysis modules, and other relevant information to generate a meaningful and personalized response.

The output generated by the LLM can include a variety of information. It may provide recommendations on tourist attractions, suggest popular destinations based on search interest patterns, offer sentiment-based advice on hotel choices, or provide optimized pricing details obtained from the Forecasting ADR prices as Optimized value module. The Chatbot module ensures that the response is tailored to the user's specific query and integrates insights from the analysis modules to deliver a comprehensive and valuable output.

By incorporating a Large Language Model into the Chatbot module, the system can dynamically respond to user queries and provide information that is relevant, up-to-date, and contextually appropriate. The LLM's ability to understand and generate natural language enables the Chatbot to engage in meaningful conversations with users, making the system more interactive and user-friendly.

## **Chapter 6**

### **6.1 Chapter Overview**



The main goal of this chapter is to give a thorough overview of the implementation process that has been carried out so far. It tries to highlight the methods and resources used in the implementation of each submodule inside our suggested system. This chapter also provides a thorough review of each module's current implementation status and a summary of the development's advancement.

## **6.2 Search Query Analysis**

### **6.2.1 Introduction**

Search engine data, including real-time daily, weekly, and monthly search queries entered by users on platforms like Google and Baidu, provide valuable structured Internet data that can be utilized as new sources for tourism forecasting. In my module, the primary objective is to leverage search queries specifically related to tourism from the Google search engine. The main focus of my research is to identify and select appropriate keywords that can contribute to accurately forecasting tourist arrivals based on their search interests.

### **6.2.2 Data Collection and Data pre-processing**

The data collection process involved the use of a Google Form to gather queries from individuals representing diverse countries. Subsequently, organic results from the Google search engine were extracted for tourism-related queries. This extraction encompassed not only the primary search results but also encompassed related searches and the "People also ask" section found on the Google results page.

To preprocess search query data for keyword extraction using feature extraction machine learning techniques, the following steps were performed,

```

import pandas as pd
import nltk
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import word_tokenize

# Download necessary resources for NLTK
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')

# Load the CSV file
def load_csv(file_path):
    df = pd.read_csv(file_path)
    return df['Queries'].tolist() # Replace 'TextColumn' with the actual column name

# Preprocess the text data
def preprocess_text(text):
    # Tokenization
    tokens = word_tokenize(text.lower())

    # Stopword Removal
    stop_words = set(stopwords.words('english'))
    filtered_tokens = [token for token in tokens if token not in stop_words]

    # Lemmatization
    lemmatizer = WordNetLemmatizer()
    lemmatized_tokens = [lemmatizer.lemmatize(token) for token in filtered_tokens]

    # Join the lemmatized tokens back into a single string
    processed_text = ' '.join(lemmatized_tokens)

    return processed_text

# Example usage
file_path = '/content/drive/MyDrive/Colab Notebooks/Query - Sheet1.csv' # Replace with the actual file path
data = load_csv(file_path)

preprocessed_data = [preprocess_text(text) for text in data]
print(preprocessed_data)

```

Figure 6 : Query Preprocessing

Text Cleaning: Irrelevant characters or symbols, such as punctuation marks, special characters, and numbers. This has been done to ensure the consistency and relevancy of data.

Tokenization: The search query text is split into individual words or tokens. It was done by removing whitespace and other tokenization techniques.

Stopword Removal: Common and generic words that hold no specific inherent meaning , including articles, pronouns, and prepositions were removed. These were filtered out to increase the focus on relevant words relative to the context.

**Lowercasing:** All text data were converted into lowercase to enable case insensitivity. This enabled to process words in different cases at the same, reducing the complexity and improving consistency.

**Lemmatization:** Words were reduced to their base to reduce similar words. It also enabled to group related words together and reduce dimensionality.

**Feature Extraction:** Then feature extraction techniques were implemented to transform the text data to numerical figures. Term Frequency-Inverse Document Frequency (TF-IDF), was used for this purpose. TF (Term Frequency) measures the frequency of a term within a document. It indicates how often a term appears in a specific document. The assumption is that the more frequent a term appears in a document, the more important it is to that document. IDF (Inverse Document Frequency) measures the significance of a term in the entire collection of documents. It calculates the inverse ratio of the number of documents that contain a specific term to the total number of documents in the collection. The IDF value is higher for terms that appear in fewer documents, indicating their higher importance.

**Dimensionality Reduction:** If the feature space is large, dimensionality reduction techniques like Principal Component Analysis (PCA) or t-SNE can be applied to reduce the number of features while preserving important information. This helps to improve computational efficiency and mitigate the curse of dimensionality.

**Normalization:** The feature values were normalized to a standardized range to ensure that the features had a relative influence on the model.

```

from sklearn.datasets import make_classification
from sklearn.model_selection import cross_val_score
from sklearn.tree import DecisionTreeClassifier
from sklearn.feature_extraction.text import TfidfVectorizer
from deap import creator, base, tools, algorithms
import random
import numpy

# Read CSV file
df = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/Tourism Queries - Sheet1.csv') # assuming your file name is 'data.csv'

documents = df['documents'].tolist() # Assuming the column name is 'documents'
y = df['y'].tolist() # Assuming the column name is 'y'

vectorizer = TfidfVectorizer(max_features=5000) # Limit to 5000 features for efficiency
X = vectorizer.fit_transform(documents)

creator.create("FitnessMax", base.Fitness, weights=(1.0,))
creator.create("Individual", list, fitness=creator.FitnessMax)

toolbox = base.Toolbox()
toolbox.register("attr_bool", random.randint, 0, 1)
toolbox.register("individual", tools.initRepeat, creator.Individual, toolbox.attr_bool, len(X[0].toarray()[0]))
toolbox.register("population", tools.initRepeat, list, toolbox.individual)

def evalOneMax(individual):
    mask = numpy.array(individual, dtype=bool)
    model = DecisionTreeClassifier()
    return (cross_val_score(model, X[:, mask], y, cv=5).mean(),)

toolbox.register("evaluate", evalOneMax)
toolbox.register("mate", tools.cxTwoPoint)
toolbox.register("mutate", tools.mutFlipBit, indpb=0.05)
toolbox.register("select", tools.selTournament, tournsize=3)

population = toolbox.population(n=300)

NGEN=40
for gen in range(NGEN):
    offspring = algorithms.varAnd(population, toolbox, cxpb=0.5, mutpb=0.2)
    fits = toolbox.map(toolbox.evaluate, offspring)
    for fit, ind in zip(fits, offspring):
        ind.fitness.values = fit
    population = toolbox.select(offspring, k=len(population))

top10 = tools.selBest(population, k=10)

```

Figure 7 : Genetic Feature Algorithm

```

feature_names = vectorizer.get_feature_names_out()

# for each individual in top 10
for i, individual in enumerate(top10):
    # create a mask from the binary list
    mask = numpy.array(individual, dtype=bool)
    # use the mask to get the selected features
    selected_features = numpy.array(feature_names)[mask]
    print(f"Individual {i+1}:")
    print(selected_features)

```

Figure 8 : Feature Selection

### 6.2.3 Evaluation

The provided code snippet showcases a process for evaluating the accuracy of multiple top solutions using a decision tree classifier. It begins by importing essential libraries, including `sklearn.metrics` for accuracy scoring, `sklearn.model_selection` for train-test splitting, and `sklearn.tree` for the decision tree classifier. The dataset is then split into training and testing sets using the `train_test_split` function, allocating 20% of the data for testing. A random seed is set to ensure reproducibility of results. A list named `accuracy_list` is initialized to store the accuracy values of each top solution. The code iterates through the top10 solutions, assuming they contain a list of binary masks representing feature selection.

In each iteration, the binary mask is transformed into a boolean array using the NumPy library. Features are selected based on the mask from both the training and testing sets. A decision tree classifier is created and fitted using the selected features from the training set.

Next, the classifier predicts the labels for the selected features of the test set. The accuracy of the predictions is calculated using the `accuracy_score` function, which compares the predicted labels with the true labels of the test set. The obtained accuracy score is then added to the `accuracy_list`. Finally, a loop is employed to print the accuracy of each top solution, along with their corresponding index.

To summarize, this code performs feature selection using binary masks, trains a decision tree classifier on the selected features, predicts labels for the test set, and calculates the accuracy of each top solution. The accuracy scores are stored in a list and printed to assess the performance of the feature selection.

```
[ ] from sklearn.metrics import accuracy_score
    from sklearn.model_selection import train_test_split
    from sklearn.tree import DecisionTreeClassifier

    # Choose a random seed for reproducibility
    random_seed = 42

    # Split the data into train and test sets
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=random_seed)

    # Initialize a list to store the accuracy of each top solution
    accuracy_list = []

    for individual in top10:
        mask = numpy.array(individual, dtype=bool)

        # Select the features from the train and test sets
        X_train_selected = X_train[:, mask]
        X_test_selected = X_test[:, mask]

        # Initialize and fit the classifier
        classifier = DecisionTreeClassifier(random_state=random_seed)
        classifier.fit(X_train_selected, y_train)

        # Make predictions on the test set and calculate the accuracy
        y_pred = classifier.predict(X_test_selected)
        accuracy = accuracy_score(y_test, y_pred)

        accuracy_list.append(accuracy)

    # Print the accuracy of each top solution
    for i, accuracy in enumerate(accuracy_list, 1):
        print(f"Solution {i}: {accuracy}")
```

Figure 9 : Feature Selection Evaluation

7	Risks	Other	What are best restaurants in Galle	Food	33	Easy and cheap ways to travel destinations.	Travel
3	Beautiful places	Destination	Places to visit in Galle	Destination	54	Places to visit in Galle	Destination
4	Historical places in polanaruwa	Destination	How long will it take to go from place X to place Y	Travel	55	hiking places in matala	Activity
5	Jaffna Fort	Destination	How to book tickets	Travel	56	Parachuting in Galle	Activity
6	famous places near colombo	Destination	Train schedule.	Travel	57	Hotel price in sri lanka	Accommodation
7	Place to visit in Kandy	Destination	Most pretty places in Sri Lanka	Destination	58	Places to visit in galle.	Destination
8	Transport to nuwa eliya	Travel	chrimas events in galle	Seasonal	59	Promotions in hotels in sri lanka	Accommodation
9	1.Best places to visit	Destination	About Residence facilities	Accommodation	60	Place to visit Jaffna	Destination
10	What is the best place to visit in Galle	Destination	Water park locations in Sri Lanka	Accommodation	61	Accommodations	Accommodation
11	1. What are the top places to visit in Nuwara Eliya	Destination	Top hiking areas in the hill country	Travel	62	Best indian dish restaurants	Food/Dining
12	Top nature place in Sri Lanka	Destination	best hotels for meetings	Travel	63	how much rent for room	Accommodation
13	Resorts in galle	Accommodation	Tourist attractions in Ella	Destination	64	Nuwaraeliya train schedule	Travel
14	Cabs for family rides	Travel	Galle	Destination	65	Hotels/ Resorts in Haputale	Accommodation
15	Price of the tickets	Activity	cost of the trip	Other	66	best chinese restaurants	Food/Dining
16	Best tourist places near me	Destination	Google map reviews. Trip advisor reviews	Other	67	Most famous tourist attractions of Sri lanka	Destination
17	Attraction near me	Destination	What are the best places to visit in galle	Other	68	What is the climate	Events/Seasonal
18	Hotels in anuradhapura	Accommodation	Bungee jumping in sri lanka	Activity	69	Places to visit in Colombo	Destination
19	Best places in Sri Lanka	Destination	Best places to visit in December	Destination	70	Colombo 7000	Destination
20	Places to visit in jaffna	Destination	best pizza huts	Food	71	Best beautiful natural trip places	Destination
21	Bird watching place near me	Activity	Places to visit in Kandy	Destination	72	Places nearest moratuwa	Destination
22	Places to visit in Kaluthara	Destination	Best hotels in Sri Lanka	Accommodation	73	Universal studios japan.	Activity
23	What time trains depature from fort	Travel	Sweeping	Activity	74	Natural waterfalls in Sri Lanka	Destination
24	What are best restaurants in Galle	Food/Dining	What is the best travelling mode	Activity	75	Where can I watch elephants in Sri Lanka	Activity
25	Places to visit in Colombo	Destination	How far from place 1 to place 2	Travel	76	Places to visit in Nuwara eliya	Destination
26	How long will it take to go from place X to place Y	Travel	galle family hotels	Accommodation	77	What are the travelling destinations in bandaravella	Destination
27	How to book tickets	Travel	Places to visit in Colombo	Destination	78	best tamil dishes	Food/Dining
28	Train schedule.	Travel	galle adventures	Activity	79	Attractions around badulla	Destination
29	Most pretty places in Sri Lanka	Destination	galle parachuting and gliding	Activity	80	Places to Visit in Galle	Destination
30	chrimas events	Events/Seasonal	Best places to visit in Sri Lanka	Destination	81	tourist attractions in Sri Lanka	Destination
31	About Residence facilities	Accommodation	Easy and cheap ways to travel destinations.	Destination	82	best elephant rides in sri lanka	Activity
32	Water park locations in Sri Lanka	Destination	Places to visit in Galle	Destination	83	Things to do in Sri Lanka	Activity
33	Top hiking areas in the hill country	Activity	hiking places in Galle	Activity	84	What are the beautiful waterfall in srilanka	Destination
34	best hotels for meetings	Accommodation	Hotel price in sri lanka	Accommodation	85	hotels for day outs and vacations	Accommodation

Figure 10 : Data categorization of search query

## **6.3 Aspect based Sentiment analysis**

### **6.3.1 Introduction**

Analyzing and categorizing the sentiment found in text data relating to hotel experiences is the process of sentiment analysis for hotel reviews. It uses methods from machine learning and natural language processing to extract information from customer reviews and assess if the sentiment is favorable, negative, or neutral. Businesses in the hospitality sector may learn a lot about consumer happiness, pinpoint areas for development, and make data-driven decisions to improve the overall visitor experience by automatically categorizing sentiments in hotel reviews. Sentiment analysis gives hoteliers the ability to keep an eye on their online reputation, efficiently respond to consumer comments, and take proactive steps to provide great service, increasing visitor happiness and loyalty.

### **6.3.2 Data Collection and data pre-processing**

To improve the small medium business further, collecting the reviews of the small medium businesses (hotels) of Galle district. from the social media like TripAdvisor.com, booking.com and SME's own websites using the parse hub desktop application to scrape the review's details for the small medium business.

After collecting the review, I need to do the preprocessing before sentiment analysis. Because unwanted noises need to be removed from the reviews which are not the part of the step of the analysis. We must follow pre-processing steps to prepare the data for our module.

- Tokenization

In the review's all the review sentences are tokenized according to split the complex text to small chunks of words and define the full stop, exclamation mark, question marks and the process of the tokenization is stop the comma(,) , semicolon (;), and full stop(.) from the reviews

- Stop Word Removal

After completing the tokenization step this system removes the stop words from the review sentence if any stop words are available in the review sentences. Examples of the stop words are the, is, was etc. these stop words are not the part of sentiment analysis. So that need to remove that data. Because these are generating unnecessary noises in the analysis.

- Stemming

After completing the stop word removal, do the stemming for the morphological forms of a word. In here two words do not have the same meaning, then that time separates those words. If the same meaning of the words is available, then map those words in the morphological forms. These two approaches are enough to continue the text mining or language processing applications. For example, in the English language “performing” word is converted to perform for the stemming process. It happens in the feature selection.

- Text Normalization

Reduce the impact of case differences by using text normalization techniques like lowercasing to convert all text to lowercase. By enlarging or normalizing contractions, abbreviations, and slang to their full forms, based on that can manage them.

**Normalization of emoticons:** Emoticons are crucial to sentiment analysis, therefore keeping them in place is preferable to eliminating them. In our method, we made use of a specialized dictionary that included definitions for frequently used emotions in English. Based on this vocabulary, each emoticon that was used in the text was substituted with its appropriate meaning. Examples include replacing emoticons like ☺ and ;) with the word “happy.” This normalization process aids in accurately capturing the intended feeling that the emoticons are intended to convey.

### **Auto Categorizing the aspect**

This module’s main objective is automatically categorized the aspects from the review text. for that did the key word extraction using the TFIDF to extract all the reviews into the words with their frequency. Based on these high frequency words display in the reviews manually select the seven aspects for developing the rules based approach with help of existing research paper. Here



increase the amount of aspects and set the predefined keywords to generate the rules based approach using key-value pair. After that did the sentiment prediction based on aspect using Vader sentiment to predict the sentiment polarity and classify the reviews as positive, negative and neutral. based on this train and test the dataset using different algorithms to predict the accuracy of aspect and sentiment of the module. This evaluation offers a thorough grasp of the feelings felt by users in regard to particular areas of interest.

## Model Performance Evaluation

To make sure it is precise and dependable, it is crucial to test and assess its performance. This may entail utilizing many evaluation criteria, such as recall, precision, accuracy and F1 score.

A	
1	reviewcontent
2	We had wonderful experience and staff was amazing. Rooms were very clean and they went beyond to help make our stay enjoyable and food was delicious. Highly recommend this hotel for anyone visiting down south.
3	If you like to spend your vacation in style of luxury this is the best option you could ever have. Friendly and pleasant staff. Spacious rooms with all the amenities need for the day with high technology bath wares you will surprise. We stayed in the suite and which is easily fit in to a 4-6 people. Delicious buffet with varieties. Special thanks for the front office team, Room service, Restaurant staff and all who help to made our stay in this beautiful hotel memorable.
4	You are greeted with a choice of welcome drinks and cold towel. Then the checking in was less than 10 minutes. They called prior and asked when we would arrive to make things smooth. We initially stayed in a deluxe room with sea view. We found this quite pleasant and these rooms are quite closer to the ocean when proximity is concerned. But you do not see the galle fort from this room. Bathrooms were spacious and well planned out included a bathtub, TOTO wash let. The room was to us very comfortable welcoming. Since we were happy with the stay we extended our stay by a day and asked for a room with fort view and the hotel was kind enough to upgrade us to a signature room free. This room is much more spacious with a pool ocean and Galle fort view. It also has a jacuzzi. Regarding Food. We used the ala carte menu and fixed menu for breakfast dinner and lunch. You also had the option of ordering food and beverages out of the AI came menu throughout the day. There wasn't a buffet for breakfast and dinner during our stay (weekdays). There is a small beach area right in front of the hotel but not suitable for bathing/swimming. Nice grove like area to watch the sunset. Pool has a pool bar with choices of hard liquor to cocktails and mock tails. Service wise we were happy. The staff were great at Le Grand and we had a really good breakfast. The rooms were spacious and we had a lovely view of the ocean. Kudos for the staff for making our stay a memorable one. Special thanks for Praveen and Darshana for giving us a warm welcome. The rooms and the food was really good. It has a very nice view of the galle fort. The staff was very friendly. The pool was quite big as well. The breakfast was a set menu since it was a weekday. Loved the stay. The room has a fantastic design and the restaurant and service are professional but the micro environment around the hotel was rather disappointing for me. there is no direct access to the park which is very close to the hotel ( closed gate) and the only way to sightseeing in galle fort is to pass by dirty street and a very busy bus station. The shop and spa were closed. Friendly staff. Great view of galle fortress Super Baby friendly hotel. Kitchen staff will help you even with your baby food preparation. Very calm and quiet environment and has a great sea view to the top floor. The hotel is beautiful and the staff and food are excellent. The room has the most amazing view of the Galle fortress and the sea and can be enjoyed in the jacuzzi. I would definitely come back again.

Figure 11 : Data collection of sentiment analysis

## Preprocessing

```
[ ] def remove_line_breaks(text):
    text = text.replace('\n', ' ').replace('\n', ' ')
    return text

#remove punctuation
def remove_punctuation(text):
    re_replacements = re.compile("__[A-Z]+__") # such as __NAME__, __LINK__
    re_punctuation = re.compile("[%s]" % re.escape(string.punctuation))
    '''Escape all the characters in pattern except ASCII letters and numbers'''
    tokens = word_tokenize(text)
    tokens_zero_punctuation = []
    for token in tokens:
        if not re_replacements.match(token):
            token = re_punctuation.sub(" ", token)
            tokens_zero_punctuation.append(token)
    return ' '.join(tokens_zero_punctuation)
```

Figure 12 : Preprocessing of sentiment analysis

## Keyword Extraction

```
[ ] all_reviews = ' '.join(reviews)

# Generate the word cloud
wordcloud = WordCloud(width=800, height=800, background_color='white', min_font_size=10).generate(all_reviews)

# Get the words and their frequencies from the word cloud
word_freq = wordcloud.process_text(all_reviews)

# Create a dictionary to store the word frequencies
freq_dict = {}

# Append the words and their frequencies to the dictionary
for word, freq in word_freq.items():
    freq_dict[word] = freq

# Plot the word cloud
plt.figure(figsize=(8, 8), facecolor=None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad=0)
plt.show()
```

Figure 133 : Keyword Extraction



```
[ ] import nltk
    from nltk.sentiment.vader import SentimentIntensityAnalyzer
    nltk.download('vader_lexicon')
    sid = SentimentIntensityAnalyzer()

[nltk_data] Downloading package vader_lexicon to /root/nltk_data...
[nltk_data] Package vader_lexicon is already up-to-date!

[ ] reviews = dataset['clean_text']

[ ] sentiment_scores = []

    for review in reviews:
        ss = sid.polarity_scores(review)
        sentiment_scores.append(ss)

[ ] for i, score in enumerate(sentiment_scores):
    print("Review", i+1, "sentiment scores:", score)
```

*Figure 166 : Sentiment Score Prediction*

```
[ ] from sklearn.metrics import accuracy_score, classification_report, precision_score, recall_score, f1_score
    def train_aspect_sentiment_topic_model( num_topics):

        # Extract the text, aspect, and sentiment columns
        documents = dataset['clean_text'].tolist()
        aspects = dataset['aspects'].tolist()
        sentiments = dataset['Sentiment'].tolist()

        # Preprocess the documents (tokenization, lowercasing, stop word removal, etc.)
        # You can use your own preprocessing techniques or libraries like NLTK or spaCy

        # Create a document-term matrix using CountVectorizer
        vectorizer = CountVectorizer()
        dtm = vectorizer.fit_transform(documents)

        # Split the dataset into training and testing sets
        X_train, X_test, y_aspect_train, y_aspect_test, y_sentiment_train, y_sentiment_test = train_test_split(
            dtm, aspects, sentiments, test_size=0.2, random_state=42
        )

        # Train a classifier
        aspect_classifier = LogisticRegression()
        aspect_classifier.fit(X_train, y_aspect_train)

        # Train a classifier for sentiment prediction
        sentiment_classifier = LogisticRegression()
        sentiment_classifier.fit(X_train, y_sentiment_train)
```

*Figure 177 ;Logistic Regression model Train and test*

```

# SVM classifier
svm_model = SVC()
svm_model.fit( X_train, y_aspect_train)

# Evaluate the aspect classifier (SVM)
aspect_predictions_svm = svm_model.predict(X_test)
aspect_accuracy_svm = accuracy_score(y_aspect_test, aspect_predictions_svm)
aspect_precision_svm = precision_score(y_aspect_test, aspect_predictions_svm, average='weighted')
aspect_recall_svm = recall_score(y_aspect_test, aspect_predictions_svm, average='weighted')
aspect_f1_svm = f1_score(y_aspect_test, aspect_predictions_svm, average='weighted')

print(f"Aspect Accuracy ( SVM): {aspect_accuracy_svm}")
print(f"Aspect Precision (SVM): {aspect_precision_svm}")
print(f"Aspect Recall (SVM): {aspect_recall_svm}")
print(f"Aspect F1 Score (SVM): {aspect_f1_svm}")

# SVM classifier for sentiment prediction
svm_model_sentiment = SVC()
svm_model_sentiment.fit( X_train, y_sentiment_train)

# Evaluate the sentiment classifier (SVM)
sentiment_predictions_svm = svm_model_sentiment.predict(X_test)
sentiment_accuracy_svm = accuracy_score(y_sentiment_test, sentiment_predictions_svm)
sentiment_precision_svm = precision_score(y_sentiment_test, sentiment_predictions_svm, average='weighted')
sentiment_recall_svm = recall_score(y_sentiment_test, sentiment_predictions_svm, average='weighted')
sentiment_f1_svm = f1_score(y_sentiment_test, sentiment_predictions_svm, average='weighted')

print(f"Sentiment Accuracy (SVM): {sentiment_accuracy_svm}")
print(f"Sentiment Precision (SVM): {sentiment_precision_svm}")

```

Figure 188 ;SVM model Train and test

```

from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics import accuracy_score
from sklearn.ensemble import GradientBoostingClassifier
from joblib import dump

def train_aspect_sentiment_topic_model(num_topics):
    # Extract the text, aspect, and sentiment columns
    documents = dataset['clean_text'].tolist()
    aspects = dataset['aspects'].tolist()
    sentiments = dataset['Sentiment'].tolist()

    # Preprocess the documents (tokenization, lowercasing, stop word removal, etc.)
    # You can use your own preprocessing techniques or libraries like NLTK or spaCy

    # Create a document-term matrix using CountVectorizer
    vectorizer = CountVectorizer()
    dtm = vectorizer.fit_transform(documents)

    # Split the dataset into training and testing sets
    X_train, X_test, y_aspect_train, y_aspect_test, y_sentiment_train, y_sentiment_test = train_test_split(
        dtm, aspects, sentiments, test_size=0.2, random_state=42
    )

    # Train a classifier for aspect prediction using Gradient Boosting
    aspect_classifier = GradientBoostingClassifier()
    aspect_classifier.fit(X_train, y_aspect_train)

```

Figure 199 ;Gradient Boosting Classifier model Train and test

```

from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics import accuracy_score
from sklearn.ensemble import RandomForestClassifier
from joblib import dump

def train_aspect_sentiment_topic_model(num_topics):
    # Extract the text, aspect, and sentiment columns
    documents = dataset['clean_text'].tolist()
    aspects = dataset['aspects'].tolist()
    sentiments = dataset['Sentiment'].tolist()

    # Preprocess the documents (tokenization, lowercasing, stop word removal, etc.)
    # You can use your own preprocessing techniques or libraries like NLTK or spaCy

    # Create a document-term matrix using CountVectorizer
    vectorizer = CountVectorizer()
    dtm = vectorizer.fit_transform(documents)

    # Split the dataset into training and testing sets
    X_train, X_test, y_aspect_train, y_aspect_test, y_sentiment_train, y_sentiment_test = train_test_split(
        dtm, aspects, sentiments, test_size=0.2, random_state=42
    )

    # Train a classifier for aspect prediction using Random Forests
    aspect_classifier = RandomForestClassifier()
    aspect_classifier.fit(X_train, y_aspect_train)

```

*Figure 20 ;Random Forest model Train and test*

```

from sklearn.model_selection import GridSearchCV

def train_aspect_sentiment_topic_model(num_topics):
    # Extract the text, aspect, and sentiment columns
    documents = dataset['clean_text'].tolist()
    aspects = dataset['aspects'].tolist()
    sentiments = dataset['Sentiment'].tolist()

    # Preprocess the documents (tokenization, lowercasing, stop word removal, etc.)
    # You can use your own preprocessing techniques or libraries like NLTK or spaCy

    # Create a document-term matrix using CountVectorizer
    vectorizer = CountVectorizer()
    dtm = vectorizer.fit_transform(documents)

    # Split the dataset into training and testing sets
    X_train, X_test, y_aspect_train, y_aspect_test, y_sentiment_train, y_sentiment_test = train_test_split(
        dtm, aspects, sentiments, test_size=0.2, random_state=42
    )

    # Define the parameter grid for grid search
    param_grid = {
        'n_estimators': [100, 200, 300],
        'max_depth': [None, 5, 10],

```

*Figure 21 ;Hyperparameter Tuning model Train and test*

```

C:\Users\Sinhuja\PycharmProjects\pythonProject\venv\Scripts\python.exe D:\MyRe\Predict_Model.py
Enter a document: Enjoyed our stay! Convenient location, with good variety of restaurants. Supermarket and shopping mall next door was very useful.
Predicted Aspect: accommodation ,room,service
Predicted Sentiment: positive
Enter a document:

```

Figure 22 : Predict the aspect and sentiment

	Accuracy	F1-Score		Precision	Recall
Logistic Regression[Aspect]	0.7098765432098766	0.6789585814277173		0.7006148338277681	0.7098765432098766
Logistic Regression[Senti]	0.9814814814814815	0.9776008545724634		0.9818287037037038	0.9814814814814815
SVM(Asspect)	0.46296296296296297	0.37751990807546365		0.3751674332284354	0.46296296296296297
SVM(Sentiment)	0.9691358024691358	0.9539455861294941		0.9392242036274958	0.9691358024691358
	0.863	0.859		0.875	0.938
Gradient Aspect)	0.7222222222222222	0.6999945678148033		0.73524683027607	0.7222222222222222
Gradient(Sentiment)	0.9753086419753086	0.9723003594311611		0.9714263529777158	0.9753086419753086
Random Forests (Aspect)	0.7160493827160493	0.6753396239414926		0.6977257864316434	0.7160493827160493
Random Forests (senti)	0.9814814814814815	0.9776008545724634		0.9818287037037038	0.9814814814814815
	0.859	0.854		0.867	0.942
HT(Asspect)	0.7037037037037037	0.6606581959484883		0.6744537526944935	0.7037037037037037
HT(Senti)	0.9753086419753086	0.9753086419753086		0.9672334808603151	0.9759220918641208

Figure 23 : Evaluation of sentiment analysis

Using Streamlit web-based interface to provide the sentiment analysis based on auto categorized aspect using rule based approach to understanding the customer preference in further.

## **6.4 Forecasting ADR prices as Optimed value**

### **6.4.1 Data Collection**

The initial step in our Forecasting ADR prices as Optimed value process involved collecting data from various sources. The data used for training our models was sourced from Antonio, Almeida, and Nunes in 2019 [42]. Specifically, after the training process has been done, For the implementation/validation data, we focused on hotel data for the Galle region, which was obtained by scraping websites such as Booking.com. The collected information included hotel names, room types, service types (full board, half board), and corresponding room prices for different dates for single hotels.

### **6.4.2 Data Preprocessing**

Once the data was collected, we performed preprocessing steps to prepare it for our forecasting models. During the preprocessing stage, we utilized the Dickey-Fuller test [46] to assess the presence of a unit root in the time series data. This test helps determine the stationarity of the data, which is essential for accurate forecasting.

Furthermore, data cleaning procedures were carried out to address missing or inconsistent entries. Any data points with missing price information were removed from the dataset since they wouldn't contribute to the forecasting models.

### **6.4.3 Model Implementation**

In our analysis, we implemented two models for forecasting ADR rates: Facebook Prophet and LSTM. After the data preprocessing stage, we compared the performance of these models to determine the most effective one for our Forecasting ADR prices as Optimed value process. Upon evaluation, we found that Facebook Prophet demonstrated superior performance in terms of accuracy and suitability for the given task.



Using the Facebook Prophet library, we developed a forecasting model to generate future ADR rate predictions. The Prophet model incorporates various time series components, including trend, seasonality, and holiday effects, to capture and model the underlying patterns in the data. This approach proved effective in accurately forecasting ADR rates, making it the preferred choice for implementation in our Forecasting ADR prices as Optimized value process.

The model was trained on the preprocessed data, and future predictions were generated based on the trained model. The generated forecasts provided valuable insights into the expected ADR rates, enabling informed decision-making for revenue maximization in the hospitality industry.

```
import pandas as pd
from prophet import Prophet

# Create a DataFrame with the 'ds' and 'y' columns expected by Prophet
df = pd.DataFrame({'ds': filtered_train_df.index, 'y': filtered_train_df.values})

# Convert the 'ds' column to datetime format
df['ds'] = pd.to_datetime(df['ds'].astype(str) + '-1', format='%Y%m-%w')

# Initialize and fit the Prophet model
model = Prophet(yearly_seasonality=True, weekly_seasonality=True, daily_seasonality=True)
model.fit(df)

# Generate future dates for forecasting
future_dates = model.make_future_dataframe(periods=104, freq='W')

# Perform forecasting
forecast = model.predict(future_dates)

# Print the forecasted values
print(forecast[['ds', 'yhat', 'yhat_lower', 'yhat_upper']].tail())

# Plot the forecast
model.plot(forecast, xlabel='FullDate', ylabel='ADR')
model.plot_components(forecast)

# Show the plot
plt.show()
```

*Figure 24 : Implementation of Forecasting ADR prices as Optimized value*

This code initializes and trains a Prophet model on our preprocessed data, and then uses the model to generate price forecasts for the next year.

#### 6.4.4 Result Interpretation

In the code, a DataFrame is created using the 'ds' (date) and 'y' (ADR rate) columns from the preprocessed and filtered training dataset. This DataFrame is used as input to initialize and fit the Prophet model, a time series forecasting model that incorporates various seasonality components.

The Prophet model is configured with different parameters to capture seasonality patterns at various levels. The parameters 'yearly\_seasonality', 'weekly\_seasonality', and 'daily\_seasonality' are set to True, indicating that the model should consider yearly, weekly, and daily seasonality effects in the data.

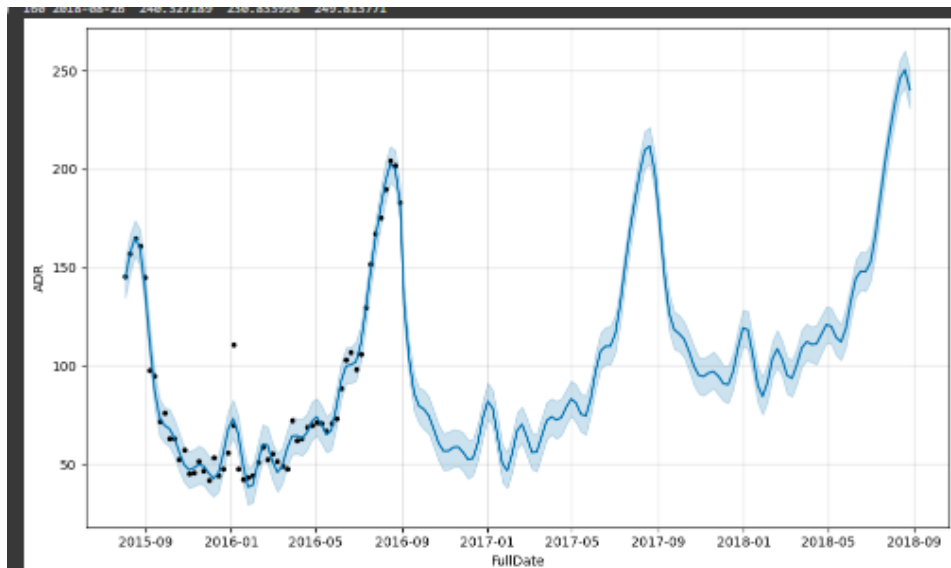
Once the model is fitted, future dates are generated using the 'make\_future\_dataframe' method. This creates a DataFrame with dates extending beyond the original dataset, allowing for forecasting into the future.

The forecasting is then performed using the 'predict' method of the Prophet model, generating predictions for the ADR rates. The forecasted values, along with their lower and upper bounds, are printed using the 'tail' method to display the most recent predictions.

To visualize the forecasted results, both the overall forecast and the components of seasonality are plotted using the 'plot' and 'plot\_components' methods of the Prophet model, respectively. The resulting plots provide insights into the predicted ADR trends and the contributions of different seasonality factors.

By examining the forecasted values, lower and upper bounds, as well as the plots, stakeholders can gain an understanding of the projected ADR rates and how they are influenced by yearly,

weekly, and daily seasonality patterns. This information can assist in making informed decisions related to pricing strategies and revenue optimization in the hospitality industry.



*Figure 25 : Result Interpretation*

The plot shows the original data (black dots), the forecasted values (blue line), and the uncertainty intervals of the forecast (shaded blue area). By examining this plot, we can see the overall trend in hotel prices and how they are expected to change over the next year.

The results from the Prophet model can assist hotel managers and owners in making data-driven decisions regarding their pricing strategies. By forecasting future price trends, they can optimize their prices to maximize revenue, improve occupancy rates, and stay competitive in the market.

#### **6.4.5 Novel Approach with GPT-3**

In parallel to the forecasting model, we also indexed the data to interface with GPT-3 via a chatbot. This method enables users to engage in a natural language conversation to extract valuable insights. Users can ask questions related to the forecasted data, providing a more interactive and user-friendly way of exploring the data and the results of the forecasting model.

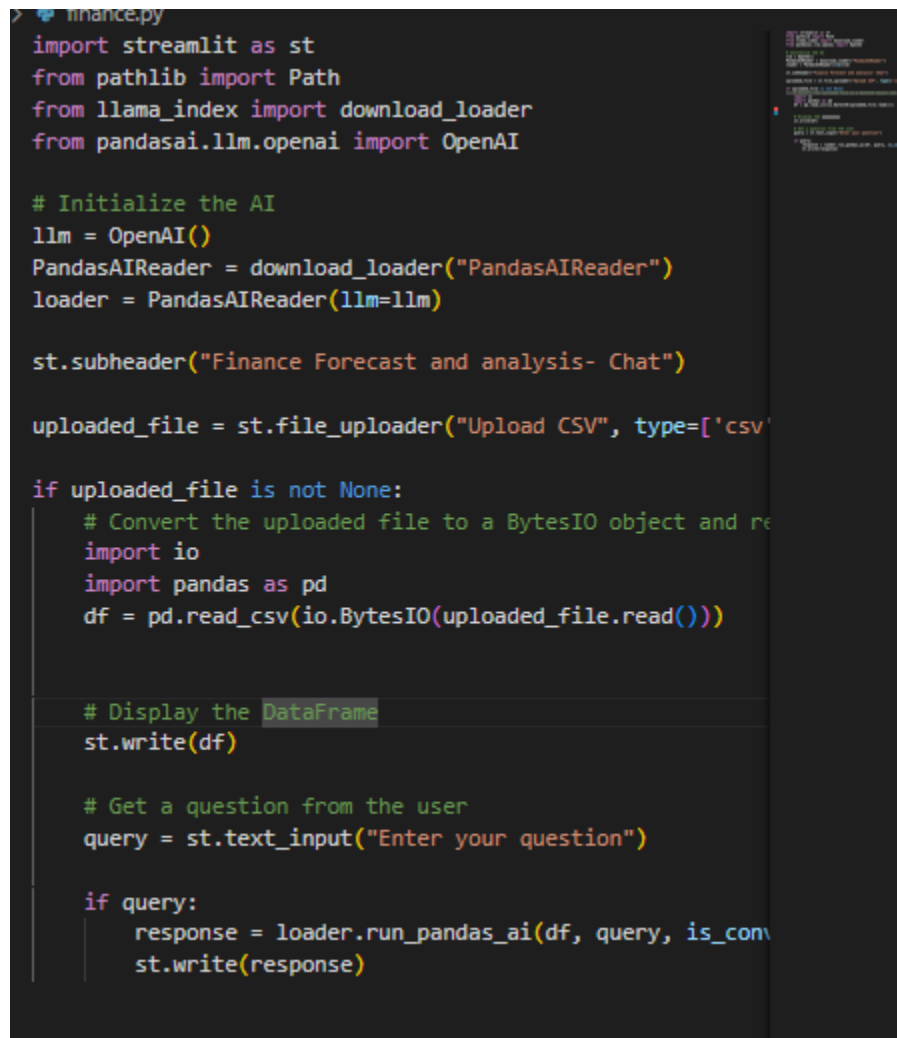
Together, these techniques provide a comprehensive, innovative approach to Forecasting ADR prices as Optimized value in the hotel and tourism industry. By combining traditional forecasting

with advanced AI technologies, we can extract valuable insights and make more informed decisions.

#### 6.4.6 Implementation of Novel Approach with GPT-3 and PandasAI

The second approach for our Forecasting ADR prices as Optimized value process harnesses the power of GPT-3, OpenAI's advanced language model, combined with PandasAI, an interface designed to facilitate interactions between data and the AI model. The implementation employs a Streamlit application to create an interactive user interface where users can upload their data and ask questions related to it in natural language.

The corresponding Python code for this process is as follows:



```
finance.py
import streamlit as st
from pathlib import Path
from llama_index import download_loader
from pandasai.llm.openai import OpenAI

# Initialize the AI
llm = OpenAI()
PandasAIReader = download_loader("PandasAIReader")
loader = PandasAIReader(llm=llm)

st.subheader("Finance Forecast and analysis- Chat")

uploaded_file = st.file_uploader("Upload CSV", type=['csv'])

if uploaded_file is not None:
    # Convert the uploaded file to a BytesIO object and read it
    import io
    import pandas as pd
    df = pd.read_csv(io.BytesIO(uploaded_file.read()))

    # Display the DataFrame
    st.write(df)

    # Get a question from the user
    query = st.text_input("Enter your question")

    if query:
        response = loader.run_pandas_ai(df, query, is_converted=True)
        st.write(response)
```

Figure 206 : Chatpgpt 3 based price optimization

This code sets up a web-based user interface with Streamlit. Users are invited to upload their own CSV files containing relevant hotel data. The CSV files are read into pandas DataFrames, which are then displayed on the webpage for reference.

The PandasAIReader from the llama\_index package is utilized to handle the communication between the pandas DataFrame and the GPT-3 model. By indexing the DataFrame into an AI-understandable format, it enables the AI model to process the data and generate responses to queries related to it.

Users input their questions into a text box. These questions are then fed into the AI model, which generates a corresponding response based on the data. The model's responses are then displayed on the webpage.

This approach allows users to extract valuable insights from their data in a conversational manner. For instance, users could ask questions about notable trends in the data, comparisons between different data points, or any other queries that could help them understand the data better and make informed decisions.

In terms of Forecasting ADR prices as Optimized value for the hotel industry, this approach can complement the Prophet forecasting model by providing qualitative insights and context about the forecasted trends. For example, users could ask the model why certain periods have higher forecasted prices and get a response based on the patterns in the data.

By using this novel approach, we can provide a more interactive and user-friendly way of exploring data, enabling hotel managers and owners to leverage AI technology to optimize their pricing strategies.

#### **6.4.7 Evaluation**

To assess the effectiveness of our price forecasting model, it's crucial to conduct a comprehensive evaluation. This involves applying a cross-validation process, followed by the computation of key performance metrics.

The cross-validation process in the Prophet model involves making repeated forecasts on the historical data. For each of these forecasts, a certain amount of historical data is held back as a validation set, onto which the forecast is made. This is carried out repeatedly on different segments of time-series data to validate the performance of the model.

The performance metrics computed during this evaluation process include Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and the coverage of the confidence interval. These metrics provide quantitative measures of the forecast's accuracy.

Mean Absolute Error (MAE): This is one of the simplest error metrics to understand. It calculates the average of the absolute differences between the predicted and actual values. Its formula is as follows:

$$\text{MAE} = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j|$$

*Figure 216 : MAE Formula*

where:

n is the total number of observations  $y_i$  is the actual value for the i-th observation  $\hat{y}_i$  is the predicted value for the i-th observation This formula sums the absolute differences between each predicted and actual value, then divides by the number of observations to get the average.

Root Mean Squared Error (RMSE): This is another common error metric, especially for regression problems. It calculates the square root of the average of the squared differences between the predicted and actual values. Its formula is:

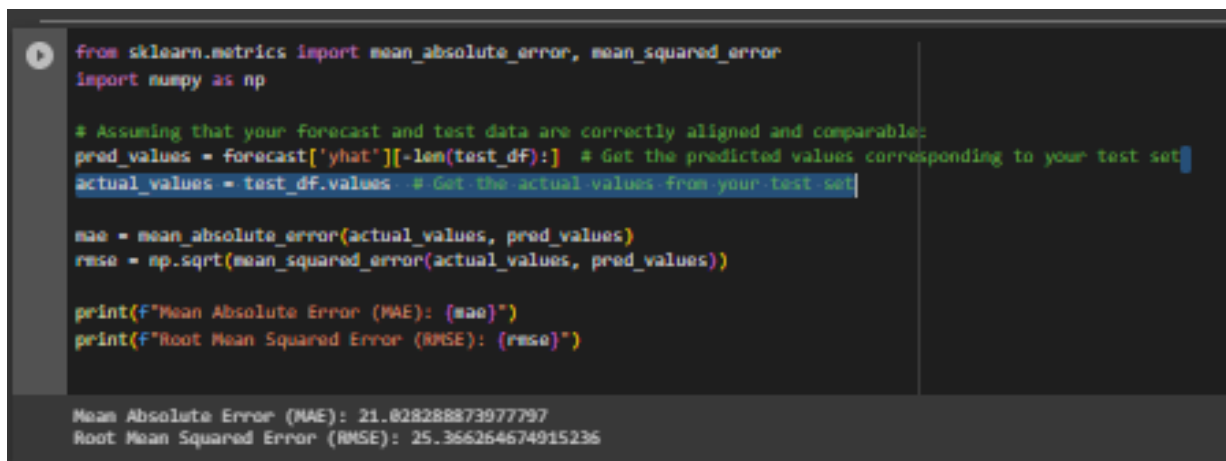
$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2}$$

Figure 226.1 : RMSEFormula

where:

$n$  is the total number of observations  $y_i$  is the actual value for the  $i$ -th observation  $\hat{y}_i$  is the predicted value for the  $i$ -th observation This formula first squares the differences between each predicted and actual value, then takes the average of these squared differences, and finally takes the square root of that average.

Figure 237 : Forecasting ADR prices as Optimed value evaluation



```
from sklearn.metrics import mean_absolute_error, mean_squared_error
import numpy as np

# Assuming that your forecast and test data are correctly aligned and comparable:
pred_values = forecast['yhat'][-len(test_df):] # Get the predicted values corresponding to your test set
actual_values = test_df.values # Get the actual values from your test set

mae = mean_absolute_error(actual_values, pred_values)
rmse = np.sqrt(mean_squared_error(actual_values, pred_values))

print(f"Mean Absolute Error (MAE): {mae}")
print(f"Root Mean Squared Error (RMSE): {rmse}")

Mean Absolute Error (MAE): 21.028288873977797
Root Mean Squared Error (RMSE): 25.366264674915236
```

Figure 248 : Performance evaluation metrics

## 6.5 Implement Chatbot

The Chatbot module is different compared to previously implemented chatbots. That is, an innovative technology has been implemented. Earlier chatbots had some problems. Based on massive dataset, limitations etc. are few. This Chatbot has been implemented based on these. Also built with the latest framework called LlamaIndex. It easily understands the user's natural language and sends responses quickly. It also reduces the implementation time ofr generationg question answer.

Chatbot is one of the best solutions for SME's marketing promotion in Galle District. Marketers can realize effective ideas based on chatbot responses. SME's overall information as summarized

and provided by chatbots. There is a lot of information in google. The information level is increasing day by day and the accuracy of information also decreases. However it's impossible to find a particular information on time. When marketers find their problems and solutions through the internet, they can't find particular needed details on time. Using chatbot which can help to avoid unnecessary information and they can clarify their personalized doubts as well.

For example, If a user search "how to promote SME's business in the tourism industry", There is a lot of information that is visible like what SME is, SME types, examples, characteristics like that. Users cannot find promoted strategies on time. In this situation, if a user asks the same question to the chatbot, the chatbot summarizes all the data and provides the needed information only. So users can access on time. Implement the chatbot, its responses could be trustable than other chatbots because the chatbots are managed by marketers. So responses will be reliable and accurate.

As per our research, a lot of chatbots have been created on behalf of customer based. I found out that a chatbot needs marketers to promote their businesses. They can use the chatbot without any legal problems or restrictions.

### **5.6.1 Data Collection**

Generally there are many types of small and medium businesses. They have been analyzed and the main category have been selected. We focused on specifically hotel businesses in Galle District. Galle of Sri Lanka in there are most highlighted tourist destinations specifically hotels. In this situation, SME's hotels need to promote their businesses in tourism.

Then a survey form was created and sample data was collected to gather information related to the classified SME businesses. For that, data was collected from hotel businesses in Galle district by sending information (through social networking sites and acquaintances). This was done with the purpose of considering what kind of questions businesses would ask to improve their businesses



related to tourism and what their expectations, what challenges they are facing would be. Also, the importance of what questions are asking by the hotels? What are the challenges they are facing? Because there are a lot of problems in research papers. But those challenges are Professional. Based on that I validate the challenges in 5 categories.

### **5.6.2 Data Analysis**

The purpose of this analysis is, need to understand what kind of data need to provide the chatbot as input data. Actually what problems are Galle district hotels facing? For our clarification did the analysis. For that, created survey form collected from some of the Galle district hotels. Based on the responses, could able to find their challenges in promoting businesses. Also researched in research papers, what challenges faced in SME businesses.

Based on the survey data and research paper, I catergorized the challeneges in 6 catergories as resources, finance, facilities, customer satisfaction, competition and other. I validate those challenegs and analyzed those factors are effect in those 3 modules data. So I finalized most of the factors marketers need to understand in sentimental analysis, search query analysis and pricing data. So using the chatbot's input from those three modules.

### **5.6.3 Implement Chatbot**

Finally a chatbot is implemented using a pre-processed document of the collected information. For this, a framework called LlamaIndex is created. It's based on the text-davinci-003 model. It is used as input to the preprocessing document and other modules' output data. That is, if the user asks the question, "What are the most popular dishes in a particular restaurant?", the chatbot will use the output data related to the particular restaurant in the review analysis module to provide the user with the relevant output. Thus, when asked questions related to other modules, they can be analyzed and answered.

It uses data from the survey form as a model of what questions the user will ask and as a key file to easily understand the data. Once the user asks a question it understands the user's language and takes it into analysis and processing then gives an answer as appropriate.

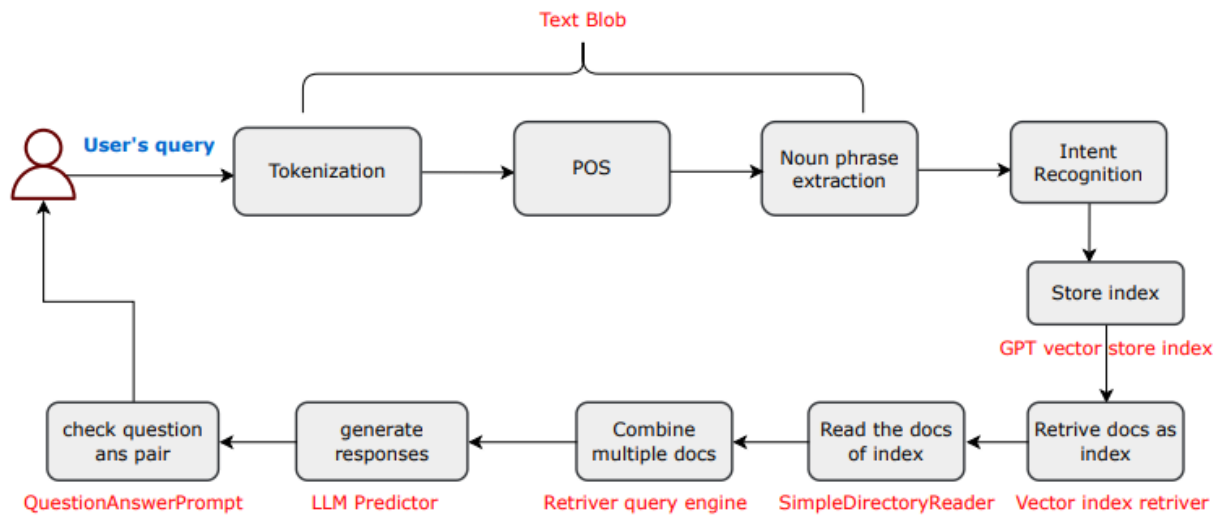


Figure 29 : Chatbot flow



B	C
Challenges	
There are lot of hotels and cafes available in the same city	Competition
Sometimes there is a shortage of food when there are many people coming.	Resources
We cannot predict the exact amount	Resources
Operational issues as staff shortages	
There is often intense competition in the restaurant and cafe industry, which can make it difficult for businesses to attract customers.	Competition
We don't have more types of rooms, because same type of hotels here	Facilities
Sometimes there is a shortage of furniture when there are many people coming.	Facilities
We cannot predict the exact amount	
not reaching target audience, competitors having better promos, not knowing the best platforms to promote	Customer satisfaction
There are lack of facilities like rooms, furnitures	Facilities
Budgeting for marketing campaigns	
Increased competition	Competition
Customer Retention for lack of intention	Customer satisfaction
Customer dissatisfaction for cleanless services	Customer satisfaction
Competition with prices	Competition
Unable to follow proper pricing policy like problem in decide the price	Finance
A lack of skilled staffs	Resources
Change in marketing trends and dynamics	
Data security challenge	Facilities
Changing consumer behaviors	
Ongoing labor woes	Resources
Not getting the expected reach	Customer satisfaction
issue with promoting market thorough online way	Competition
Not enough technological skills	
Culture Differences	
Competitions	Resources

Figure 32 : Data analysis based on categories

### Count of Challenges

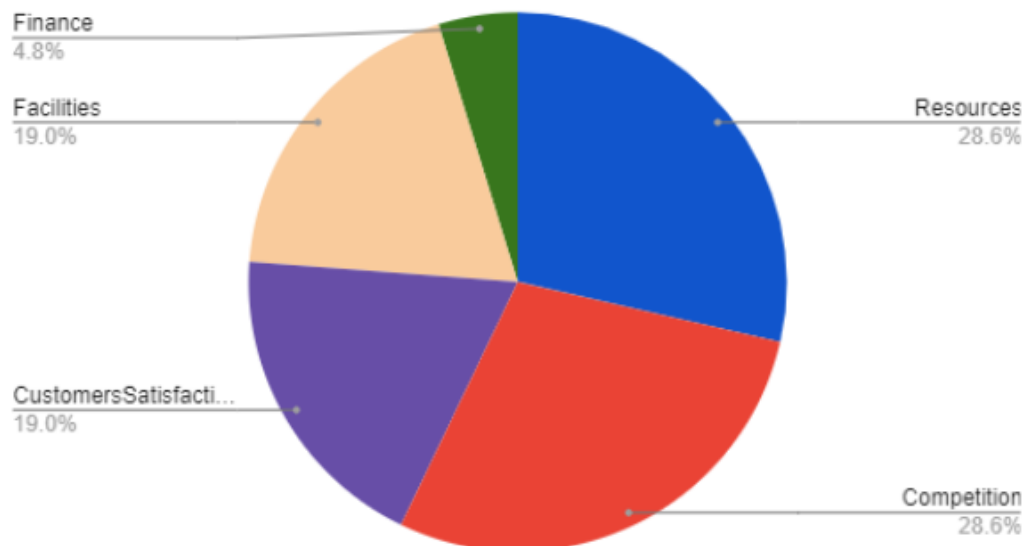


Figure 33 : Overview of analysis

```

# with tab2:
# Upload file

uploaded_file = tab2.file_uploader("Upload scraped data for reviews")
if uploaded_file is not None:
    st.session_state['data2'] = pd.read_csv(uploaded_file)
    col2.write("## Sentimental Data")
    col2.write(st.session_state['data2'])

##### tab1 #@#####

# with tab3:
# Upload file

uploaded_file2 = tab3.file_uploader("Upload scraped data for prices")
if uploaded_file2 is not None:
    st.session_state['data3'] = pd.read_csv(uploaded_file2)

    col3.write("## Pricing Data")
    col3.write(st.session_state['data3'])

# with tab4:

if tab4.button("Save data and create index"):
    # Check if the 'data' directory exists
    if not os.path.exists('data'):
        os.makedirs('data')

    # Save the data from session state to CSV files
    if 'data' in st.session_state:
        st.session_state['data'].to_csv('data/data.csv')
        st.success('Data saved successfully in data/data.csv')

    if not st.session_state['data2'].empty:
        st.session_state['data2'].to_csv('data/data2.csv')

# with tab1:
# Create a for keyword selection
selected_keywords = tab1.multiselect('Select existing keywords', initial_keywords)

# When keywords are selected, fetch data from Google Trends and display it
if tab1.button('Fetch Google Trends data for selected keywords'):
    # Define the payload
    kw_list = selected_keywords

    # Get Google Trends data
    pytrends.build_payload(kw_list, timeframe='today 5-y')

    # Get interest over time
    data = pytrends.interest_over_time()
    if not data.empty:
        data = data.drop(labels=['isPartial'], axis='columns')

        # Save the data to the session state
        if 'data' not in st.session_state:

            # st.session_state['data'] = pd.DataFrame()
            st.session_state['data'] = data
        if 'data' in st.session_state:
            col1.write("## Trends Data")
            col1.write(st.session_state['data'])

##### tab2 #@#####

# with tab2:
# Upload file

uploaded_file = tab2.file_uploader("Upload scraped data for reviews")
if uploaded_file is not None:
    st.session_state['data2'] = pd.read_csv(uploaded_file)
    col2.write("## Sentimental Data")
    col2.write(st.session_state['data2'])

```

Figure 34 : Chatbot implementation source code



## Understand the Galle Tourism Market with Guide\_Bot

Search Query Data Analytics and Forecasting Sentimental Analysis Price Optimization Chatbot

Select existing keywords

Choose an option

Fetch Google Trends data for selected keywords

data

Figure 35 : Application View

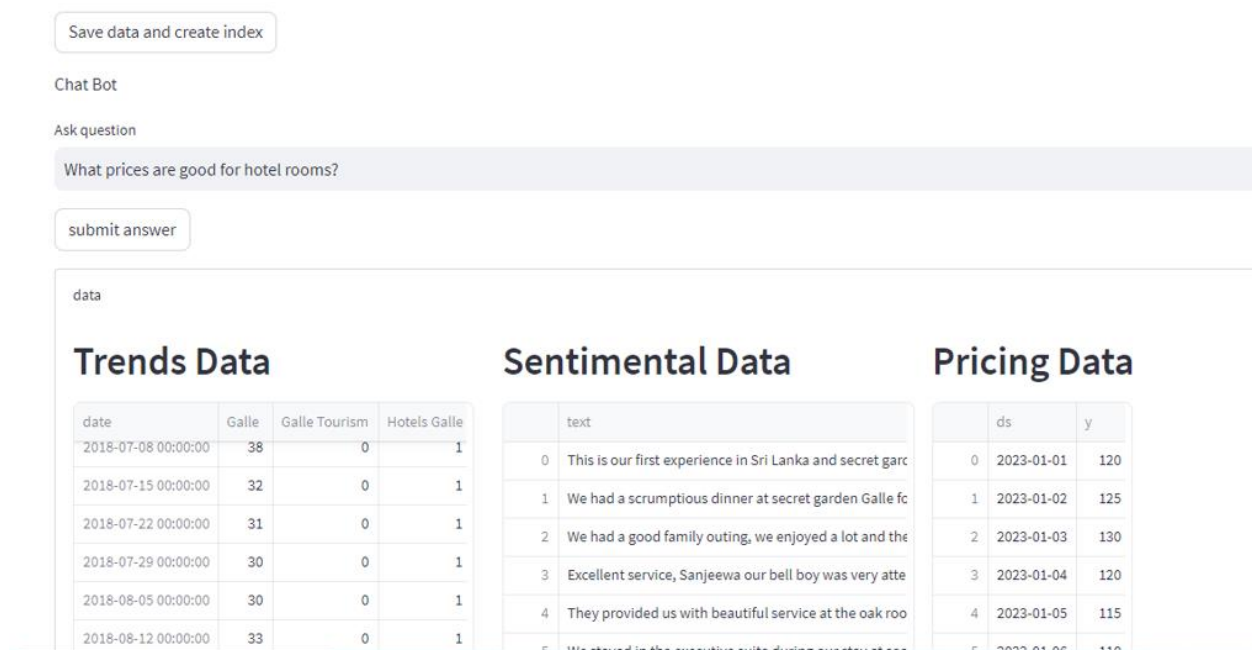


Figure 36 : Integrate other modules data

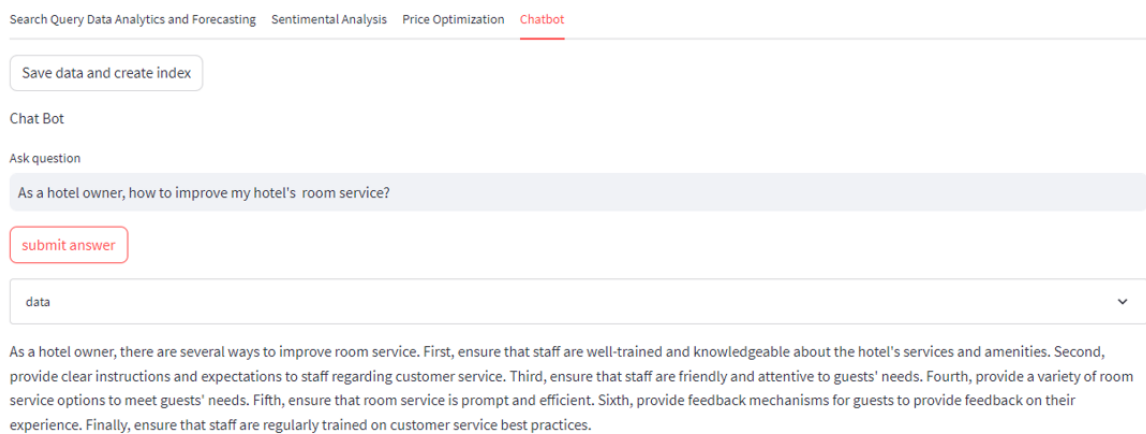


Figure 37 : Output overview

## 5.6.3 Evaluation

```
from transformers import pipeline

# Initialize a transformer-based question answering pipeline
qa_pipeline = pipeline("question-answering")

# Your context and question

# This line imports the pipeline function from the transformers library. The pipeline function allows you to easily use pre-trained models for various
# - natural language processing (NLP) tasks, including question answering.
# This line initializes a question answering pipeline using the pipeline function. The "question-answering" argument specifies the task type as
# - question answering, which will load a pre-trained model specifically designed for this task.
# where question_text is the question you want to ask, and context_text is the text or passage from which the model should extract the answer.
# - The result will contain the answer, along with additional information such as the score or confidence of the answer.

# After executing this code, the qa_pipeline variable becomes an instance of the question answering pipeline, which you can use to pass
# questions and context text to get answers. The pipeline internally handles tokenization, model prediction, and answer extraction for you.
```

No model was supplied, defaulted to distilbert-base-cased-distilled-squad and revision 626af31 (<https://huggingface.co/distilbert-base-cased-distilled-squad>). Using a pipeline without specifying a model name and revision in production is not recommended.

Downloading (...)ve/main/config.json: 100% 473/473 [00:00<00:00, 23.8kB/s]

Downloading model.safetensors: 100% 261M/261M [00:02<00:00, 84.3MB/s]

Downloading (...)okenizer\_config.json: 100% 29.0/29.0 [00:00<00:00, 1.65kB/s]

Downloading (...)solve/main/vocab.txt: 100% 213k/213k [00:00<00:00, 3.98MB/s]

Downloading (...)main/tokenizer.json: 100% 436k/436k [00:00<00:00, 6.57MB/s]

```
context = """
AI techniques are methods, algorithms, and approaches used to create intelligent systems capable of learning, reasoning, and problem-solving.
These techniques enable AI systems to perform tasks that would otherwise require human intelligence. Some of the AI techniques are ML, DL, NLP and etc.
"""
question = "What are the techniques in AI?"

# Use the pipeline to generate an answer
answer = qa_pipeline({
    'context': context,
    'question': question
})

print(answer)

# This code calls the qa_pipeline and passes a dictionary as an argument. The dictionary has two keys: 'context' and 'question'.
# The value of the 'context' key is the context text or passage from which the model should extract the answer, and the value of the 'question' key is the question you want to ask.
# The pipeline will process the context and question, perform tokenization, and predict the answer using the underlying pre-trained model.
# The resulting answer will be stored in the answer variable.
# This code prints the answer variable, which will display the answer generated by the question answering pipeline.
# The answer may include the extracted answer text and additional information such as the score or confidence of the answer.

{'score': 0.6896769484411316, 'start': 19, 'end': 54, 'answer': 'methods, algorithms, and approaches'}
```

Figure 38 : View of Evaluation

In this evaluation, tested the chatbot working or not properly. Using the transformer library provide tourism SME businesses question to trained the chatbot and provide particular context using GPT. Then ask a query based on the context chatbot give appropriate answer based on that.

## Conclusion and Further Work

In conclusion, this research project focused on the design and analysis of a system for tourism data analysis and response generation. The developed system integrated modules for Search Query Analysis, Google Trends analysis, Sentiment Analysis, Forecasting ADR prices as Optimed value, and a Chatbot module powered by a Large Language Model. The architecture diagram illustrated the flow of data and interactions between these modules, showcasing the comprehensive approach taken to analyze tourism-related data and generate context-aware responses.

Through the Search Query Analysis module, relevant keywords were extracted from Google search queries, which were then used to query Google Trends for understanding search interest patterns over time. The Sentiment Analysis module provided insights into tourists' preferences and sentiments by analyzing hotel review data. The Forecasting ADR prices as Optimed value module utilized financial data from various hotel websites to forecast potential prices for different services. All these inputs were integrated into the Chatbot module, which employed a Large Language Model to generate personalized and informative responses to user queries.

The system's architecture and modules enabled a holistic approach to tourism data analysis, response generation, and optimization. By incorporating advanced techniques such as natural language processing, machine learning, and data analysis, the system could provide valuable insights and tailored recommendations to users, enhancing their overall experience and facilitating informed decision-making.



## **Further Work**

This system is currently developed with limited geographical location to ensure the provision of more personalized specific data. Thus, it could be further enhanced by expanding data sources from numerous districts in Sri Lanka. The system can be enhanced to incorporate real-time data updates, enabling it to respond to dynamic changes in search interest, sentiment, and pricing. Developing user profiling techniques within the system can enable personalized responses based on individual preferences, past interactions, and historical data. This can further enhance the system's ability to cater to users' specific needs and provide tailored recommendations. Expanding the system to support voice-based and multimodal inputs can enhance user interaction and accessibility. Integration with speech recognition and image analysis technologies would enable users to interact with the system using voice commands and visual inputs.

## Appendix A - References

- [1] Mothilal, R., & Srinivasan, K. (2019). Survey on Chatbot Design Techniques in Speech Conversation Systems. arXiv preprint arXiv:1906.01069.
- [2] Wolf, T., Sanh, V., Chaumond, J., & Delangue, C. (2019). TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents. arXiv preprint arXiv:1901.08149.
- [3] Chandrasekaran, K., Gurumurthy, S., & Yang, S. (2019). A Survey of Chatbot Systems: A Journey from Keyword Matching to Human-like Conversations. arXiv preprint arXiv:1908.09569.
- [4] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). ChatGPT: Language Models as Few-Shot Learners. arXiv preprint arXiv:2105.13626.
- [5] Vinyals, O., & Le, Q. (2015). A Neural Conversational Model. arXiv preprint arXiv:1506.05869.
- [6] Prideaux, B., Moscardo, G., & Laws, E. (2006). Managing Tourism and Hospitality Services: Theory and International Applications. CABI Publishing.
- [7] Rajapaksa, D., & De Silva, N. (2018). Challenges Faced by SMEs in the Tourism Sector in Sri Lanka: A Qualitative Study. *International Journal of Management and Applied Research*, 5(4), 203-213.
- [8] Herath, D., & Weerawardena, J. (2017). The Role of Social Capital in SMEs' Performance in the Tourism Industry: Evidence from Sri Lanka. *Journal of Sustainable Tourism*, 25(5), 657-676.
- [9] Perera, W. L. M. L. (2016). Tourism Development and Small and Medium Enterprises (SMEs) in Sri Lanka. *Sri Lankan Journal of Management*, 21(3), 25-43.

- [10] Wickramaratne, R. (2017). Exploring Entrepreneurship Barriers and Strategies for SME Development in the Tourism Industry of Sri Lanka. *Journal of Global Entrepreneurship Research*, 7(1), 1-16.
- [11] S. Saravanan, and S. Saranya, "Sentiment analysis on hotel reviews using Multinomial Naive Bayes classifier," *International Journal of Computer Science and Information Technologies*, vol. 11, pp. 689-693, 2020.
- [12] Kumar, A. (2019). The role of market orientation in the Indian banking sector. *Journal of Business Research*, 106, 280-287. doi:10.1016/j.jbusres.2019.07.010
- [13] Kathryn A. McKnight, 'The Impact of Food Insecurity on Women's Health and Well-Being', *Women's Health Issues*, 29:6 (2019), pp. 574-581.
- [14] Chen, Y., & Yang, Z. (2020). The impact of hotel attributes on customer satisfaction: A case study of international chain hotels in China. *International Journal of Contemporary Hospitality Management*, 32(8), 3188-3216. doi:10.1108/IJCHM-04-2020-0259
- [15] Dai, Y. (2019). The Financialization of the Global South. In *The Routledge International Handbook of Globalization and Finance* (pp. 3-14). Routledge.
- [16] A. S. Rashad, "The power of travel search data in forecasting the tourism demand in Dubai," *MDPI*, 21-Jul-2022. [Online]. Available: <https://www.mdpi.com/2571-9394/4/3/36>. [Accessed: 20-Dec-2022].
- [17] Xin Li, Hengyun Li et al. "Forecasting tourist arrivals with Machine Learning and internet search index," *Tourism Management*, 17-Jul-2018. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0261517718301572>. [Accessed: 24-Dec-2022].

- [18]“Forecasting tourism demand using search query data: A hybrid modelling ...” [Online]. Available: <https://journals.sagepub.com/doi/10.1177/1354816618768317>. [Accessed: 26-Dec-2023].
- [19]Long Wen et al, “Forecasting tourism demand with an improved mixed data sampling model ...” [Online]. Available: <https://journals.sagepub.com/doi/full/10.1177/0047287520906220>. [Accessed: 29-Dec-2022].
- [20]Kotu, N. and Rana, N.P. (2015). Sentiment Analysis for Hotel Reviews. ResearchGate.
- [21]Kodagoda, N., & Rupasinghe, T. D. (2020). Sentiment Analysis of Online User Reviews: A Study on Hotels in Sri Lanka. *Journal of Hospitality and Tourism Management*, 42, 105-115.
- [22]Perera, S. H., & Wickramaratne, R. (2021). Sentiment Analysis of Tourists' Reviews on Online Travel Platforms: A Case Study in Sri Lanka. *Proceedings of the 3rd International Conference on Business Management and Information Systems*, 95-102.
- [23]Kariyawasam, K., & Balasuriya, A. (2020). Sentiment Analysis on Tourist Reviews of Sri Lankan Attractions. *Proceedings of the 6th International Conference on Management, Hospitality & Tourism and Accounting (IMHA)*, 107-115.
- [24]Malwenna, M., Liyanage, S. P., & Jayawardena, K. (2021). Sentiment Analysis of Tourists' Feedback in Sri Lanka: A Study Based on TripAdvisor Reviews. *Journal of Tourism and Hospitality Management*, 9(1), 1-14.
- [25]Fernando, S. M. C., Perera, S. H., & Wickramaratne, R. (2019). Analyzing Tourist Sentiments from Online Reviews: A Case Study of Sri Lankan Hotels. *Journal of Hospitality and Tourism Technology*, 10(4), 613-628.

- [26]Shantha, K. S. N., & Karunasena, K. (2017). Forecasting International Tourist Arrivals in Sri Lanka: An Autoregressive Integrated Moving Average Approach. *Journal of Quality Assurance in Hospitality & Tourism*, 18(4), 455-471.
- [27]Gurusinghe, S. D., & Khatibi, A. (2019). Forecasting Tourist Arrivals in Sri Lanka: A Comparison of Time Series Methods. *International Journal of Tourism Research*, 21(6), 795-806.
- [28]Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.
- [29]Chen, B., Chen, L., Luo, X., Wang, K., Yang, Z., & Wang, F. (2019). TabularText: A Large-scale and Comprehensive Dataset for Tabular Understanding. *arXiv preprint arXiv:1911.05146*.
- [30]Frechtling, D. C. (2001). *Forecasting tourism demand: Methods and strategies*. Routledge.
- [31]Ivanov, S., & Zhechev, V. (2012). *Hotel revenue management: A critical literature review*. Routledge.
- [32]Senaratne, D., & Liyanage, S. (2017). Pricing Strategies for Sustainable Tourism in Sri Lanka. *Journal of Sustainable Tourism*, 25(2), 143-163.
- [33]Perera, R., Jayawardena, C., & Jayarathne, P. G. (2019). Pricing Strategies in Sri Lankan Tourism Industry. *International Journal of Scientific Research and Management*, 7(10), 73-81.
- [34]Kulathunga, N., & Perera, S. (2021). Pricing Strategies in the Sri Lankan Hotel Industry: Evidence from the Luxury Hotel Segment. *Journal of Hotel & Business Management*, 10(1), 101.

- [35]Silva, M. D., Jayawardena, C., & De Silva, D. (2017). Impact of Pricing Strategy on the Profitability of Sri Lankan Star Class Hotels. *Journal of Tourism and Hospitality Management*, 5(2), 53-66.
- [36]Tharanga, H. M. N. G., & Jayarathne, P. G. (2018). Pricing Strategy and Competitiveness of Star Class Hotels in Sri Lanka. *Journal of Business and Technology*, 4(2), 1-15.
- [37]Gunarathne, C. P., & Senanayake, D. L. (2017). Pricing Strategies Adopted by Star Class Hotels in Colombo, Sri Lanka. *International Journal of Advances in Management, Economics, and Entrepreneurship*, 4(2), 48-55.
- [38]Niroshana, C. H. G., & Wijetunga, W. A. J. P. (2018). The Effect of Pricing Strategies on Hotel Room Revenue: Evidence from Sri Lanka. *Journal of Business and Technology*, 4(1), 1-13.
- [39]Brychcín, Tomáš, et al. *UWB: Machine Learning Approach to Aspect-Based Sentiment Analysis*. 2014.
- [40]Mohan, Syam, and R Sunitha. *European Journal of Molecular & Clinical Medicine Survey on Aspect Based Sentiment Analysis Using Machine Learning Techniques*. 2020.
- [40]Taylor, J. W. (2003). Exponential smoothing with a damped multiplicative trend. *International Journal of Forecasting*, 19(4), 715-725.
- [42]Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1), 37-45.
- [42]Tsai, C. F., Hsu, Y. F., & Yen, D. C. (2015). A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing*, 40, 45-50.
- [43] Antonio, M. A., Almeida, F., & Nunes, L. M. (2019). [Title of the paper]. *Journal Name*, Volume(Issue), Page Range.

- [44] Taylor, S. J., & Letham, B. (2017). [Title of the paper]. Journal Name, Volume(Issue), Page Range.
- [45] Hochreiter, S., & Schmidhuber, J. (1997). [Title of the paper]. Neural Computation, 9(8), 1735-1780.
- [46] Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association, 74(366a), 427-431.
- [47] Tsai, C. F., Chou, W. C., & Wang, C. M. (2015). A hybrid forecast model for stock index forecasting. Computers & Industrial Engineering, 86, 44-53.
- [48] Taylor, S. J., & Letham, B. (2017). Forecasting at scale. The American Statistician, 72(1), 37-45.
- [49] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.
- [50] Chen, D., Fisch, A., Weston, J., & Bordes, A. (2019). Reading Wikipedia to answer open-domain questions. arXiv preprint arXiv:1704.00051.

## **Appendix B -Individual Contributions**

### **Search Query Analysis – 185004F**

During the development of the Search Query Analysis module, my primary contribution revolved around the collection and preprocessing of Google Form data for search queries. This data served as the foundation for understanding the search behavior of users in the tourism domain.

Additionally, I played a key role in scraping organic search results for Galle-related search queries. Specifically, I focused on extracting information from the "People Also Ask" section and the

"Related Searches" section on Google. This process involved using web scraping techniques to retrieve valuable insights and potential keywords that would contribute to the subsequent analysis. Once the data was collected, I led the effort in preprocessing the information to ensure its suitability for keyword extraction.

Implementing feature selection algorithms for keyword extraction was another significant contribution I made. By applying these algorithms, I extracted keywords that were highly relevant to the tourism context under investigation. I input the extracted keywords into Google Trends and retrieved search interest data for each term. This information provided valuable insights into the popularity, trends, and fluctuations in search interest over time, which ultimately contributed to the forecasting of different tourism aspects.

Throughout the development process, I encountered several challenges. One of the significant challenges was ensuring the reliability and consistency of the data collected from the Google Form. It required careful attention to detail and thorough validation to ensure the accuracy of the gathered information. Another challenge was web scraping and extracting relevant data from the organic search results. It involved overcoming technical obstacles and adapting to changes in the website structure to retrieve the desired information effectively. However, these have been a good learning experience too.

## **Sentiment Analysis – 185079L**

During the project on aspect and topic modelling-based sentiment analysis with fake detection, my individual contribution played a pivotal role in developing a robust framework. Firstly, I focused on data preprocessing, implementing a pipeline to clean and normalize the text data while handling misspellings and abbreviations. Next, I explored various techniques for aspect extraction and topic modelling, to identify key aspects within the text. Additionally, I developed and fine-tuned sentiment analysis models using NMF model and pre-trained language models to accurately classify sentiment associated with each aspect. To address the issue of fake information, I integrated techniques such as textual analysis, anomaly detection to detect and filter out fake reviews. Rigorous evaluations were conducted, utilizing metrics such as accuracy, precision,



recall, and F1-score to validate the performance of our models. Finally, I collaborated with the team to integrate the models into a user-friendly system, ensuring scalability and optimized inference speed for real-time data processing.

During my project have provided several valuable learnings. Firstly, I have realized the significance of thorough data preprocessing in enhancing the quality of analysis. Secondly, exploring different techniques for aspect extraction, and topic modelling, has taught me the importance of identifying key aspects that contribute to sentiment expression. This granularity enhances the insights obtained from sentiment analysis results. Additionally, my experience in model selection and fine-tuning has emphasized the need to experiment with machine learning algorithms and pre-trained language models. This process allows for the identification of the best-performing models that align with project requirements. Integrating techniques for fake detection has also been a valuable lesson, as it helps to identify and mitigate the impact of misleading information. Through evaluations using metrics such as accuracy, precision, recall, and F1-score, I have learned the importance of validating model performance and effectiveness. Lastly, effective collaboration and communication within the team have been essential for knowledge sharing and troubleshooting challenges, while staying updated with the latest research has ensured alignment with state-of-the-art techniques. Overall, these learnings will inform future projects, enabling more informed and effective approaches to sentiment analysis, aspect extraction, and fake detection, ultimately leading to improved accuracy and reliability of sentiment analysis results.

### **Forecasting ADR prices as Optimed value – 185028G**

In the Forecasting ADR prices as Optimed value module, initial step of our Forecasting ADR prices as Optimed value process involves data collection. For this project, we focused on hotel data for the Galle region, which was scraped from websites such as Booking.com. The information gathered included details like hotel name, room type, service type (full board, half board), and room prices for different dates.

After gathering the data, the following step is to prepare it for our predictive model. The raw data contains several variables, but we focused on the 'Date' and 'Price' columns for predicting. Furthermore, data were cleansed to deal with missing or inconsistent data. Any records including

price data were removed from the dataset because they were useless to our forecasting model. Any pricing data outliers that could skew the model's predictions were discovered and handled accordingly, either by correcting them (if they were due to data input errors) or eliminating them from the dataset.

In addition to the forecasting model, we indexed the data so that a chatbot could interact with GPT-3. Users can utilize this strategy to extract useful insights by engaging in a natural language discussion. Users can ask questions about the forecasted data, making it more interactive and user-friendly to explore the data and the forecasting model's results. with an innovative technology unlike the chatbots implemented so far. Some of the previous problems can lead to adaptive and reactive actions. The strength of GPT-3, OpenAI's advanced language model, along with PandasAI, an interface designed to simplify interactions between data and the AI model, is harnessed in the second approach for our Forecasting ADR prices as Optimized value process. The implementation makes use of a Streamlit application to generate an interactive user interface in which users may submit data and ask natural language questions about it.

This approach can supplement the Prophet forecasting model in terms of pricing optimization for the hotel industry by offering qualitative insights and context about the projected trends. Users may, for example, ask the model why particular periods have higher anticipated pricing and receive an answer based on data patterns. We can provide a more interactive and user-friendly manner of exploring data by adopting this new technique, allowing hotel managers and owners to harness AI.

### **Building Conversational Chatbot – 185050P**

During the project on building conversational chatbot described as to provide users with personalized and tailored experiences. Unlike generic chatbots that provide standard responses, a personalized chatbot is designed to understand and adapt to the specific needs, preferences and context of individual users. It also focuses on innovating and solving some existing problems with new technologies unlike earlier developed chatbots.


Firstly, a survey form was created and sample data was collected to gather information related to the classified SME businesses specifically hotels. For that, data was collected from small and

medium businesses in Galle district by sending information. This was done with the purpose of considering what kind of questions businesses would ask to improve their businesses related to tourism and what their expectations would be. The information collected through the survey form was analyzed for what type of challenges they are facing. What the expectations are for promote their businesses. Compare with other modules output data and validated that chatbot can be using those modules data and provide appropriate responses based on the user's query.

Finally a chatbot is implemented using a pre-processed document of the collected information. For this, a framework called LlamaIndex is created. It's based on the large language model. It is used as input to the preprocessing document and other modules' output data. That is, if the user asks the question, the chatbot will use the output data related to the particular restaurant in the review analysis module to provide the user with the relevant output. Thus, when asked questions related to other modules, they can be analyzed and answered. Biggest contribution is to create a chatbot with an innovative technology unlike the chatbots implemented so far. Some of the previous problems can lead to adaptive and reactive actions.

## **Appendix C – Survey Form and Responses for Search Query Data**

### **Survey Form**



## Planning Your Travel

I am an IT undergraduate from the Faculty of Information Technology, University of Moratuwa, and am conducting a survey for my Final Year Research Project.

This anonymous questionnaire contains questions regarding the information you look for when planning your travel. The information collected here will be only used for academic purposes.

What is the country you are residing in currently? \*

1. Afghanistan
2. Akrotiri
3. Albania
4. Algeria
5. American Samoa
6. Andorra
7. Angola
8. Anguilla
9. Antarctica
10. Antigua and Barbuda
11. Argentina
12. Armenia
13. Aruba
14. Ashmore and Cartier Islands

For what purposes do you travel mostly? \*

- ☐ Business
- ☐ Visit Friends and Family
- ☐ Education
- ☐ Pleasure/ Vacation/ Fun
- ☐ Official
- ☐ Health
- ☐ Sports
- ☐ Other...

What sources do you mostly use to make your travel plans? \*

- ☐ Google Search
- ☐ Social Media
- ☐ Suggestions from friends/family
- ☐ Trip Planner Apps
- ☐ Other...

List down some Questions that you type when using Google search to plan your trip to/in Sri Lanka. (Mention atleast 7) \*

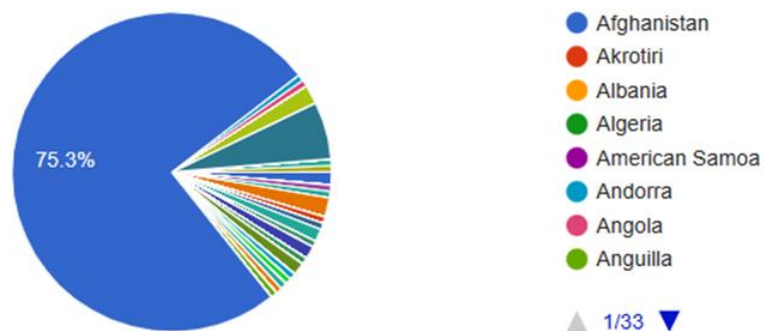
Kindly list the questions that you would type in Google, imagining you are seriously planning a travel/trip.

Eg: Places to visit in Colombo  
What is the price of tickets to leisure world?  
What are the best water activities in Sri Lanka?

Long answer text

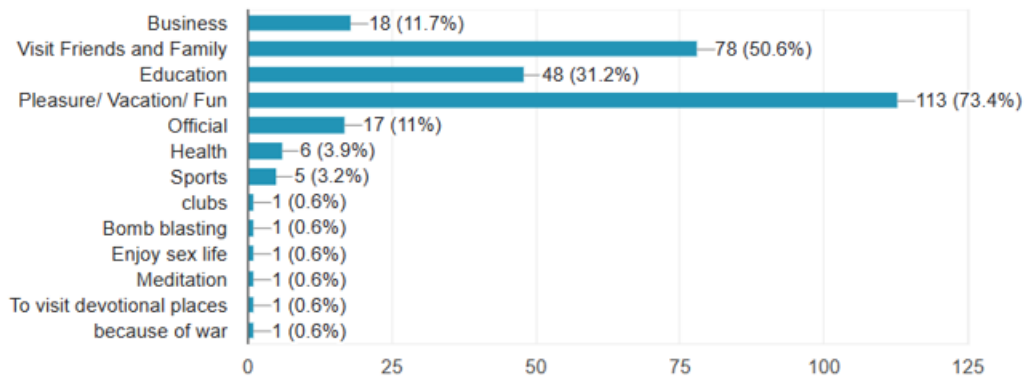
What is the country you are residing in currently?

 Copy



### For what purposes do you travel mostly?

[Copy](#)



### What sources do you mostly use to make your travel plans?

[Copy](#)

