Machine Learning-Based Sentiment Analysis Pipeline for Evaluating Hajj Food Service Quality

Amjad Enad Almutairi, Aisha Yaqoub Alsobhi, Abdulrhman M Alshareef
Faculty of Computing and Information Technology-Information System Department,
King Abdulaziz University, Jeddah 21589, Saudi Arabia

Abstract—**Pilgrimage, also known as Hajj, brings together millions of people each year, creating significant challenges in managing, organizing, and maintaining the quality of various services. Among these essential services, food provision plays a vital role in shaping pilgrims' overall experience and satisfaction. Despite its importance, research focusing on food services using sentiment analysis during Hajj remains limited. Existing studies often rely on social media data, which may not accurately capture the genuine opinions of pilgrims. This study addresses this gap by analyzing food service text reviews collected from Google Maps within the Hajj context. It contributes a new dataset collected for evaluating food services provided to pilgrims after the Hajj season, along with an empirical benchmark for Arabic Hajj food reviews. The dataset consists of 4,018 Google Maps reviews from 160 Hajj campaigns conducted between 2022 and 2025. After data preprocessing, the reviews were classified using several classical machine learning algorithms as empirical baselines, including support vector machine (SVM), logistic regression (LR), naïve Bayes (NB), decision tree (DT), and random forest (RF). The experimental results demonstrate that LR achieved the highest accuracy of 93.6% among the evaluated models, followed by SVM and RF with accuracies of 92.9% and 92.2%, respectively. The analysis also shows that positive sentiment dominated across all studied years, indicating an overall improvement in pilgrims' satisfaction with food services. However, the persistence of food-related issues highlights the need for continued attention and improvement in service quality.**

Keywords*—*Sentiment analysis; service quality; machine learning; Hajj; food service

I. Introduction

Every year, Mecca attracts over two million pilgrims who gather to perform the Hajj, one of the largest religious gatherings in the world. The magnitude of this event requires extensive logistical planning and highly efficient service management across multiple sectors, including healthcare, transportation, accommodation, and food services. The scale and complexity of the Hajj demand meticulous organization to ensure safety, comfort, and satisfaction for millions of pilgrims. Among these services, food quality plays a vital role in maintaining pilgrims' health and overall well-being, as poor management or inadequate quality can lead to dissatisfaction or even health-related risks.

Recently, sentiment analysis, a significant subfield of natural language processing (NLP), has become an essential tool for understanding public opinion and satisfaction levels [1]. It classifies text into emotional categories—positive, negative, or neutral—based on subjective expressions such as reviews, posts, or comments [1]. The exponential growth of online platforms has generated massive volumes of opinion-rich data, making sentiment analysis an effective means of identifying trends and evaluating public perception in business, politics, and social sciences [1].

Within the context of the Hajj, several studies have explored the quality of services such as accommodation [2], transportation [3], and healthcare [4]; however, research focusing on food service quality—especially through sentiment analysis in Arabic—remains limited [5]. Despite these efforts, significant research gaps persist. Most prior studies have relied on data from social media platforms such as X (Twitter) [6], [7], [8], which may not accurately reflect the current state of service quality following recent developments in line with Saudi Arabia's Vision 2030. Furthermore, there is a lack of studies specifically analyzing Arabic-language reviews related to food services during the Hajj, leaving a valuable source of user-generated feedback, such as Google Maps reviews, underutilized. This study aims to address these gaps by employing machine learning–based sentiment analysis on pilgrims' Arabic textual reviews of food services collected after the Hajj.

The specific objectives of this study are: first, to extract and describe the principal food-service issues reported by pilgrims (e.g., availability, safety, cleanliness, and nutrition); second, to estimate the overall sentiment distribution toward food service quality (positive, negative, or neutral); and third, to benchmark multiple machine learning algorithms—including SVM, LR, NB, DT, and RF —to determine the best-performing model for sentiment classification. Accordingly, the main research questions are: What food service issues do pilgrims report during the Hajj?, What sentiments (positive, negative, or neutral) are associated with these issues across pilgrims' reviews?, and Which machine-learning approach most effectively classifies sentiment in this dataset?

This research contributes a new dataset collection applied for the first time to evaluate food services provided to pilgrims after the Hajj season. To the best of our knowledge, it is the first study to analyze Google Maps reviews within the Hajj context with a specific focus on food services. Previous work has seldom relied on Google Maps as a data source in this domain.

The scope of the research is restricted to text-based analysis of reviews about food services during the Hajj. It excludes multimedia data, such as images or videos, and does not address external factors such as politics or finance. The main focus is on textual sentiment to provide an evidence-based understanding of food service quality.

However, there remains a lack of focused research applying these methods specifically to food services in the unique context of the Hajj, creating a need for an updated, domain-focused analysis that utilizes pilgrim-authored reviews beyond social media and rigorously evaluates sentiment models for text to produce actionable insights.

The structure of this research is organized as follows: Section I presents the introduction and problem statement and outlines the research objectives, questions and contributions; Section II reviews related literature on service quality and sentiment analysis in the context of the Hajj; Section III describes the methodology; Section IV presents and discusses the results; and Section V concludes the study with key findings, limitations, and recommendations for future research.

II. Literature Review

Every year, the Hajj attracts a huge number of pilgrims; a total of 1,673,230 pilgrims were recorded in 2025 [9], and the quality of services provided plays a major role in achieving pilgrim satisfaction. The literature review of this research has been divided into two sections: the service quality and sentiment analysis in the Hajj.

A. Services Quality in Hajj

Service quality consists of two component words. “Service” refers to the basic features of a specific function, while “quality” refers to the application of a user-based approach [10]. Together, “service quality” refers to the value of service to the customer [10]. A measure of service quality shows how well a company meets the expectations of its clients by comparing actual service delivery to client expectations [11]. Consumers have explicit wants and needs, as well as occasionally having latent or concealed demands [11]. It is the responsibility of the company to recognize these demands and meet them, either above or beyond the expectations of the client. One of the most crucial factors in determining a company’s success or failure—and, in the event of success, the degree of that success—is service quality [11].

These studies investigate the quality of services provided during the Hajj, including food, accommodation, transportation, and healthcare services.

Food service provision during the Hajj introduces unique challenges and opportunities to improve. Research indicates that there are varying levels of compliance with food safety practices among different food provider companies. Professional training was found to considerably enhance adherence to safety protocols, with a compliance rate between 72.67% and 88.3% [12]. The researchers proposed automated food production using blockchain and centralized plants to control and improve food service quality, reducing service delivery times by 54.46% [5]. During the Hajj, the quality of food service is a critical concern for pilgrims’ satisfaction and health. About 20% of pilgrims suffer from chronic diseases, with 15.7% needing special diets due to chronic health conditions; however, research has found that most caterers do not provide special meals for them [13]. All of these efforts, such as training and automated food production, aim to enhance the quality of food, particularly during the Hajj. Yezli et al. [14] investigated the prevalence of gastrointestinal symptoms among pilgrims during the 2019 Hajj and assessed their knowledge and practices regarding food and water safety. They also aimed to identify knowledge gaps and risky behaviors that could increase the risk of food- and water-borne diseases (FWBDs). The findings showed that nearly 10% (n = 133) of pilgrims reported at least one gastrointestinal symptom during the Hajj, with diarrhea being the most common (51.9%). Most symptoms lasted for less than two days, and 78.2% of affected individuals sought medical assistance [14]. During the 2023 Hajj, Msrahi et al. [15] investigated the relationship between foodborne disease (FBD) and food safety knowledge and practices among pilgrims. They employed a self-reporting online questionnaire, which comprised four sections: demographic data, pilgrims’ food safety knowledge and practice, and FBD symptoms. The dataset collected involved 409 pilgrims from 15 countries. They found that the pilgrims demonstrated an above-average level of food safety knowledge, while their food safety practices were below average [15]. Alfiah et al. [16] investigated the nutritional completeness of food consumed by Indonesian Hajj pilgrims, identifying the factors associated with incomplete nutrition and highlighting the importance of nutritional education and policy. The study included a sample of 231 Indonesian respondents who were Hajj or Umrah pilgrims aged 18 years and above. The researchers found that a significant majority (87%) of Indonesian Hajj pilgrims consumed nutritionally incomplete food, meaning that one or more components (carbohydrates, fat, protein, vegetables, or fruit) were missing [16].

The provision of accommodation during the Hajj plays a crucial role in enhancing pilgrims’ satisfaction and their overall experience. Mohammed et al. [17] used the analytic hierarchy process (AHP) to prioritize the service quality dimensions that are most important for improving pilgrim satisfaction. The most important dimension was found to be tangibility, reflected in the physical facilities, equipment, and the appearance of personnel.The reliability dimension demonstrated the ability to deliver promised services consistently and dependably [17]. Providing high-quality accommodation services, comfortable transportation, and reliable religious scholars all enhance the pilgrimage experience, and authorities should focus on these [18]. Jouda et al. [2] aimed to investigate how service quality affects the satisfaction levels of Hajj pilgrims in Saudi Arabia’s hotel sector and identified key service dimensions that influence the pilgrim’s experience. The authors made suggestions for improving hotel services aligned with religious tourism expectations. They found that service quality significantly impacted pilgrims’ satisfaction, while the dimensions of responsiveness, reliability, and empathy were the most influential in determining satisfaction levels [2].

Large numbers of participants and spatiotemporal restrictions lead to significant challenges in the transportation of pilgrims during the Hajj [3], [19]. Trains and pedestrian routes are intensely used modes of transport during the Hajj [19]. Hussain et al. and Owaidah et al. [3], [19] proposed strategies to improve transport shuttle bus service efficiency and evaluated services by using simulations of discrete events for pilgrim transport modeling between sites. Weber et al. [20] aimed to assess pilgrims’ satisfaction with metro services during the Hajj. They defined the key factors affecting satisfaction, offered proposals and recommendations to improve metro services in the future, and emphasized the role of metro services in mitigating transport crowds during the

Hajj season. They found high satisfaction among pilgrims concerning the quality of the service, with the metro helping to mitigate overcrowding, a positive perception of safety through the measures implemented in the metro system, and improvements to pilgrim mobility; however, the study also found that elderly and disabled pilgrims faced difficulties in accessing metro services [20]. Almujibah et al. [21] aimed to evaluate the efficiency of the Almashaaer Al Mugaddassah Metro (MMMSL) in transporting pilgrims, assess the efficacy of the metro system in meeting its intended objectives, analyze the safety measures implemented in the MMMSL and their effectiveness, and provide recommendations for enhancing performance and safety. They concluded that the metro worked with high efficiency during off-peak times, faced challenges during peak periods, was successful in reducing traffic, and had effective safety measures in place for managing the movement of passengers. The authors also provided a number of improvement recommendations [21]. These studies emphasize the need for innovative solutions to address the complex transportation challenges faced during the Hajj.

Digital transformation initiatives implemented by the Saudi Ministry of Health have led to improved healthcare services [4]. However, integrating various healthcare applications remains a challenge due to the lack of a unified electronic platform [4].

B. Sentiment Analysis in Hajj

Sentiment analysis is a method that uses a vast number of textual data sources to identify positive and negative views about particular goods, services, or events [22]. These research articles targeted particular domains, for instance, sentiment analysis in the education [23], [24], Hajj domain [6], [7], [8], [25], [26], [27], [28], medical domain [29], or business domain [30], [31]. Sentiment analysis offers several advantages, such as gaining insights into students' opinions about the educational process delivered by teachers; understanding people's opinions about specific events, such as the Hajj; helping in gathering patients' feedback on hospital services; and allowing businesses to understand customers' opinions about specific products or services offered in a restaurant or store. These benefits help companies and service providers improve their products and services. The following studies discuss sentiment analysis in relation to the Hajj:

Alghamdi [6] presented a comprehensive sentiment analysis of Arabic tweets related to the Hajj over a six year period from 2017 to 2022 to investigate the prevailing sentiments related to Hajj events before, during, and after the Hajj. The study showed a significant surge in both positive and negative tweets during the Hajj period. It applied different classifications of machine learning and deep learning, with the bidirectional encoder representations from transformers (BERT) model achieving the highest accuracy of 93.8% in sentiment classification. Thus, BERT was found to be one of the most effective models in sentiment analysis, especially for capturing the complexities of Arabic text. The study's findings indicated an increase in both positive and negative sentiments during the Hajj period through engagement and emotional expression [6].

Ottom et al. [8] applied sentiment analysis using a series of tweets related to the Hajj. The researchers collected data from 1-8-2018 to 25-8-2018 using the Twitter application programming interface (API) and conducted two experiments based on the machine learning approach and the lexicon approach to compare the accuracy of the method of analysis. The machine learning-based approach worked better than the lexicon-based approach in the classification and sentiment analysis, with SVM outperforming the other models with a high degree of accuracy (84%) [8].

During the COVID-19 pandemic, a small number of pilgrims were undertaking hajj. Shambour [25] used deep learning methods to discuss people's impressions about the 1442 Hajj season. A dataset was compiled based on approximately 4,300 texts from social media, including X (Twitter) and YouTube, posted over 13 days. Based on the experiment results, the convolutional neural networks and long short-term memory (CNN-LSTM) model outperformed the other models in sentiment classification, achieving 97% accuracy. The researcher provided a clear framework for understanding public perceptions by using a five-point scale for sentiment classification. The overall results showed highly positive feelings about the Hajj season, with ratings exceeding 4 out of 5 [25].

In another study, Gutub et al. [26] investigated the emotional impact of the COVID-19 pandemic on individuals in Makkah and Madinah during the 2021 Hajj season, utilizing deep learning techniques to analyze X (Twitter) data. It aimed to understand the sentiments of residents in these holy cities, which were significantly affected by reduced pilgrim numbers due to the pandemic. The dataset comprised more than 22,000 tweets collected from Twitter accounts in Makkah and Madinah during the 2021 Hajj season (1442 AH) between 28 Dhu al-Qi'dah and 23 Dhu al-Hij'jah. They found "strange negative feelings" toward the Coronavirus pandemic in tweets from the holy cities. The sentiment analysis rate for COVID-19-related tweets in both cities decreased below neutrality, with Makkah tweeters expressing more negative sentiments [26].

Albahar et al. [27] used deep learning to predict satisfaction levels with high accuracy, which influenced the quality of accommodation services. The study collected 15,859 Arabic reviews from the Booking.com website and proposed a model using a deep learning approach that achieved 97% accuracy. The study employed sentiment analysis techniques based on the expectation-confirmation paradigm. This method made it possible to gain a complex understanding of how pilgrims' expectations affected how satisfied they were with the hospitality services provided [27].

Chelloug et al. [28] presented an advanced social media classification system designed to enhance Hajj and Umrah services that utilized predictive deep learning and particle swarm optimization to improve accuracy and offer immediate feedback for better decision-making and service quality management. The researchers analyzed 5,000 tweets, 25,000 Facebook posts, and 10,500 Instagram posts. The proposed model demonstrated superior performance, achieving an accuracy of 98.85% and outperforming other models in classifying social media posts related to Hajj and Umrah services [28].

Several studies have explored food services [14], sentiment [7], [26], [28], and health issues [15], [16] among Hajj and Umrah pilgrims using advanced analytical methods.

In food services, Alasmari et al. [7] aimed to create a novel approach for sentiment analysis of pilgrims by utilizing a hybrid deep learning model combining CNN and LSTM models. The study used a novel and specialized dataset comprising 4,669 tweets collected from X (Twitter) during the 1442 Hajj (2021 AD) that discussed pilgrims' opinions about catering services. The proposed CNN-LSTM model achieved an accuracy of 92%, outperforming other machine learning models (SVM, RF, LR, and DT) and individual deep learning models (CNN and LSTM) [7].

C. Studies Analysis

Most of the studies examined here did not specify the type of service analyzed, focusing instead on general experiences. Table I summarizes previous studies that conducted sentiment analysis during the Hajj using various online platforms, such as X (Twitter). Most of these studies focused on general topics, for example, [25], [8], [6], [26] related to Hajj services while only one study [27] focused on hospitality services. On the other hand, ref. [28] focused on religious rites, management, safety, well-being, and services, but only one study [7] focused on catering services, and it was limited to Hajj 1442 (2021 AD). In terms of data collection periods, the studies collected data during short and specific Hajj seasons, for example, 2018 [8], 2021 [7], and 2023 [27], or over limited durations, such as 120 hours [28]. Concerning platform use, most previous studies have relied on X (Twitter) [8], [25], [7], [28], [26] as the primary source of data, with some incorporating YouTube [25], Facebook [28], or Instagram [28]. With respect to language coverage, previous studies show that sentiment analyses were conducted in Arabic, English, or both.

TABLE I. SENTIMENT ANALYSIS IN HAJJ

Ref.	Platform	Data Collection Period	Languages	Service Type
[6]	X(Twitter)	2017-2022	Arabic	Not specified
[7]	X(Twitter)	Hajj 1442 (2021)	Arabic	Catering services
[8]	X(Twitter)	1-8-2018 to 25-8-2018	English	Not specified
[25]	X(Twitter), YouTube	Hajj 1442	Arabic	Not specified
[26]	X(Twitter)	Hajj season (1442 AH) between 28 Dhu al-Qi'dah and 23 Dhu al-Hij'jah	Arabic, English	Not specified
[27]	Booking	Hajj 2023	Arabic	Hospitality service
[28]	X(Twitter), Facebook, Instagram	collected over a continuous period of 120 hours	Arabic, English	Religious rites, management, safety, well-being, and services

These studies depend on data from social media platforms in sentiment analysis, which does not offer straightforward conclusions about the quality of services provided to pilgrims during the Hajj. The results showed the effectiveness of using machine learning algorithms and deep learning algorithms in sentiment analysis. In addition, many of these studies were conducted several years ago, which may limit their relevance to the Hajj at the present time. Table II shows the differences between studies based on number of datasets, machine-learning models used, sentiment analysis used, and the levels of accuracy achieved by the machine-learning or deep learning models in each study. Table I presents the differences between studies that apply sentiment analysis in relation to the Hajj. The study has aimed to measure the effectiveness of machine learning techniques in improving the quality of food services that are provided to pilgrims based on their sentiment analysis.

In sum, our research stands out in three key ways when compared with previous studies in the Hajj field, as shown in Table I. First, in terms of specificity, it focuses directly on the food service aspect of the Hajj, unlike most prior works. Second, in terms of recency and duration, our review covers the 2022–2025 period, providing a modern and multi-year perspective. Lastly, in terms of practical relevance, this study uses sentiment analysis and machine learning techniques to contribute practical insights for identifying the main problems related to food services that pilgrims face during the Hajj.

TABLE II. SUMMARY OF RELATED RESEARCH OF MACHINE LEARNING AND SENTIMENT ANALYSIS IN HAJJ

Ref.	Dataset	Classifier	Features Extraction	Sentiment Polarity	Accuracy
[6]	Over 80,000 tweets	LR, RF, KNN, SVM, NB, XGBoost, CNN, LSTM, BERT	TF-IDF, BoW, Word2Vec Embeddings	Positive, Negative, Neutral	93.8% BERT
[7]	4669 tweets	SVM, DT, RF, LR, CNN, LSTM, CNN-LSTM	One Hot Encoder	Positive, Negative, Neutral	92% CNN-LSTM
[8]	3175 tweets	SVM, KNN, NB, Text Blob	BoW, N-Grams, TF-IDF	Positive, Negative, Neutral	84% SVM
[25]	2996 tweets, 1293 comments	CNN-LSTM	CNN	five-point scale for sentiment classification	Reach to 97%
[26]	Over 22,000	CNN-LSTM	CNN	Scale score from 1 to 5	Not specified
[27]	15,859 reviews	CNN, LSTM, proposed model	Convolutional Layer	Positive, Negative, Neutral	97% proposed model
[28]	15,000 tweets, 25,000 Facebook posts, and 10,500 Instagram posts	SVM, RF, DT, XGBoost, K-NN, BERT, ALBERT, ROBERTA, XLNet, Proposed model	Principal Component Analysis (PCA) and Pearson Correlation Coefficient (PCC)	based on relevant attributes: service-level, and scores	98.85% proposed model

III. METHODOLOGY

This research used a sentiment analysis approach applied to pilgrims' reviews about food service. Sentiment analysis is a widely recognized approach in natural language processing (NLP) that focuses on identifying and interpreting emotions or opinions expressed in text data [32], [1]. Its main objective is to classify written content—such as reviews, tweets, or comments—into sentiment categories such as positive, negative, or neutral [32], [1]. With the exponential rise of social media and digital communication, massive amounts of opinion-rich data are generated daily, making sentiment analysis an essential tool for understanding public attitudes, market trends, and consumer behavior [33].

This approach plays a crucial role in various domains, including business, politics, and social research, by helping organizations analyze feedback, monitor reputations, and enhance decision-making [33], [32]. Sentiment analysis approaches can be divided into three categories, as shown in Fig. 1: machine learning, deep learning, or ensemble learning [1]. In this research, the focus is on the machine learning approach.

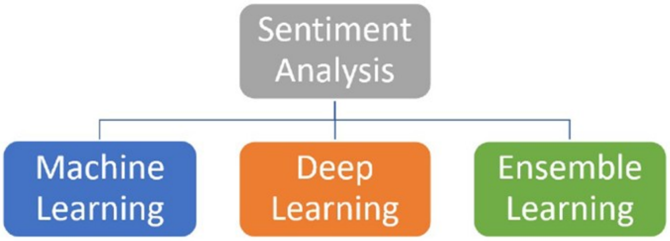

Fig. 1. Sentiment analysis approach.

Sentiment analysis typically involves several key stages [1], as illustrated in Fig. 2. It starts with data acquisition, then perform data preprocessing to clean and normalize the text. After that, feature extraction converts text into numerical representations using methods like TF–IDF (term frequency–inverse document frequency) or word embeddings. Finally, machine learning algorithms such as SVM for classification are used [1]. By integrating natural language processing and machine learning techniques, sentiment analysis provides a powerful framework for automatically extracting and interpreting human emotions and opinions from unstructured text data [33], [32], [1].

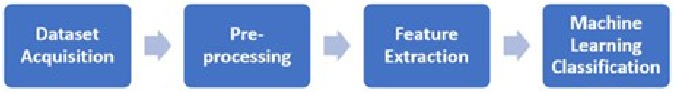

Fig. 2. Sentiment analysis process.

This section accentuates our research methodology, which includes the dataset collection, labeling, and preprocessing, as well as the subsequent steps of feature extraction and model classification, as shown in Fig. 3.

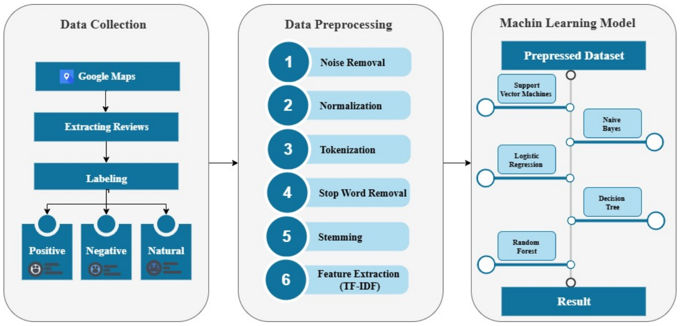

Fig. 3. Methodology overview.

A. Dataset Collection

The foundation for sentiment analysis has been established by collecting text-based feedback from Hajj pilgrims. This feedback covered sentiments related to food service quality. The data have been gathered from a range of sources between 2022 and 2025, from 160 Hajj campaigns. The research data for this study have been collected from Google Maps reviews related to campaigns at Hajj sites using the Outscraper website [34]. This tool offers rapid, reliable access to data on Google Maps locations through a simple and intuitive interface [34]. Additionally, we utilized Instant Data Scraper, a feature of Google Chrome that serves as an automated data extraction tool for any website. The dataset was garnered from Google Maps. We initially gathered 6,450 reviews; after removing duplicates, non-text items, and entries with missing or off-topic content, we retained 4,018 reviews explicitly referencing food or catering. The data collection process involved several steps. First, specific keywords were searched on Google Maps to identify Hajj campaigns, such as "حملات" (campaigns), "حملة حج" (hajj campaign), "حجاج الداخل" (domestic pilgrims), and "حج" (hajj). Based on the search results, relevant campaigns were selected, and all available reviews were opened and extracted using the Instant Data Scraper tool. The collected reviews were then downloaded, copied, and combined into a single file. This process was carried out individually for each campaign, which required a considerable amount of time due to the manual nature of the task. After completing the data collection, the dataset was cleaned and standardized to ensure consistency. The data preparation process included unifying all reviews into Arabic, standardizing the review dates into an annual format, and organizing the dataset for data labeling and preprocessing. Table III presents the dataset attributes and their types used in this study.

TABLE III. Potential Attributes and Data Type

Attribute	Data Type
Textual review	Text
Review datetime	Numerical
Sentiment analysis	Categorical

Fig. 4 shows a bar chart illustrating the relationship between review years and sentiment categories. The chart shows a gradual increase in the number of reviews over the years. There are relatively few reviews in 2022, with a noticeable rise in 2023 and a sharp increase in 2024 and 2025. Positive reviews remain the most frequent across all years, followed by negative and neutral reviews.

1) Dataset Labeling: The dataset has been organized into three class single-label sentiment classification (positive/neutral/negative) to encompass the entire spectrum of classification. The positive class comprises reviews that are commendatory or appreciative of the pilgrim' satisfaction with the food or service, such as food quality, good taste, cleanliness, variety, organization, or speed of service. On the other hand, the negative class encompasses reviews that are derogatory and critical and reflect dissatisfaction or a poor experience, such as poor taste, delayed service, lack of variety, poor cleanliness, or poor service by servers. Finally, a neutral class expresses a descriptive opinion or information without positive or negative emotion (i.e., an objective comment that does not express obvious satisfaction or dissatisfaction); for instance, three meals are served daily to pilgrims. Table IV shows examples of three categories of sentiment analysis. The

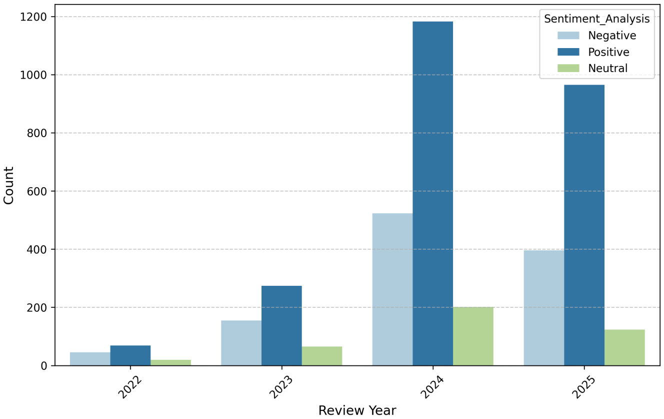

Fig. 4. Dataset distribution.

dataset was manually labeled by assigning a sentiment label to each review individually using an Excel file, which was later converted into a CSV format. The final dataset contains 4,018 Arabic reviews provided by pilgrims. To ensure labeling reliability, all reviews were annotated by three native Arabic speakers. One annotator has a background in information technology (IT), while the other two are in information systems (IS). All annotators were between 28 and 32 years old and participated independently in the labeling process. During data preparation, only the review text and review date were retained, while star ratings, no campaign identifiers or reviewer IDs were excluded to avoid potential data leakage risks and bias. Each annotator was provided with the same dataset and instructed to classify each review into one of the three sentiment categories (positive, negative, or neutral) based on the predefined definitions. After all annotators completed the labeling process, the final label for each review was determined using majority voting. For example, if two annotators labeled a review as positive and one labeled it as neutral, the final label was assigned as positive. In ambiguous cases, the annotators discussed the review collectively to decide the appropriate sentiment class, and reviews that could not be confidently resolved were removed from the dataset.

Fig. 5 illustrates the distribution of sentiments and the number of reviews in each class in our dataset. The largest portion of the chart represents positive sentiments, which account for 62.00% (2,491 reviews) of the total. The negative sentiments make up 27.85% (1,119 reviews), while the neutral sentiments represent the smallest portion, with 10.15% (408 reviews). It is clear that positive reviews dominate the dataset, followed by negative and then neutral reviews.

B. Dataset Preprocessing

Preprocessing techniques are important in natural language processing tasks [35]. The quality of preprocessing has a direct impact on the performance and precision of NLP tasks such as text classification and sentiment analysis [35]. One of the most widely used libraries in NLP is the Natural Language Toolkit (NLTK) [36]. It is a Python library that offers a variety of linguistic resources, like text processing, and performs operations such as tokenization, stemming, and classification [36]. To prepare the data for analysis, preprocessing tasks, as shown in Table V, such as text cleaning and tokenization, have been performed to ensure consistency. The reviews obtained for this study consisted of feedback written in Arabic as well as several other languages. To standardize the linguistic input and avoid model bias arising from multilingual text, we translated all non-Arabic reviews into Arabic using Google Translate. This translation step served to unify the dataset under a single language representation and enable the application of Arabic-specific preprocessing and sentiment-analysis techniques. Accordingly, the following steps were applied, before building the machine learning model, to ensure the efficiency of its accuracy.

TABLE IV. Examples of Review in Different Sentiment Classes

Sentiment	Review	English translation
Positive	حملة جدا رائعة وممتازين في الخدمة والاعاشة اقيمهم خمس نجوم 🤩	A truly wonderful campaign; excellent service and catering. I rate them five stars 🤩
	شركه ممتازة ونظافة والاكل كان رائع كل التوفيق لكم 🩶 🙏	Excellent company, very clean, and the food was great. Wishing you all the best 🙏 🩶
Negative	البوفيه سيء جدا خيارات محدودة وما يناكل الله يعز النعمة	The buffet was very bad, limited options, and barely edible. May God bless the food we have.
	الحملة جدا سيئة لا ترتيب ولا تنظيم ولا سناكات والاكل جدا سيء نفس الاصناف مكررة يوميا والله الإعاشة زي الزفت بصراحة 👎	The campaign was really bad — no order, no organization, no snacks, and the food was terrible, same dishes repeated daily. Honestly, the catering was awful 👎
Neutral	كان الطعام جيدًا ولكنه تقليدي بشكل أساسي	The food was good but mostly traditional.
	طبخهم للوجبات لا بأس به اهم شي ما تموت يعني مو لازم الطعم يكون ١٠٠ % 😅	Their cooking is okay — at least you won't die, the taste doesn't have to be 100% 😅

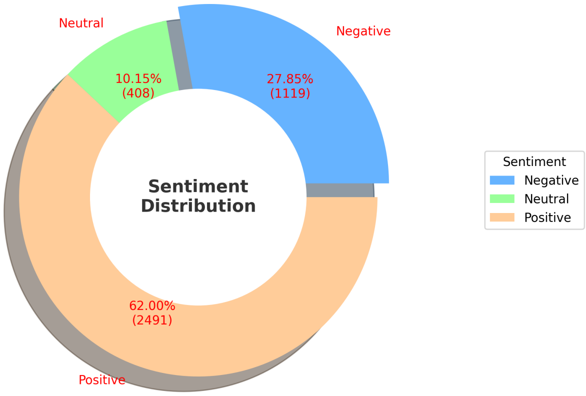

Fig. 5. Number of positive, negative and natural reviews in our dataset.

1) Noise removal: According to [37], removing noise enhances analysis speed and accuracy. This involves removing any URLs, special characters, punctuation, and numbers from the reviews. To address these complexities in our dataset, we implemented a Python function called *remove_noise(text)*, which relies on regular expressions and the “sub” method from the “re” package [37] to systematically preprocess Arabic reviews, and applies it to a panda DataFrame. This function removes unwanted elements such as digits, punctuation marks, and unusual symbols. It also cleans emojis, reduces elongated Arabic vowels, and normalizes white space. The cleaned text is then saved into a new column called *clean_text*, ensuring a clean and readable dataset for further analysis.

2) Normalization: Text normalization is employed to reduce noise within Arabic data and to correct spelling inconsistencies, as noted in prior research [38]. This procedure has been shown to enhance model performance in multiple studies [38], [6]. For example, the authors of [38] reported improved accuracy after applying normalization along with other preprocessing steps. Arabic text presents additional complexities, as some letters can appear in multiple forms depending on their position within a word. To address this, we normalized these letter variants—for example, converting ة (taa' marbuta) to ه, standardizing the different forms of آ, أ and ا to a single ا (alif), converting alif maqsura ى into ي, and changing letters with hamza (ؤ and ئ) into a stand-alone hamza ء. Arabic diacritics (such as fatha, damma, and kasra) are also removed, as they don't contribute significantly to sentiment analysis.

3) Tokenization: Tokenization is a fundamental preprocessing step in Arabic text analysis because it reduces typographical variation and prepares the text for reliable linguistic processing [39], [38]. Tokenization is the process of splitting the text into individual words or tokens [22]. For Arabic text, tools like NLTK, can be used to perform this step efficiently. In summary, this step converts the normalized Arabic text into separate words, making it easier to process in later tasks such as removing stop words, stemming, or performing deeper text analysis. Effective tokenization facilitates more precise counting of word occurrences and enhances the overall structure and interpretability of the text for NLP models [39], [38].

4) Stop word removal: Removing stop words—common words like في, على, من , etc.—is another key task, as these words don't carry much useful information for sentiment analysis [22]. We used the NLTK stop words corpus to filter out common Arabic words that usually carry little semantic meaning [39]. First, it downloads and loads the Arabic stop words provided by NLTK into a set called *stop_words*. Then, it extends this list with a custom set of additional Arabic stop words that the user defines manually (e.g., pronouns, prepositions, and frequently used words like من, على, في , etc.). After combining both sets, a function called *remove_stopwords* is defined, which takes a list of tokens (words) and returns only those that are not in the stop words set.

5) Stemming or lemmatization: Stemming, or lemmatization, is performed to reduce words to their root forms, ensuring variations of the same word [22]; for example, "وجبات" becomes "وجب". In this research, we apply stemming to Arabic words using NLTK's ISRIStemmer as implemented in [39], which is a rule-based algorithm specifically designed for the Arabic language. This step refines the text preprocessing pipeline by unifying word variants under a common root, which enhances the effectiveness of tasks such as text analysis and machine learning by treating semantically related words as the same concept.

TABLE V. Example of Data Preprocessing

Data preprocessing	**Example**
Original Text	♥ حملة جج رائعة ومتميزة جدا ما ينقص المخيم اي شيء، عشاء وغداء وفطور بوفيه كامل متكامل
Noise Removal	حملة جج رائعة ومتميزة جدا ما ينقص المخيم اي شي عشاء وغداء وفطور بوفيه كامل متكامل
Normalization	حمله جج رائعه ومتميزه جدا ما ينقص المخيم اي شي عشاء وغداء وفطور بوفيه كامل متكامل
Tokenization	حمله، جج ، رائعه ، ومتميزه ، جدا، ما، ينقص، المخيم ، اي، شي، عشاء، وغداء، وفطور ، بوفيه ، كامل، متكامل
Stop-Word Removal	حمله، جج ، رائعه، متميزه، ينقص، مخيم ، عشاء، غداء، فطور، بوفيه ، كامل، متكامل
Stemming	حمل، جج ، روع، ميز، نقص، خيم ، عش، غد، فطر، بف، كمل، تكمل

6) Feature extraction: After preprocessing, relevant features have been extracted for input in machine learning models. This process converts raw text into numerical features that machine learning models can understand. Examples of these features include bag-of-words (BoW), term frequency–inverse document frequency (TF–IDF), and word embeddings.

For the sentiment analysis task, the text data were transformed into numerical features using the TF–IDF technique, which quantifies the importance of words or phrases within a collection of documents. This process prepared both the text data and sentiment labels for machine learning. The categorical sentiment labels ("positive," "negative," and "neutral") were first converted into numerical form through a mapping dictionary, where the values 0, 1, and 2 represented positive, negative, and neutral sentiments, respectively. This encoding allowed the data to be used effectively in algorithms that require numerical input. The textual data were then vectorized using TF–IDF to assign weighted values to each term, emphasizing those that are more relevant to individual documents while downplaying common or less informative terms. To optimize feature extraction, several parameters were configured: The n-gram range was set from 1 to 3 to capture single words, bigrams, and trigrams; the minimum document frequency (min_df=3) excluded rare terms occurring in fewer than three documents; the maximum document frequency (max_df=0.9) removed overly common words appearing in more than 90 percent of documents; sublinear term frequency scaling (sublinear_tf=True) was applied to reduce the influence of very frequent terms; and the maximum number of features was limited to 5,000 to manage the dimensionality of the feature space. The resulting TF–IDF matrix represented each document as a numerical vector corresponding to the relative importance of its terms, and the matrix dimensions along with the vocabulary size were recorded to validate the transformation.

C. Machine Learning Models

Machine learning, a vital subfield of artificial intelligence (AI), enables computers to learn and enhance their performance through experience without requiring explicit programming [40]. It is designed to help systems automatically identify patterns, make predictions, and make informed decisions based on data [40]. We used five machine learning algorithms to evaluate our dataset.

1) Support Vector Machine (SVM): Support vector machine is a supervised machine learning algorithm used for both classification and regression tasks [41]. The goal of SVM is to find the best possible hyperplane that divides the data into distinct classes, while maximizing the margin, or the distance, between the hyperplane and the closest data points from each class [41].

2) Naive Bayes (NB): Naive Bayes classifiers are a group of algorithms based on Bayes' Theorem, which is used to calculate probabilities for classification tasks [42]. Rather than referring to a single method, "Naive Bayes" covers a family of models that make the simplifying assumption that the features are independent of each other [42]. This assumption makes the models computationally efficient, especially when working with large datasets [42].

3) Logistic Regression (LR): Logistic regression is a supervised learning technique commonly used for classification problems, particularly when the goal is to estimate the probability that a given instance belongs to a specific class [43]. It works by calculating a value between 0 and 1, representing the likelihood that an event will occur, based on the relationship between the input features and the target variable [43].

4) Decision Tree (DT): A decision tree is a supervised machine learning technique that can be applied to both classification and regression problems [44]. Its structure resembles a hierarchy, beginning with a root node and branching into internal nodes, edges, and leaf nodes [44]. Much like a flowchart, it guides the decision-making process step by step: Internal nodes perform tests on attributes, branches denote the values of these attributes, and leaf nodes provide the final outcome or prediction [44]. Decision trees are commonly employed because they are easy to interpret, are adaptable, and require minimal data preprocessing [44].

5) Random Forest (RF): Random forest is a powerful machine learning technique that combines the predictions of multiple decision trees to improve accuracy [45]. During training, it builds several decision trees, each working with a different subset of the data and features. This "ensemble" approach helps make the model more robust, reducing the risk of overfitting and enhancing its ability to generalize well to new data [45].

D. Performance Measures

The quality of sentiment analysis has been assessed using various metrics and indicators to measure its performance. In this section, multiple machine learning algorithms have been compared based on performance metrics such as accuracy, precision, recall, and F1-score to determine which algorithm best predicts sentiment. Each metric has a unique formula, and we evaluated how well classifiers performed by using mathematical formulas. The following definitions of symbols are used in the formulas. "TP" refers to instances that are correctly classified as positive, "TN" signifies instances that are accurately classified as negative, "FP" represents instances that are incorrectly classified as positive, and "FN" denotes instances that are incorrectly classified as negative.

Accuracy refers to the overall proportion of correct classifications made by the model, including both positive and negative predictions [6].

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

The recall is the percentage of actual positive cases that the model correctly identifies as positive [46].

$$\text{recall} = \frac{TP}{TP + FN} \tag{2}$$

Precision is the proportion of positive predictions made by the model that are actually correct [46].

$$\text{precision} = \frac{TP}{TP + FP} \tag{3}$$

The F1-score is a way to combine precision and recall into a single metric. It is the harmonic mean of the two, balancing the trade-off between them [46].

$$\text{f1} - \text{score} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{4}$$

This comparison has led to a final interpretation, visualization, and evaluation of the different approaches. By gathering feedback data, preprocessing it for machine learning models, and applying various algorithms, issues with food service quality can be identified.

IV. Results and Discussion

The following sections present the detailed results obtained from each model, followed by a comparative discussion that interprets their performance and highlights the most effective approach for analyzing pilgrims' opinions about food service quality during the Hajj.

A. Machine Learning Evaluation

Table VI presents the performance metrics for all classifiers, including accuracy, precision, recall, and F1-score. These results were obtained using the optimal parameters determined through a grid-search procedure. We also implemented an adaptive hyperparameter tuning strategy that automatically selected the appropriate parameter grid based on the characteristics of each model type. This approach ensured that each classifier received targeted parameter optimization focused on the most relevant hyperparameters. Using f1_weighted as the scoring metric facilitated balanced performance across all sentiment classes. To ensure the robustness and reliability of the findings, the dataset was preprocessed and divided into training and testing subsets using a 70/30 split ratio (total dataset 4018, of which 2812 were for training and 1206 for testing), as illustrated in Table VII, followed by fivefold cross-validation. Stratified sampling was employed to maintain the original class distribution in both subsets. Examination of the training data revealed class imbalance, with some sentiment categories being underrepresented. To address this issue, the synthetic minority oversampling technique (SMOTE) was applied to the training set. SMOTE generates synthetic samples of minority classes, thereby achieving a balanced dataset without duplicating existing entries [47]. The resulting balanced dataset ensured that the classification models were trained on an equal representation of all sentiment categories and no duplicate reviews across splits, thereby reducing bias and improving the reliability of the predictive outcomes.

TABLE VI. CLASSIFICATION RESULT

Model	Accuracy	Precision	Recall	F1-score
LR	**0.936153**	**0.926571**	**0.862949**	**0.888955**
SVM	0.928690	0.900183	0.859686	0.877759
RF	0.922056	0.916339	0.839022	0.870070
NB	0.912106	0.846642	0.855347	0.850854
DT	0.865672	0.825404	0.792528	0.807579

TABLE VII. CLASS DISTRIBUTION IN TRAINING AND TEST SETS

Dataset Split	Positive	Negative	Neutral	Total
Full dataset	2491	1119	408	4018
Training set (70%)	1743 (61.98%)	783 (27.84%)	286 (10.17%)	2812
Testing set (30%)	748 (62.03%)	336 (27.86%)	122 (10.12%)	1206

The comparison of machine learning models for sentiment analysis of food services during the Hajj demonstrates notable variations in performance across algorithms. Fig. 6 presents a bar chart that compares the performance metrics of different models. Among them, the LR classifier achieved the highest performance, with an accuracy of 93.6%, a precision of 92.7%, a recall of 86.3%, and an F1-score of 88.9%. This strong balance between precision and recall indicates that LR not only minimizes false positives but also captures most true positives, which is an ideal outcome for sentiment classification. The SVM and RF classifiers also performed competitively, with accuracies of 92.9% and 92.2%, respectively. SVM achieved 90% precision, 86% recall, and an F1-score of 87.8%, whereas RF achieved 91.6% precision, 83.9% recall, and an F1-score of 87%. The NB classifier demonstrated moderately strong performance, with an accuracy of 91.2%, a precision of 84.7%, a recall of 85.5%, and an F1-score of 85.1%. In contrast, the DT classifier delivered the weakest performance, achieving 86.6% accuracy, 82.5% precision, 79.3% recall, and an F1-score of 80.8%. As shown in Table VI, LR achieved the highest overall accuracy (93.6%) and F1-score (88.9%), indicating its strong ability to correctly classify pilgrims' opinions on food services. The competitive performance of SVM and RF suggests that these models also provide effective alternatives for text classification tasks. In contrast, the relatively low performance of the DT model may be attributed to its susceptibility to overfitting.

These results indicate that LR offers the most equitable and dependable performance for the analysis of textual feedback. Therefore, it can be considered the most suitable model for identifying and interpreting sentiments expressed by pilgrims regarding food quality, distribution, and overall satisfaction during the Hajj season.

Table VIII reports class-wise precision, recall, and F1 scores for each model, along with the macro-averaged F1 score. Most models achieve strong performance on the positive and negative classes, whereas performance on the neutral class is consistently lower, as expected given its smaller sample size (408 compared to 1119 and 2491). Among the evaluated models, LR achieves the highest macro-F1 score, indicating more balanced classification performance across the sentiment categories.

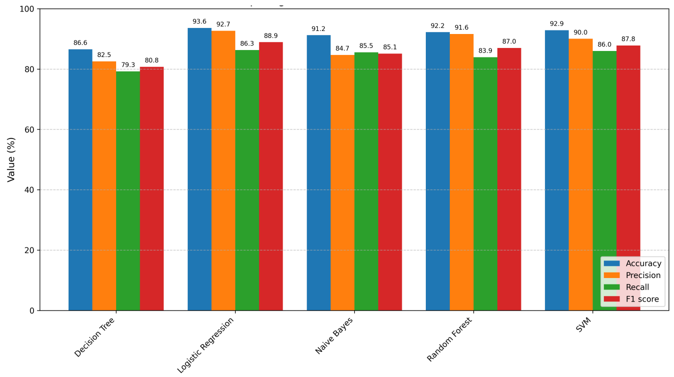

Fig. 6. Comparing the performance metrics of different models.

To assess the stability of the results, LR was trained and evaluated over ten runs using different random seeds, as shown in Table IX. The macro-averaged F1 score was reported as the mean across runs, along with a 95% confidence interval. To evaluate the robustness of the LR model, the experiment was repeated ten times using different random seeds with stratified train–test splits. The macro-averaged F1 score across runs ranged from 0.859 to 0.907, with a mean value of 0.887 and a standard deviation of 0.015. The corresponding 95% confidence interval [0.878, 0.897] indicates limited variability in performance, suggesting that the model's balanced classification ability is stable and not sensitive to a particular data split. This result is consistent with the macro-F1 value reported in Table VIII, confirming that the observed performance is reproducible and not an artifact of a single evaluation split.

TABLE VIII. CLASSIFICATION RESULT IN EACH CLASS

Model	Class	Precision	Recall	F1-score	Macro-F1
LR	positive	0.951031	0.986631	0.968504	0.888955
	negative	0.905605	0.913690	0.909630	
	neutral	0.923077	0.688525	0.788732	
SVM	positive	0.948320	0.981283	0.964520	0.877759
	negative	0.909091	0.892857	0.900901	
	neutral	0.843137	0.704918	0.767857	
RF	positive	0.928030	0.982620	0.954545	0.869824
	negative	0.911315	0.886905	0.898944	
	neutral	0.908046	0.647541	0.755981	
NB	positive	0.959569	0.951872	0.955705	0.850854
	negative	0.892857	0.892857	0.892857	
	neutral	0.687500	0.721311	0.704000	
DT	positive	0.895541	0.939840	0.917156	0.807579
	negative	0.830671	0.773810	0.801233	
	neutral	0.750000	0.663934	0.704348	

Fig. 7 illustrates the accuracy results across different machine learning models, with LR achieving the highest score of 93.6% and DT achieving the lowest score of 86.6%.

Fig. 8, Fig. 9, and Fig. 10 present the performance comparison of several machine learning models—SVM, NB, LR, DT, and RF—used for sentiment analysis of food service

TABLE IX. Stability Analysis of LR using Repeated Runs

Metric	Value
Number of runs	10
Macro-F1 scores	[0.88305793 0.85866379 0.8780987 0.88270304 0.90666875 0.88068048 0.90701475 0.87978505 0.89923511 0.89673202]
Mean Macro-F1	0.8873
Standard Deviation	0.0151
Minimum Macro-F1	0.8587
Maximum Macro-F1	0.9070
Confidence Interval 95%	[0.8779, 0.8966]

experiences during the Hajj. The sentiment categories considered in this analysis are positive, negative, and neutral, and performance is evaluated using recall, precision, and F1-score.

In Fig. 8, the positive and negative classes achieved high precision across most models, indicating that when the models predicted a positive or negative sentiment, they were usually correct. The neutral class showed greater variability, with LR performing best for this category. This finding suggests that neutral predictions are more prone to misclassification compared to the other sentiment types.

Fig. 9 shows that all models achieved high recall for the positive class, particularly LR and SVM, both exceeding 0.95. This indicates that these models were highly effective in correctly identifying positive sentiments. For the negative class, recall values were slightly lower but still strong, typically around 0.85–0.9, meaning that most negative sentiments were also captured well. However, the neutral class consistently exhibited the lowest recall across all models, suggesting that distinguishing neutral opinions remains a challenge.

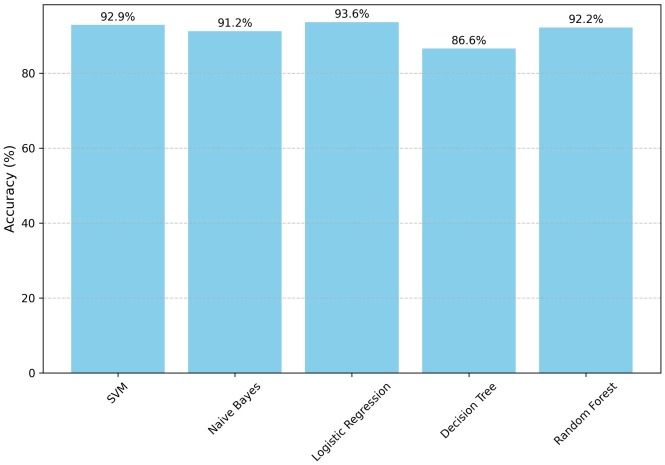

Fig. 7. Accuracy results comparing the performance of several machine learning models.

The F1-score results, shown in Fig. 10, reveal a similar pattern. The positive class achieved the highest F1-scores across all models, followed by the negative class. The neutral class again received the lowest F1-scores, reflecting the overall difficulty of detecting neutral feedback in sentiment analysis tasks. Overall, the results demonstrate that LR and SVM provided the most balanced and consistent performance across all sentiment categories. These findings suggest that these models are well-suited for analyzing pilgrims' feedback on food services during the Hajj, as they reliably distinguish between positive and negative sentiments.

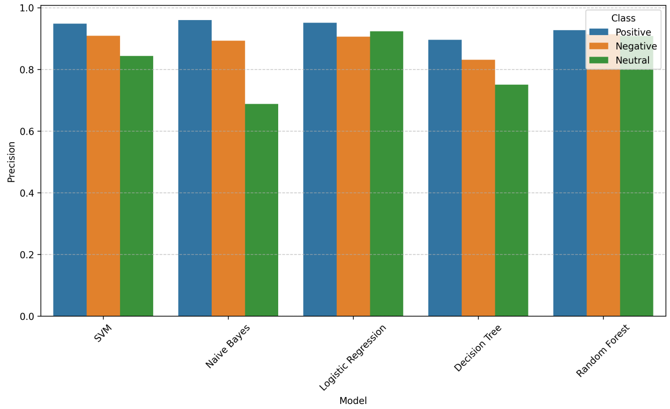

Fig. 8. Precision results comparing the performance of several machine learning models across sentiment categories.

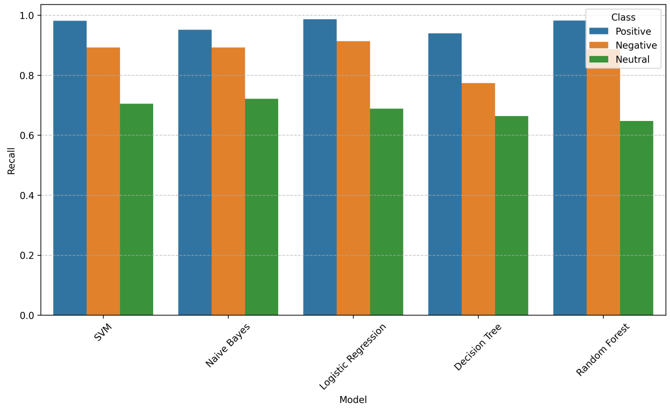

Fig. 9. Recall results comparing the performance of several machine learning models across sentiment categories.

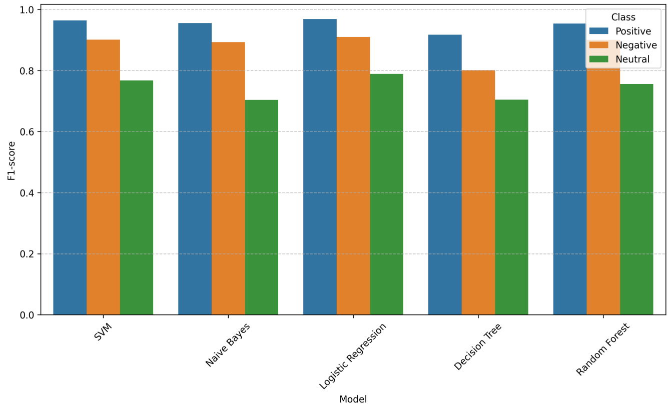

Fig. 10. F1-score results comparing the performance of several machine learning models across sentiment categories.

B. *Visualization*

Word frequency, word cloud, and confusion matrix—these visualization outputs contribute to data inspection and communication of results, supporting both preprocessing validation and model evaluation stages. Within Hajj-related sentiment analysis, authors have used these tools to display word-frequency patterns and illustrate sentiment trends in user-generated content on social platforms [6], [7], [8].

1) Word frequency: The analysis of word frequency in the textual data was conducted to identify the most common terms used by pilgrims when describing their experiences with the food services during the Hajj. As illustrated in Fig. 11, the word “الأكل” (food) appeared most frequently, with 1,315 occurrences, indicating that the majority of comments were directly related to the quality, availability, and overall experience of the food provided. The next most frequent word was “جداً” (very), which appeared 789 times, suggesting that many participants used intensifiers to emphasize their opinions, whether positive or negative. Other common words such as “ممتازة” (excellent), “الله” (God), “الحملة” (the campaign), and “البوفيه” (buffet) also appeared prominently, reflecting positive sentiments and references to the service settings. In addition, terms such as “خدمة” (service), “الطعام” (meal), and “الخدمات” (services) were frequently mentioned, showing that pilgrims often discussed both the food and the quality of service delivery. The presence of words like “رائعة” (wonderful) and “ممتاز” (excellent) further suggests that the general sentiment expressed in the comments was positive, aligning with the results of the sentiment analysis. Overall, this word-frequency analysis highlights the main topics of concern and appreciation among pilgrims—particularly food quality, buffet arrangements, and service efficiency—offering valuable insights for improving food service management during the Hajj season.

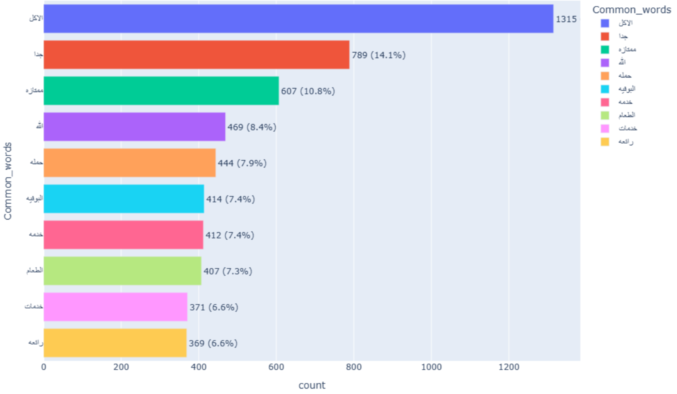

Fig. 11. Word frequency analysis showing the most common terms used by pilgrims in their feedback about food services during Hajj.

2) Word cloud: The next figure presents three Arabic word clouds categorized by sentiment: positive, negative, and neutral. Each word cloud visually displays the frequency of words appearing in pilgrim feedback, where larger words indicate higher occurrence.

Fig. 12 illustrates a word cloud of positive Arabic reviews, highlighting the most frequently used terms related to food services. In the visualization, larger words represent higher frequency. The most prominent terms include “الأكل” (food), “خدمة” (service), and “النظافة” (cleanliness), indicating that praise is most often directed toward these three aspects. Other common expressions consist of strong positive adjectives such as “ممتاز” (excellent), “رائع”(great), “متوفر” (available), and “تنظيم” (organization), along with intensifiers like “جداً” (very) and expressions of gratitude “شكراً” (thank you). Operational references, such as “البوفيه” (buffet), “كل شيء” (everything), “الموظفين” (staff), and “على مدار الساعة” (around the clock), also appear frequently, reflecting appreciation not only for food quality but also for efficient logistics and staff support. Overall, the word cloud underscores a dominant theme of positive sentiment centered on cleanliness, high-quality food, and attentive service.

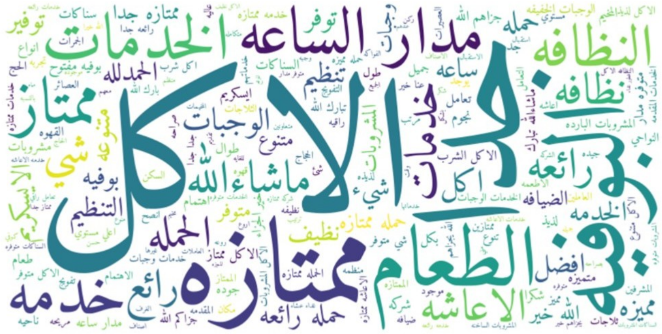

Fig. 12. Word cloud showing the most frequent words in positive Arabic reviews about food services in Hajj.

In the neutral word cloud, as shown in Fig. 13, terms such as “الطعام” (food), “جيد” (good), “الخدمات” (services), and “الجودة” (quality) appear most frequently, indicating that many reviews described the experience as satisfactory or acceptable. These comments primarily reflect practical assessments of meal quality, cleanliness, and organizational processes rather than emotionally driven opinions.

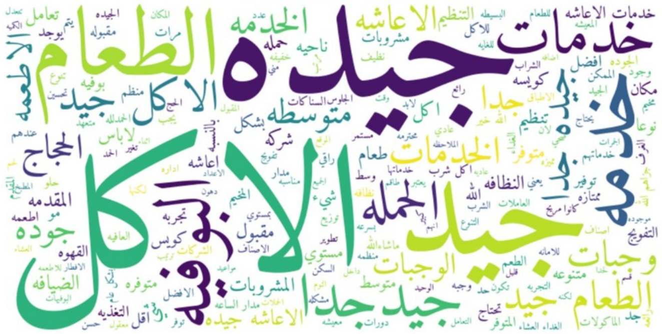

Fig. 13. Word cloud showing the most frequent words in neutral Arabic reviews about food services in Hajj.

The negative word cloud, presented Fig. 14, displays frequently used words in reviews expressing dissatisfaction with food services. The most noticeable terms, “الأكل” (food),

“سيئة” (bad/poor), “البوفيه” (buffet), “تنظيم” (organization), “لا يوجد” (not available), and “بارد” (cold), point to dissatisfaction related to meal quality, limited variety, and weak coordination. The prominence of these words highlights key areas of concern among pilgrims. Many complaints relate to food quality, temperature, availability, and organizational issues. Phrases like “لا يوجد تنوع” (no variety) and “الطعام بارد” (cold food) are indicative of operational and logistical shortcomings rather than fundamental dissatisfaction with the concept of the service itself. Moreover, the recurrence of words such as “الأسف” (unfortunately) and “المستوى” (standard) suggests unmet expectations, particularly given the significance of the Hajj pilgrimage and the challenging conditions under which meals are prepared and served. A closer analysis of negative sentiments also reveals several challenges faced by pilgrims, including the lack of specially prepared meals for individuals with chronic illnesses, such as diabetes or hypertension, as well as for those with specific dietary needs. Pilgrims also reported occasional delays in meal service and unexpected changes in serving times without prior notice. In addition, the absence of designated dining areas often forced them to eat in their rooms or in shared hallways. Overall, the visualizations clearly show that pilgrim opinions, whether positive, negative, or neutral, center around the same core themes: food quality, buffet service, and the overall service experience.

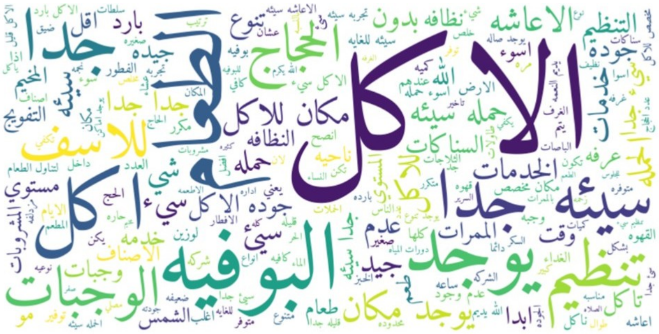

Fig. 14. Word cloud showing the most frequent words in negative Arabic reviews about food services in Hajj.

3) *Confusion matrix:* We evaluated the performance of the SVM, LR, NB, DT, and RF models using confusion matrices, as illustrated in Fig. 15, to analyze their classification accuracy across the positive, negative, and neutral sentiment classes. The diagonal values in each matrix represent correctly predicted instances, whereas the off-diagonal values indicate misclassifications. Each confusion matrix shows how well a model correctly classified instances across the positive, negative, and neutral categories. The results indicate that LR achieved the most balanced performance, correctly predicting 738 positive, 307 negative, and 84 neutral samples. The SVM model produced similar accuracy, with slightly fewer correct neutral predictions. The RF model performed well overall, although it exhibited minor misclassifications in the negative and neutral categories. NB also achieved strong results, with 712 positive, 300 negative, and 88 neutral correct predictions, while the DT model performed comparatively lower, especially in the negative class. Overall, LR and SVM provided the most stable and accurate sentiment classifications across all categories.

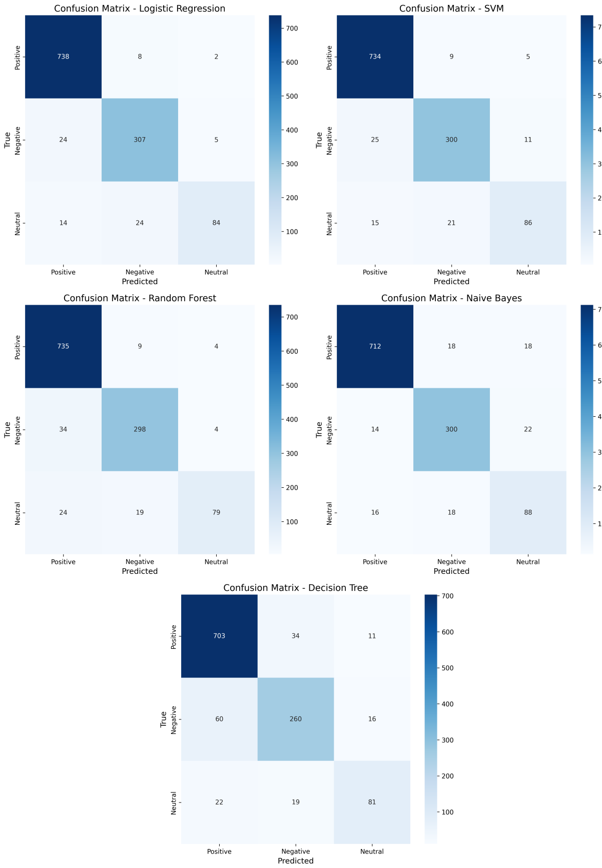

Fig. 15. Confusion matrices for five machine learning models: LR, SVM, RF, NB, and DT.

C. Hyperparameter Tuning

We performed hyperparameter tuning on the best-performing classification model obtained from previous evaluations. Based on the selected model type, we defined a specific set of parameters to optimize its performance, as shown in Table X. Each algorithm—SVM, LR, NB, DT, and RF—was assigned a distinct parameter grid containing key hyperparameters and their possible values. We used the GridSearchCV method with fivefold cross-validation to systematically evaluate all parameter combinations. The weighted F1-score was selected as the evaluation metric to effectively handle class imbalance. This process allowed us to identify the most suitable hyperparameter configuration for the chosen model, improving its overall accuracy and generalization.

D. Comparison with Previous Studies

Compared with previous studies, the present research provides a distinctive contribution by focusing specifically on food service quality during the Hajj, an area that has received limited attention in prior sentiment analysis works. While earlier studies, such as [8], [6], and [26], mainly analyzed general service sentiments on platforms like Twitter and YouTube, this study utilized Google Map reviews collected directly from Hajj campaigns between 2022 and 2025.

TABLE X. Hyperparameter Settings for Each Model

Model Type	Hyperparameters Tuned	Parameter Values Tested

SVM	Regularization parameter (C),
	Class weight
		C = [0.1, 1, 10, 100];
		class_weight = ['balanced', None]

NB	Smoothing parameter (alpha),
	Class prior fitting
		alpha = [0.1, 0.5, 1.0, 2.0];
		fit_prior = [True, False]

LR	Regularization (C),
	Penalty type,
	Solver,
	Class weight
		C = [0.1, 1, 10, 100];
		penalty = ['l1', 'l2'];
		solver = ['liblinear'];
		class_weight = ['balanced', None]

DT	Maximum depth,
	Minimum samples split,
	Minimum samples leaf,
	Class weight
		max_depth = [None, 5, 10, 15];
		min_samples_split = [2, 5, 10];
		min_samples_leaf = [1, 2, 4];
		class_weight = ['balanced', None]

RF	Number of estimators,
	Maximum depth,
	Minimum samples split,
	Minimum samples leaf,
	Class weight
		n_estimators = [100, 200, 300];
		max_depth = [None, 10, 20];
		min_samples_split = [2, 5, 10];
		min_samples_leaf = [1, 2, 4];
		class_weight = ['balanced', None]

This approach ensures that the dataset reflects pilgrims' authentic experiences with catering services. This approach provides a more targeted view of food-related satisfaction. Methodologically, while most earlier works (e.g., [7], [25], [28]) relied on deep learning models such as CNN, LSTM, or hybrid CNN-LSTM , achieving accuracies ranging from 92% to 98%, the present study employed traditional machine learning algorithms (SVM, LR, NB, DT, RF) combined with TF-IDF and n-grams for feature extraction. Despite being less computationally intensive, the LR classifier achieved an accuracy of 93.6%, demonstrating that high performance can be obtained with optimized classical methods when the data is domain-specific and linguistically normalized. Another key distinction lies in the language unification process. While several studies, such as [6] and [28], addressed multilingual datasets or mixed Arabic–English content, this research standardized all data into Arabic, ensuring consistency in sentiment polarity detection. Moreover, the scope of this study—covering a four-year data collection period (2022–2025)—extends beyond the shorter, event-based windows (e.g., a few weeks during Hajj 1442) used in earlier works, allowing for more stable and generalizable insights. Overall, this research bridges a critical gap by integrating machine learning–based sentiment analysis with a domain-specific assessment of food services during the Hajj, offering both methodological rigor and practical value for service management.

V. Conclusion

Through our analysis, we demonstrated that sentiment analysis is an effective method for understanding pilgrims' opinions regarding food services during the Hajj. We addressed a gap in the existing research, as previous work tended to rely on social media data and did not specifically focus on food services. In this study, we analyzed pilgrims' feedback using Arabic textual reviews collected from Google Maps, providing a direct and representative source of post-Hajj evaluations. Our primary objective was to assess the performance of five machine learning algorithms in classifying sentiments related to food services. To achieve this, we collected approximately 4,018 reviews across 160 Hajj campaigns conducted between 2022 and 2025. After applying preprocessing techniques and extracting features using TF-IDF with n-grams, we trained and evaluated the classifiers. In addition to model performance, we examined the key issues reported by pilgrims and analyzed sentiment distributions related to food service quality.

The results revealed that many pilgrims still face issues related to food, such as a lack of specially prepared meals for individuals with chronic illnesses, limited variety, and weak coordination, indicating that this remains an ongoing challenge. The application of machine learning classifiers showed excellent overall performance. Among the models, LR achieved the highest accuracy of 93.6%, outperforming the other classifiers. The SVM and RF models produced comparable results, with accuracies of 92.9% and 92.2%, respectively. Furthermore, sentiment analysis indicated that positive sentiments predominated across all years, from 2022 to 2025, reflecting an overall improvement in pilgrims' satisfaction with food services over time.

The limitations of this research primarily stem from data-related constraints. First, the volume of available reviews was not sufficient to support the training of more complex deep learning models. Second, the dataset is limited in scope, as it includes reviews from domestic Hajj campaigns only and therefore does not fully represent the diversity of pilgrims. Finally, some reviews contained mixed or ambiguous sentiments, complicating the classification process.

For future research, we intend to employ more advanced feature extraction techniques—such as word embeddings and contextual language models—to better capture semantic nuances in Arabic text. Furthermore, we plan to extend the scope of the analysis to cover other essential services provided during the Hajj, such as transportation and accommodation. Finally, exploring deep learning methods may further improve classification performance and provide deeper insights into sentiment patterns.

References

[1] Tan, K. L.; Lee, C. P.; Lim, K. M. A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. Applied Sciences 2023, 13 (7), 4550. https://doi.org/10.3390/app13074550

[2] Jouda, A. A.; Aziz, H. B. A.; Almasradi, R. B.; Alsharif, A. H. The Relationship Between Service Quality and Pilgrims Performers' Satisfaction: An Empirical Evidence from the Hotel's Industry in Saudi Arabia. Baltic Journal of Law & Politics, 2022, 15(1), 1129-1156.https://www.researchgate.net/publication/365687426

[3] Hussain, O.; Felemban, E.; Rehman, F. U. Optimization of the Mashaer Shuttle-Bus Service in Hajj: Arafat-Muzdalifah Case Study. Information 2021, 12 (12), 496. https://doi.org/10.3390/info12120496

[4] Aljohani, A.; Nejaim, S.; Khayyat, M.; Aboulola, O. E-Government and Logistical Health Services during Hajj Season. Bulletin of the National Research Centre 2022, 46 (1). https://doi.org/10.1186/s42269-022-00801-4

[5] Dhefaf Radain; Saliha Almalki; Almarghalani, S. A.; Elhag, S. Towards Achieving the 2030 Vision, the Case Study of Automating the Food Production Services during the Hajj Season and Quality Control Using the Blockchain Technology. The 5th International Conference on Future Networks & Distributed Systems 2021. https://doi.org/10.1145/3508072.3508096

[6] Alghamdi, H. M. Unveiling Sentiments: A Comprehensive Analysis of Arabic Hajj-Related Tweets from 2017–2022 Utilizing Advanced AI Models. Big data and cognitive computing 2024, 8 (1), 5–5. https://doi.org/10.3390/bdcc8010005

[7] Alasmari, A.; Farooqi, N.; Alotaibi, Y. Sentiment Analysis of Pilgrims Using CNN-LSTM Deep Learning Approach. PeerJ Computer Science 2024, 10, e2584. https://doi.org/10.7717/peerj-cs.2584

[8] Ottom, M. A.; Nahar, K. M. O. Social Media Sentiment Analysis: The Hajj Tweets Case Study. Journal of Computer Science 2021, 17 (3), 265–274. https://doi.org/10.3844/jcssp.2021.265.274

[9] Hajj — DataSaudi. Datasaudi.sa. https://datasaudi.sa/en/sector/hajj

[10] Ali, B. J.; Saleh, P. F.; Akoi, S.; Abdulrahman, A. A.; Muhamed, A. S.; Noori, H. N.; Anwar, G. Impact of Service Quality on the Customer Satisfaction: Case Study at Online Meeting Platforms. International Journal of Engineering, Business and Management 2021, 5 (2), 65–77. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3839031

[11] Udayalakshmi, C.; Sridevi, J. Service Quality Models: A Review with Respect to Fast Food Restaurants. BOHR International Journal of Social Science and Humanities Research 2023, 2 (1), 92–96. https://doi.org/10.54646/bijsshr.2023.30

[12] Ruyuf Alnafisah; Fahad Alnasiri; Saleh Alzaharni; Ibrahim Alshikhi; Alqahtani, A. Food Safety Practices during Hajj: On-Site Inspections of Food-Serving Establishments. Tropical Medicine and Infectious Disease 2023, 8 (10), 480–480. https://doi.org/10.3390/tropicalmed8100480

[13] Turkistani, A. M. S. The Special Dietary Needs of Pilgrims and Practices of Agencies Regarding Food Quality and Safety during the Hajj. Biosciences Biotechnology Research Asia 2022, 19 (3), 757–766. https://doi.org/10.13005/bbra/3028

[14] Yezli, S.; Yassin, Y.; Mushi, A.; Aburas, A.; Alabdullatif, L.; Alburayh, M.; Khan, A. Gastrointestinal Symptoms and Knowledge and Practice of Pilgrims Regarding Food and Water Safety during the 2019 Hajj Mass Gathering. BMC Public Health 2021, 21, 1288. https://doi.org/10.1186/s12889-021-11381-9

[15] Msrahi, N.; Arwa Turkistani. Assessment the Association between Foodborne Disease and Food Safety Knowledge and Practices among Pilgrims during Hajj in Saudi Arabia. Journal of Health Population and Nutrition 2025, 44 (1). https://doi.org/10.1186/s41043-025-00964-6

[16] Alfiah; A.; Hapsari; R. A. F.; Wulandari; E. Analysis of Food consumed by Indonesian Moslem Hajj Pilgrims: Complete and Incomplete Nutrition. The Avicenna Medical Journal 2023, 4(2), 31-40. https://journal2.uinjkt.ac.id/index.php/amedj/article/view/31740CrossRef

[17] Masoud Mohammed, R. Y.; Ahmed Zaid, W. M.; Bahurmoz, Prof. A. Enhancing Hajj Pilgrim Satisfaction: A Strategic Analysis of Service Quality Dimensions Using the Analytic Hierarchy Process in Alignment with Saudi Vision 2030. Global Journal of Management and Business Research 2024, 1–23. https://doi.org/10.34257/GJMBRAVOL24IS3PG1

[18] HASSAN, T. H.; ABDOU, A. H.; ABDELMOATY, M. A.; NOR-EL-DEEN, M.; SALEM, A. E. THE IMPACT of RELIGIOUS TOURISTS' SATISFACTION with HAJJ SERVICES on THEIR EXPERIENCE at the SACRED PLACES in SAUDI ARABIA. GeoJournal of Tourism and Geosites 2022, 43 (3), 1013–1021. https://doi.org/10.30892/gtg.43321-915

[19] Owaidah, A.; Olaru, D.; Bennamoun, M.; Sohel, F.; Khan, N. Transport of Pilgrims during Hajj: Evidence from a Discrete Event Simulation Study. PLOS ONE 2023, 18 (6), e0286460. https://doi.org/10.1371/journal.pone.0286460

[20] Weber, L.; Müller, S.; Haase, K. Pilgrims' Satisfaction with Metro Operations during Hajj. SSRN Electronic Journal 2022. https://doi.org/10.1007/s12469-023-00323-w

[21] Almujibah, H. R.; Nistorescu, C. G. Analysis of Public Transportation System in Makkah: Evaluation of Efficiency, Efficacity and Safety Operation of Almashaaer al Mugaddassah Metro (MMMSL). World Journal of Advanced Science and Technology 2022, 2 (1), 011–021. https://doi.org/10.53346/wjast.2022.2.1.0038

[22] BUSINESS INTELLIGENCE and ANALYTICS.System for Decision Support; RAMESH SHARDA; DURSUN DELEN; EFRAIM TURBAN; TENTH EDITION; http://seu1.org/files/level8/IT445/IT445%20BOOK%20EDIT.pdf.

[23] Dake, D. K.; Gyimah, E. Using Sentiment Analysis to Evaluate Qualitative Students' Responses. Education and Information Technologies 2022. https://doi.org/10.1007/s10639-022-11349-1

[24] Giang, N. T. P.; Dien, T. T.; Khoa, T. T. M. Sentiment Analysis for University Students' Feedback. Advances in Intelligent Systems and Computing 2020, 55–66. https://doi.org/10.1007/978-3-030-39442-4_5

[25] Mohd Khaled Shambour. Analyzing Perceptions of a Global Event Using CNN-LSTM Deep Learning Approach: The Case of Hajj 1442 (2021). PeerJ. Computer science 2022, 8, e1087–e1087. https://doi.org/10.7717/peerj-cs.1087

[26] Adnan Gutub; Mohd Khaled Shambour; Abu-Hashem, M. A. Coronavirus Impact on Human Feelings during 2021 Hajj Season via Deep Learning Critical Twitter Analysis. Journal of Engineering Research 2023, 11 (1), 100001–100001. https://doi.org/10.1016/j.jer.2023.100001

[27] Albahar, M.; Gazzawe, F.; Thanoon, M.; Albahr, A. Exploring Hajj Pilgrim Satisfaction with Hospitality Services through Expectation-Confirmation Theory and Deep Learning. Heliyon 2023, 9 (11), e22192. https://doi.org/10.1016/j.heliyon.2023.e22192

[28] Samia Allaoua Chelloug; Saleh, M.; Jamil, F.; Mehdhar S. A. M. Al-Gaashani; Soha Alhelaly; Aziz, A.; Ammar Muthanna. Enhancing Hajj and Umrah Services through Predictive Social Media Classification. IEEE Access 2025, 13, 67220–67238. https://doi.org/10.1109/ACCESS.2025.3559204

[29] Rahim, A. I. A.; Ibrahim, M. I.; Chua, S.-L.; Musa, K. I. Hospital Facebook Reviews Analysis Using a Machine Learning Sentiment Analyzer and Quality Classifier. Healthcare 2021, 9 (12), 1679. https://doi.org/10.3390/healthcare9121679

[30] Dey, S.; Wasif, S.; Tonmoy, D. S.; Sultana, S.; Sarkar, J.; Dey, M. A Comparative Study of Support Vector Machine and Naive Bayes Classifier for Sentiment Analysis on Amazon Product Reviews. IEEE Xplore. https://doi.org/10.1109/IC3A48958.2020.233300

[31] Saad, A. I. Opinion Mining on US Airline Twitter Data Using Machine Learning Techniques. 2020 16th International Computer Engineering Conference (ICENCO) 2020. https://doi.org/10.1109/ICENCO49778.2020.9357390

[32] What is Sentiment Analysis? GeeksforGeeks. Available online:https://www.geeksforgeeks.org/machine-learning/what-is-sentiment-analysis/. (accessed on 2025-07-12).

[33] Dataset for Sentiment Analysis. GeeksforGeeks. Available online: https://www.geeksforgeeks.org/nlp/dataset-for-sentiment-analysis/#1-imdb-reviews-dataset (accessed 2024-05-23).

[34] Outscraper - get any public data from the internet. outscraper.com. https://outscraper.com/.

[35] Nafea, A. A.; Muayad, M. S.; Majeed, R. R.; Ali, A.; Bashaddadh, O. M.; Khalaf, M. A.; Sami, A. B. N.; Steiti, A. A Brief Review on Preprocessing Text in Arabic Language Dataset: Techniques and Challenges. Babylonian Journal of Artificial Intelligence 2024, 2024, 46–53. https://doi.org/10.58496/BJAI/2024/007

[36] NLTK NLP. GeeksforGeeks. Available online: https://www.geeksforgeeks.org/python/NLTK-NLP/. (accessed on 2025-08-08).

[37] Musleh, D. A.; Alkhwaja, I.; Alkhwaja, A.; Alghamdi, M.; Abahussain, H.; Alfawaz, F.; Min-Allah, N.; Abdulqader, M. M. Arabic Sentiment Analysis of YouTube Comments: NLP-Based Machine Learning Approaches for Content Evaluation. Big Data and Cognitive Computing 2023, 7 (3), 127. https://doi.org/10.3390/bdcc7030127

[38] Rezwanul, M.; Ali, A.; Rahman, A. Sentiment Analysis on Twitter Data Using KNN and SVM. International Journal of Advanced Computer Science and Applications 2017, 8 (6). https://doi.org/10.14569/IJACSA.2017.080603

[39] Almaqtari, H.; Zeng, F.; Mohammed, A. Enhancing Arabic Sentiment Analysis of Consumer Reviews: Machine Learning and Deep Learning Methods Based on NLP. Algorithms 2024, 17 (11), 495. https://doi.org/10.3390/a17110495

[40] Machine Learning with Python Tutorial. GeeksforGeeks. Available online:https://www.geeksforgeeks.org/machine-learning/machine-learning-with-python/. (accessed on 2025-07-23).

[41] SVM vs KNN in Machine Learning. GeeksforGeeks. Available online: https://www.geeksforgeeks.org/svm-vs-knn-in-machine-learning/?ref=header_outind (accessed 2024-07-17).

[42] Naive Bayes Classifiers. GeeksforGeeks. Available online: https://www.geeksforgeeks.org/naive-bayes-classifiers/?ref=header_outind. (accessed on 2024-07-10).

[43] Logistic Regression in Machine Learning. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/understanding-logistic-regression/ (accessed on 2024-06-20)

[44] Decision Tree in Machine Learning. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/decision-tree-introduction-example/(accessed 2025-08-04).

[45] Random Forest Algorithm in Machine Learning. GeeksforGeeks. Available online: https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/?ref=header_outind. (accessed on 2024-07-12).

[46] Google. Classification: Accuracy, recall, precision, and related metrics. Google for Developers. https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall.

[47] ML Handling Imbalanced Data with SMOTE and Near Miss Algorithm in Python. GeeksforGeeks. Available online: https://www.geeksforgeeks.org/machine-learning/ml-handling-imbalanced-data-with-smote-and-near-miss-algorithm-in-python/. (accessed on 2025-07-12).