big data and cognitive computing

MDPI

Article

Unveiling Sentiments: A Comprehensive Analysis of Arabic Hajj-Related Tweets from 2017–2022 Utilizing Advanced AI Models

Hanan M. Alghamdi

Department of Computers, College of Engineering and Computing Al Qunfidhah, Umm Al-Qura University, Mecca 24382, Saudi Arabia; hmhghamdi@uqu.edu.sa

Abstract: Sentiment analysis plays a crucial role in understanding public opinion and social media trends. It involves analyzing the emotional tone and polarity of a given text. When applied to Arabic text, this task becomes particularly challenging due to the language's complex morphology, right-to-left script, and intricate nuances in expressing emotions. Social media has emerged as a powerful platform for individuals to express their sentiments, especially regarding religious and cultural events. Consequently, studying sentiment analysis in the context of Hajj has become a captivating subject. This research paper presents a comprehensive sentiment analysis of tweets discussing the annual Hajj pilgrimage over a six-year period. By employing a combination of machine learning and deep learning models, this study successfully conducted sentiment analysis on a sizable dataset consisting of Arabic tweets. The process involves pre-processing, feature extraction, and sentiment classification. The objective was to uncover the prevailing sentiments associated with Hajj over different years, before, during, and after each Hajj event. Importantly, the results presented in this study highlight that BERT, an advanced transformer-based model, outperformed other models in accurately classifying sentiment. This underscores its effectiveness in capturing the complexities inherent in Arabic text.

Keywords: sentiment analysis; Hajj tweets; machine learning; deep learning

Citation: Alghamdi, H.M. Unveiling Sentiments: A Comprehensive Analysis of Arabic Hajj-Related Tweets from 2017–2022 Utilizing Advanced AI Models. *Big Data Cogn. Comput.* **2024**, *8*, 5. https://doi.org/10.3390/bdcc8010005

Academic Editors: Zuchao Li and Min Peng

Received: 13 November 2023
Revised: 26 December 2023
Accepted: 27 December 2023
Published: 2 January 2024

1. Introduction

Hajj is an obligatory rite of pilgrimage in Islam, which is one of the five fundamental tenets of the religion. Hajj is a religious obligation that Muslims must fulfill, which involves traveling to the city of Mecca in Saudi Arabia and carrying out a series of activities known as Hajj formalities. It is one of the religious obligations for Muslims specified in the last month of the Hijri Calendar. Around 3–4 million Muslims from 180 countries traveled to Mecca for this year's Hajj [1,2]. All Muslims who are physically and financially able to do so must perform Hajj at least once in their lifetime.

The proposed study holds great importance due to the Saudi government's commitment, as stated in its "Vision 2030" plan, to ensure the smooth execution of Umrah for 15 million Muslims annually [3]. The Ministry of Hajj and Umrah (www.haj.gov.sa/en/Home/Index (accessed on 20 December 2023)) makes significant efforts to facilitate and serve pilgrims better. It continuously upgrades the services provided by utilizing new technologies and coordinating with all relevant agencies to ensure a smooth Hajj and Umrah experience.

Organizations and businesses offer a range of services to pilgrims during Hajj, including housing, meals, transportation, and other amenities. Many service providers strive to gather feedback from pilgrims regarding the facilities provided and their overall satisfaction with them. They are actively seeking ways to enhance the quality of services for future pilgrimages.

Extensive research has been conducted to provide solutions and suggestions for the challenges faced during the Hajj season. One of them such as work done by [4] where they used computer vision and image processing to propose a missing-and-found computerized system. Another study suggested using the Internet of Things (IoT) [5] to build smart fire monitoring and detection application [6] increased the performance of an agent-based crowd to simulate pilgrim movement during the rites of Hajj to simulate large crowds and to predict whether the developed event plan is viable or not. A work by Bhuiyan et al. [7] proposed a comprehensive system for analyzing moving scene crowd videos using the Convolutional Neural Networks (CNN) model. It was used for Hajj applications and introduced a system for counting and estimating crowd density.

In their research, Bati [8], Ottom and Nahar [9] and Shambour [10] explored the utilization of big data tools, sentiment analysis techniques, and data visualization in relation to Twitter users' views on the Hajj. The author emphasizes the need to collect, monitor, and analyze feedback from customers in order to identify areas for improvement in terms of health and safety standards. Having access to these data is a great advantage for companies, as it helps them understand their customers better and serve them in the best possible way. The author outlined a set of steps for collecting and analyzing Hajj-related tweets but did not employ these methods in their research.

Researchers from around the world have proposed the use of IT solutions to address challenges in Hajj and Umrah [11]. These solutions involve utilizing smartphones and social media apps to offer personalized services and location-based assistance for pilgrims. Additionally, other research explores tracking and navigation, expert systems, and learning to provide real-time answers to the questions of Hajjis. In addition, technology is being used to digitize the Hajj experience. This involves tasks such as document management, disease prevention measures, crowd control, intelligent transportation systems, and traffic simulation. These technologies aim to improve the infrastructure of Hajj sites, while other research focuses on the literature that addresses cybersecurity and privacy concerns related to Hajj. Specifically, it explores the protection of pilgrims' data.

Sentiment analysis of texts related to the Hajj can be beneficial in understanding the unique characteristics and attractions associated with it. Texts found on social media about the services provided during this annual pilgrimage can also be analyzed for their positivity or negativity. Social media platforms can be utilized to observe, capture, and analyze individuals' and communities' opinions, perceptions, and feelings. This data provides a valuable source of insights that decision-makers and officials can use to develop more informed decisions [12]. Sentiment analysis of texts related to the Hajj pilgrimage can give us valuable insight into the services offered at this event.

Social media is an essential platform for Muslims to connect and dialogue with each other across the globe. Communities on Twitter have notably taken up the responsibility of providing support and advice to fellow Muslims who are embarking on Hajj, including information related to worship, before, during, and after their journey. Social media has an impact beyond the direct reach of its users. The open dialogue nature of social media enables people to gain new insights into various topics, both generally and specifically. Social media has not only allowed Muslims to connect with each other across the globe but also helped them gain a better understanding of what is happening in their communities at home and abroad.

Arabic is the sixth most spoken language in the world and is used by more than 200 million people across the world [13]. Diacritical marks (Harakat) are Arabic language symbols, which are utilized to distinguish between words that have similar spelling to represent vowel sounds [14].

Arabic Sentiment Analysis (ASA) approaches can be classified into three main categories: corpus-based (supervised, unsupervised, and hybrid learning), lexicon-based (unsupervised learning), and hybrid- based [15]. Corpus-based techniques seek to leverage a large amount of Twitter data available to construct machine learning models in order to accurately classify sentiments associated with individual tweets. The lexicon-based ap-

proach uses sentiment lexicons to identify and classify words or phrases in a text as having positive, negative, or neutral sentiments. The sentiment value of each word or phrase is then used to calculate the overall sentiment of the text. Nevertheless, a hybrid-based process is a combination of lexicon and corpus techniques for ASA. In the proposed approach to evaluating tweet sentiment, this paper employs a lexicon-based method. It utilizes an existing word list, where each word is linked to a specific sentiment. Our lexicon-based strategy follows a fundamental sequence, starting with the initial preprocessing of each tweet. It initializes an initial polarity score (P) to zero. It then examines each token in the text, checking if it corresponds to an entry in the sentiment dictionary, and assigns a polarity value to P accordingly. If the total polarity score surpasses 0, we classify the text as positive. Conversely, if it falls below 0, it interprets the text as negative. If the score equals 0, it labels the text as neutral.

In a review of sentiment analysis research in Arabic language, Oueslati et al. [16] found that, for Arabic, the most used algorithms are NB and SVM. Deep learning is still in the early stages of exploration for Arabic sentiment analysis and has not yet been used as much as for English sentiment analysis.

The primary objective of this research is to conduct sentiment analysis studies on Twitter texts written about Hajj spanning a period of six years. By analyzing people's opinions, the Ministry of Hajj and Umrah aims to gain valuable insights into how individuals cope with the current circumstances. This heightened level of awareness can significantly assist the Ministry in enhancing the services provided to pilgrims, ultimately improving their overall experience and efficiency. To achieve this research goal, an Arabic tweet dataset was obtained from Twitter using targeted keywords focused on Hajj across various years before, during, and after each Hajj event. The collected tweets underwent several pre-processing and feature extraction techniques to appropriately prepare them for analysis. Different machine learning and deep learning techniques were implemented to analyze these tweets, followed by evaluating and contrasting these distinct approaches with each other. By conducting sentiment analysis on a vast volume of Twitter texts over an extended timeframe, this research endeavors to provide valuable information that can contribute towards informed decision-making within the Ministry of Hajj and Umrah. The contributions of the presented work can be summarized as follows:

- To collect a Twitter dataset in Arabic text related to the Hajj in different phases (before, after, and during) over six years and label them as positive, negative, or neutral.
- To study various Arabic sentiment analyzers. This analyzer first identifies the subjective text obtained from Twitter, preprocesses it, and then determines the polarity of the subjective text. This is followed by different feature extraction methods, such as unigram and bigram TF-IDF weighting, Bag-of-Words (BOW), or Word2Vec word embedding. After that, different machine learning classifiers were employed, including Logistic Regression, Support Vector Machine (SVM), Naïve Bayes, Random Forest, XGBoost, and K-Nearest Neighbor (KNN).
- To introduce an architecture based on Deep Learning (DL) for Arabic Sentiment Analysis. The architecture utilizes different feature extraction methods, including unigram and bigram TF-IDF weighting, Bag-of-Words (BOW), and Word2Vec word embedding. Various models, such as CNN, LSTM, and Bert mini for Arabic, are also employed.
- To compare deep learning models, this paper will use convolutional neural networks (CNN) and long short-term memory (LSTM) models. These models will be trained with feature extraction techniques and will be evaluated exclusively using uncleaned tweets.
- To evaluate the classification procedure of models applied to Arabic tweets using various machine learning and deep learning methods, with consideration for evaluation metrics such as recall, precision, and F-measure.
- Investigate the sentiment orientation of textual features and emojis-based features in Arabic comments posted on Twitter before, during, and after the Hajj event.

- To identify key topics associated with positive and negative sentiment that can be utilized by the Ministry of Hajj and Umrah, as well as other Hajj service providers, to enhance the overall experience of pilgrims.

The paper is organized as follows: The Related Works in Section 2 provides a thorough examination of prior research conducted in the same field. The study objectives and sentiment analysis architecture used in this research are explained in the Method and Experiments in Section 3. The Results and Discussion in Section 4 then presents the findings obtained from the data analysis. The Conclusion (Section 5) of the study summarizes the findings and offers recommendations for future research endeavors. The concluding section of the research paper provides a concise summary of the main findings and presents insightful conclusions derived from the research outcomes. It also serves to propose potential areas for future research, emphasizing avenues that warrant further investigation and exploration.

2. Related Works

Hajj is the most popular yearly event that people discuss on social media sites, and sentiment analysis of Hajj-related tweets can help build strategic plans for improving the Hajj season and developing better services [17]. Sentiment analysis allows for a more in-depth understanding of what is happening during a particular time period. Sentiment analysis is also used to measure the acceptability of policies and ideas, which helps with the advancement of society [18].

Sentiment analysis is one topic of Natural Language Processing and is also considered a text mining technique that can be used to detect favorable and unfavorable opinions or feelings collected from various sources about a particular subject [18–20]. Sentiment analysis refers to the process of measuring the sentiment expressed in a text, with positive sentiment defined as positivity and negative sentiment defined as negativity. The goal of sentiment analysis is to determine whether a given piece of text contains more positive or negative sentiments. Sentiment analysis is used in a wide range of areas, from market research to legal proceedings [21].

Social media has emerged as a highly influential platform for freedom of speech and the expression of emotions. Popular social media platforms like Facebook, Twitter, Instagram, and YouTube provide individuals with the opportunity to freely express their emotions [22]. Several academic studies have extensively examined the sentiment analysis of Twitter messages, primarily due to the platform's large and diverse user base, which regularly expresses opinions on various topics [23]. Compared to other social media platforms, Twitter data are particularly valuable for extracting crucial insights during times of crises [24]. This platform serves as a reliable source of up-to-date and genuine information, as tweets are directly shared by individuals without any modification or bias. Moreover, the tweets of active users often provide insightful details about their whereabouts and travel experiences [22]. Compared to other platforms, Facebook and Instagram are perceived as semi-private, whereas Twitter is regarded as a more public platform. Furthermore, communication on Twitter is typically more dispersed than on Facebook and Instagram [25]. As a consequence of the substantial reduction in data accessibility via Instagram APIs and Facebook APIs, academic researchers are now faced with the challenge of effectively accessing this data source [26–30]. Fortunately, the Twitter API platform provides various endpoints that can be utilized for different purposes. For instance, the Twitter Streaming API allows for fetching real-time data, while the Twitter Search API enables the retrieval of past tweets [31]. The current method of handling YouTube Data API requests is inefficient and time-consuming [32]. To retrieve relevant reviews, filtering out irrelevant results is necessary. Since YouTube comments are posted in multiple languages, there is a need for a language detection API that can accurately identify and retain only those comments written in the desired language [33]. These findings highlight the limitations of using other digital platforms for sentiment analysis when compared to Twitter.

Several machine learning techniques have been developed for Sentiment Analysis from text content found on Twitter. Ottom and Nahar [9] conducted a study to find the most effective method for Sentiment Analysis using a dataset collected by the researchers during the 2020 pilgrimage season. The findings demonstrate that machine learning techniques work better than the lexicon approach in classifying and analyzing Hajj-related tweets. The Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Naïve Bayes (NB) are used as supervised algorithms for a machine-learning approach. The Text Blob analyzer is used as a lexicon-based approach, where all the words occurring in the corpus are assigned their respective meanings. The difficulty faced is the limited availability of a Hajj tweets corpus dataset.

A study done by [34] recently presented the ArHajj-21 dataset, which consists of more than 200,000 tweets and accompanying transmission networks associated with Hajj events in 2021. This marks the first time a dataset dedicated to Hajj activities has been made available for public access. ArHajj-21 was developed to support research on Arabic Hajj-related tweets in various aspects, such as social computing, machine learning, natural language processing, and information retrieval. These include retweeting and conversational threads for the most popular topics. Furthermore, the authors applied K-means clustering algorithms using TF-IDF, Word2Vec, and Doc2Vec document vectorizations, as well as Latent Dirichlet Allocation (LDA) and Gibbs Sampling for Dirichlet Multinomial Mixture model (GSDMM) to acquire useful information for decision-makers. Elgamal [35] proposed another sentiment analysis method for Hajj-related tweets in the English language to classify the type of each tweet into positive or negative classes. A Naïve Bayes classifier was applied together with N-gram feature selection.

With recent developments in deep learning algorithms, researchers have been able to explore and address various issues using this technology. One issue is the use of deep learning algorithms to gain insights into data sets and uncover hidden patterns. This study will examine some prior research using deep learning-based algorithms for analyzing Hajj-related tweets to tackle this issue.

Aldhubaib [12] explores the social aspects of Makkah during Hajj 1441 H, with a focus on how changes due to COVID-19 have impacted the community's lifestyle compared to previous years. Data from the Makkah community were collected through a questionnaire and Twitter. A CNN model was employed to identify features that the embeddings had not indicated, while a Long-Short Term Memory (LSTM) model identified the correlation between those extracted features. A CNN-LSTM deep learning technique was employed [10] in order to analyze the attitudes of pilgrims and others during the 1442 AH Hajj period. More than 11,000 Twitter and YouTube posts related to the Hajj were collected for this purpose. The findings from the research demonstrate a few intriguing insights concerning the activities that happen during the Hajj.

An analysis of 5 million tweets in Arabic and English showed that Hajj evoked different emotional reactions in tweeters, with positivity, negativity, and neutrality varying based on certain events during the pilgrimage [36]. These events were associated with spiritual and faith values that generated positive impressions. In contrast, the stampede in Mina, which caused some deaths, represents a negative impression. However, the study failed to detail how the data were classified or how sentiment analysis was conducted.

In the period of Dhul-Hijjah 1–13, 1442 (11–23 July 2021), a total of 4300 Hajj-related posts and interactions were identified on social media platforms such as Twitter and YouTube [10]. This includes tweets, comments, likes, shares, and other interactions related to the Hajj. These posts are expected to provide valuable insights into the Hajj experience of pilgrims and help to identify key trends and topics among Hajj-goers. This research investigates the opinions of Hajj pilgrims and people within and outside the Makkah community on various topics related to Hajj, especially in relation to the spread of COVID-19, which has led to a decrease in pilgrims and kept them in their homes. An investigation of the sentiment expressed within and outside the Makkah community during the 1442 Hajj season was conducted using a Convolutional Neural Network-Long Short-Term Memory

(CNN-LSTM) deep learning model. This model demonstrated more precise results than other models.

A CNN-LSTM deep learning model was applied to analyze 22,000 tweets from Makkah and Madinah, the two holy cities, demonstrating the value of social media in understanding public opinion [37]. The research examined the emotions of Hajj pilgrims in the context of COVID-19 health measures. An analysis of tweets from two cities revealed similarities and differences in the characteristics of their Twitter users. It appears that Twitter engagement with COVID-19 topics is influenced by external factors, such as news reports and events. An analysis of tweets from the Makkah regions and tweeters' sentiments provides further insights into how individuals view various situations and predict their future behavior.

Regarding the research conducted by ASA, there are noteworthy recent studies that deserve attention due to their various objectives and methodologies. One particular model that has gained attention is BERT, known for its advanced capabilities [31,38–40]. Notably, BERT has proven to be effective in pre-training the Arabic language, as demonstrated in the works referenced [31,38,40]. A notable study by the authors [31] examined the detection of hate speech in Arabic using the Arabic BERT-Mini Model (ABMM). This study yielded promising results, achieving an accuracy score of 0.986. The main objective of their research was to identify instances of hate speech, abuse, and normal language on Twitter, addressing the important issue of online toxicity and harmful language. There has been another research paper that explores the application of active learning in Arabic sentiment analysis tasks [40]. The authors applied an active learning approach to choose training data and utilized the MARBERTv2 pretrained model for generating sentence embeddings. AraBERT is utilized to represent the input data as embeddings [39]. The article introduces a deep-learning ensemble model for sentiment analysis, which consists of three base classifiers: Gated Recurrent Unit (GRU), LSTM, and Bidirectional LSTM (BiLSTM). The stacking ensemble model captures long-range dependencies in the embeddings for each class. Lastly, SVM serves as the meta-classifier, combining the predictions from stacking deep learning models. A recent study [41] undertakes an investigation into the effectiveness of ensemble learning in the field of ASA by combining two distinct deep learning classifiers. The researchers specifically employ the BiLSTM model and a Generative Pre-trained Transformers (GPT) model to conduct their analysis. Another study [42] developed BiLSTM language models to extract crucial semantic data from Arabic tweets and label them as Positive, Negative, or Neutral.

The C-Support Vector Classification (C-SVC) method is commonly used in classification tasks because of its flexibility in selecting kernel type and regularization strength [43]. In this study, the SVM Sentiment Analysis for Arabic Students' Course Reviews (SVM-SAA-SCR) algorithm was specifically designed to analyze course reviews written in Arabic. It employs the sophisticated technique of Support Vector Machines (SVM), which is further fine-tuned to optimize its performance. This fine-tuning involves adjusting parameters, such as kernel type and textual comments and grading-based feedback; the algorithm provides a more precise and comprehensive understanding of the sentiments expressed by students.

An extensive analysis was conducted to evaluate the sentiments expressed in Arabic YouTube comments across various videos [44]. This study utilized six specific supervised machine learning text classifiers: SVM, Naïve Bayes, Logistic Regression, KNN, Decision Tree, and Random Forest. In order to analyze the data, a range of N-grams and TF-IDF methods were employed to extract features. The comments in the dataset were then manually classified into "Positive" and "Negative" categories after preprocessing. It is clear that TF-IDF can enhance the performance of most classifiers utilized in this study, with the exception of Naïve Bayes. When utilizing only n-grams and a count-vectorizer, Naïve Bayes demonstrated satisfactory results.

Compared to other languages, such as English, there are few studies and research works on Arabic language sentiment analysis as shown in Table 1. When considering Hajj-related tweets, only a limited number of research papers focus on their analysis [45].

Table 1. Summary of Related Works for Tweet Sentiment Analysis.

Study	Model Used	Features Extraction Used	Dataset	Tweets Language
[9]	SVM, KNN and NB	BOW, N-Grams and TF-IDF	3175 tweets	English language
[10]	CNN-LSTM	Skip-gram model	2996 tweets	Not specified.
[31]	BERT-Mini Model	Word embedding	9352 tweets	Arabic
[34]	K-Means clustering, LDA and GSDMM	TF-IDF, word embeddings, and document embeddings	200 K tweets	Arabic
[35]	CNN model and LSTM model	Word embedding	45,000 tweets	Not specified.
[36]	Not specified	Not specified	5 million tweets	Arabic and English
[37]	CNN-LSTM	Skip-gram for word embeddings, (CNN)	More than 22,000 tweets	
[39]	GRU, LSTM, BiLSTM, and SVM	word embedding model (AraBERT)	60,000 tweets	Arabic
[42]	bi-LSTM	BOW	2500 tweets	Arabic

Sentiment analysis of Hajj-related tweets can help build strategic plans to improve the Hajj season and develop better services. Using sentiment analysis to discover the sentiment polarity of customers and pilgrims through Hajj-related tweets can help uncover gaps between pilgrims' expectations and the services delivered. It can help in monitoring the facility's performance and in understanding the main causes of negative customer sentiment scores about the services. Therefore, with the resulting sentiment analysis, organizations and business providers can enhance services to provide excellent customer experiences, as they gain a deeper understanding of customer feedback [26–28].

Despite its clear importance, the impressions of Hajj pilgrims on social media have yet to be deeply explored and understood. Moreover, there are few studies that provide satisfactory methods for analyzing user emotions and exploring how they relate to certain events during the Hajj period. There is a lack of research on Hajj-related topics, including an analysis of user sentiments expressed in Arabic. Examining user sentiments expressed in Arabic regarding Hajj is essential for understanding the perspectives of individuals engaging in one of the most influential social events worldwide. Conducting research on this subject can provide valuable insights into the thoughts and opinions of Hajj participants.

3. Proposed Method

This section presents the proposed methodology for developing a reliable system to analyze Arabic tweets related to Hajj. The approach consists of five phases: data collection, data cleaning and preprocessing, feature extraction, classification, and evaluation. The following paragraphs provide a detailed description of each stage in their respective subsections. Figure 1 provides a summary of the methodology.

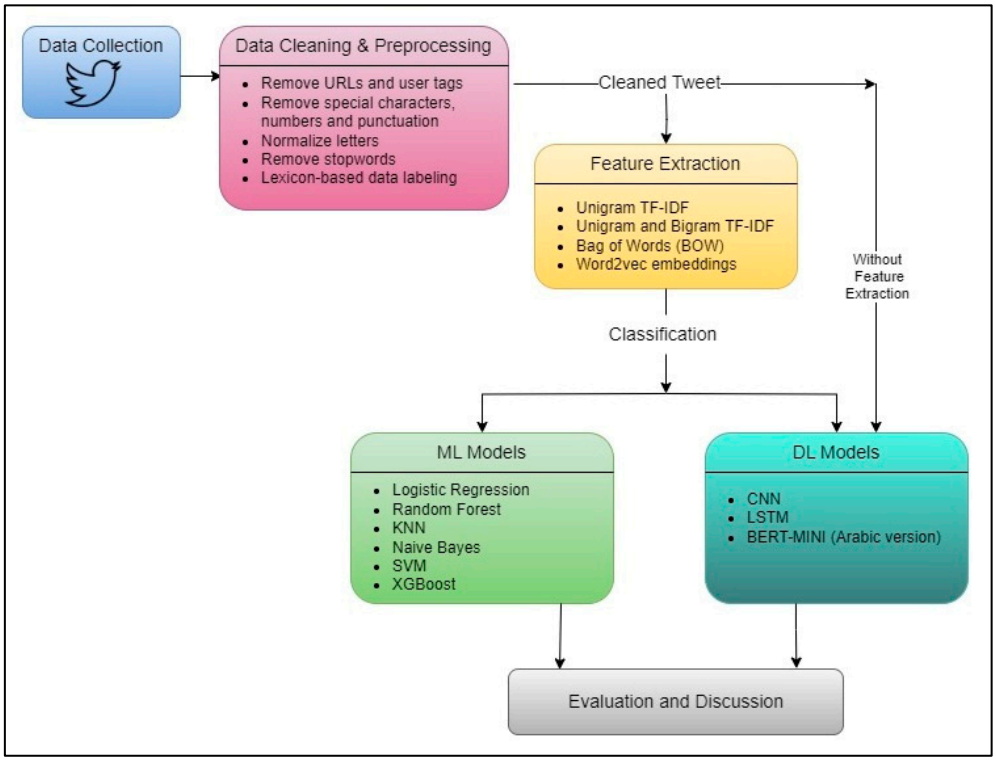


Figure 1. Proposed Methodology.

3.1. Data Collection

This research selected Python to retrieve Arabic tweets from Twitter's Search Application Programming Interface (API) [29,30], in order to build a dataset. The author opted for Python as the programming language to gather data due to its versatility, accessibility, and the availability of relevant libraries for working with the Twitter API. The decision not to use monitoring systems for a more comprehensive sample may have been influenced by our specific study constraints or objectives. For instance, the research aims to analyze a defined period or specific themes within Hajj-related tweets; periodic batch processing with Python has sufficed without the need for real-time monitoring systems. Resource constraints and the ease of implementation have further contributed to this choice, enabling a pragmatic and customized approach to building the dataset tailored to the study's goals.

To remain pertinent to our target audience, we started off by using hashtags related to the Hajj topic. The hashtags used for the search are #Hajj, #حج, #موسم_الحج, #الحج or some hashtags along with year of Hajj, such as #Hajj2021, #حج١٤٤٢, and #حج2021. These initial seed terms were used with Tweepy API to retrieve 86,682 related tweets and some tweets' information.

The aim of this collection was to allow for a longitudinal, continuous study of Hajj sentiment in Saudi Arabia. Data were gathered continuously for six years (2017–2022) over three different durations. The period from January to June was the first cycle before Hajj. During Hajj, which spanned from July to August, was the second cycle. The final cycle was in the months of September through November after Hajj had concluded. It is proposed that the duration of the study must span six years (2017–2022) in order to capture customer sentiment before, during, and after each Hajj event. The categories of the collected tweets were divided into three durations: before Hajj (January–June), during Hajj (July–August), and after the Hajj event (September–November).

This study was conducted to analyze customer sentiment across a range of Hajj events over several years in Saudi Arabia. Its results provide a better understanding of the status and customs surrounding Hajj, as well as of how customer behaviors have evolved throughout its duration.

The size of the dataset was over 80,000 tweets, as shown in Figure 2, which were in Arabic. This research used a large dataset size to build a classification model, then applied the model to study the sentiment, as recommended by [21]. The datasets that were collected align with the findings of previous studies. These studies have shown that datasets comprising more than 20,000 tweets are deemed adequate for the development of cutting-edge systems for Twitter Sentiment Analysis (SA) [46,47]. In this research, the author collected only the text of the tweets, along with their respective time and location data, without gathering any additional user-related information.

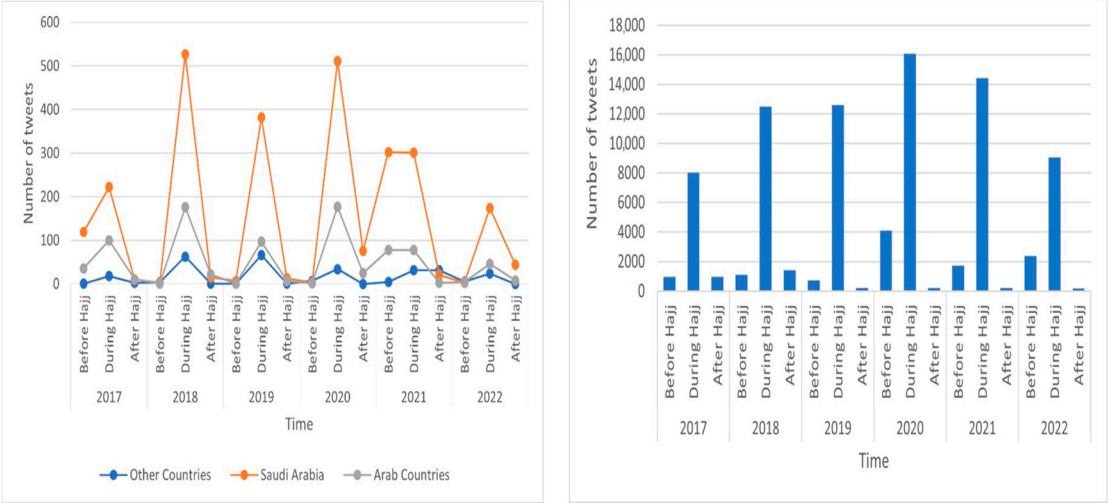


Figure 2. Number of tweets in the dataset over the years from different locations (on the **left**) and without locations provided (on the **right**).

3.2. Dataset Cleaning and Pre-Processing

To remove unwanted data from the dataset, a thorough cleaning process was conducted. One of the techniques employed for this purpose was the removal of spam, specifically tweets that included a Uniform Resource Locator (URL). As outlined by Alayba et al. [48], this particular strategy was adopted due to the prevalence of news or spam tweets containing URLs in the dataset. Furthermore, this paper followed the recommendations of [10,21] and Arabic language sentiment analysis by excluding repetitive information, such as retweets. In addition, to guarantee the classifier's precision and productivity, a dedicated language filter tailored to Arabic (lang: AR) was employed to eliminate non-Arabic tweets from the dataset. The decision to adopt this approach was made in favor of translation methods, as translations often have the potential to introduce errors that could negatively affect overall performance.

Before analysis, the datasets underwent pre-processing to eliminate unnecessary elements within tweets that could potentially undermine accuracy. This involved removing user mentions (e.g., "user"), numbers, special characters (such as +, =, ~$), and stop words (such as ",", ".", ";"). These actions were carried out in accordance with the suggestions provided by Aljabri et al. [21] and Oueslati et al. [16], For normalization and tokenization purposes, the datasets underwent processing utilizing the Natural Language Toolkit (NLTK) library in the Python programming language.

Subsequently, the tweets underwent tokenization, wherein sentences were divided into separate words to facilitate analysis. Lastly, the tweets were normalized. When dealing with Arabic text, normalization entailed the standardization of certain Arabic letters that may have had varying shapes, ensuring consistent representation [16,49,50]: Substitute the Arabic letters "إ", "أ", and "آ" with the bare alif "ا". Substitute the letters "ئ", "ئ", and "ئ" with the bare ya "ئ". Replace the final letter "ة" with its common form, which is also "ة".

In cases where a word begins with the letter "ء", it replaces with an initial alif, which is denoted as "أ". Lastly, replace instances of the letter "ؤ" with its common form "و".

Following the initial step, the data were labeled in order to implement supervised learning techniques. To assess the sentiment of a tweet, a lexicon-based approach was employed. This method utilizes a preexisting word list in Arabic language where each word is linked to a distinct sentiment. The specific dictionaries employed for sentiment analysis were taken from an open-source dictionary [51]. For this research, all collected tweets are considered true negative or positive based on dictionaries and polarity scores calculated. However, we will examine more on how to differentiate between true and fake positive and negative tweets for the next research. The lexicon-based approach adheres to a fundamental sequence wherein each tweet undergoes preprocessing initially. After that, the initial polarity score (P) is set to zero. The process then examines each token in the text to determine whether it can be found in the sentiment dictionary and assigns a polarity value to P accordingly. If the total polarity score ends up being greater than 0, the text is classified as positive. Conversely, if it is less than 0, the text is seen as negative. Should it equal 0, the text is labeled as neutral. Nevertheless, through experimentation, it was noted that including neutral tweets had an adverse impact on model training. As a result, neutral tweets were excluded from the dataset during the training process, and training was conducted solely on positive and negative tweets.

In this study, a conventional data splitting strategy was utilized. Specifically, 80% of the dataset was designated for training purposes, while the remaining 20% was set aside for testing previously unseen text. This approach played a crucial role in the methodology, as it facilitated a thorough assessment of the sentiment analysis model's performance. During the training phase, the model assimilated knowledge from most of the data, enabling it to understand and identify various nuances in Arabic sentiment expressions and their associated features. Through the evaluation of the model's performance on the 20% of data that has been intentionally withheld, it is possible to determine how effectively it can apply its acquired knowledge to unobserved text.

3.3. Feature Extraction

Within the domain of natural language processing and sentiment analysis, the careful selection of feature extraction methods plays a pivotal role in capturing significant information from textual data. This section delves into the application of several techniques for extracting features from Arabic text, such as Term Frequency-Inverse Document Frequency (TF-IDF) features (both Unigram and Unigram-Bigram), Bag of Words (BoW), and Word2Vec embeddings.

3.3.1. Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF is a well-established and efficient technique for feature extraction that highlights the significance of words in a group of documents. In the context of Arabic text analysis, TF-IDF can be leveraged to extract features by considering the frequencies of Unigrams (single words) and Bigrams (pairs of consecutive words) [34,42]. By utilizing TF-IDF features based on Unigrams alone, valuable insights can be obtained regarding the importance of individual words. However, incorporating bigrams into the model enhances its ability to capture meaningful word associations and specific Arabic phrases. In this paper, TF was calculated as described in Equation (1).

$$TF = \frac{number\ of\ times\ the\ term\ appears\ in\ the\ document}{total\ number\ of\ terms\ in\ the\ document} \quad (1)$$

The IDF of a term reflects the proportion of documents in the corpus that contain the term. Words unique to a small percentage of documents receive higher importance values than words common across all documents. IDF is calculated as described in Equation (2).

$$IDF = \log(\frac{number\ of\ documents\ in\ the\ corpus}{number\ of\ documents\ in\ the\ corpus\ containing\ the\ term}) \quad (2)$$

The TF-IDF of a term is calculated by multiplying TF and IDF scores. This approach proves valuable in recognizing the importance of specific Arabic terms in expressing sentiment, particularly when considering the nuances of Arabic linguistics.

3.3.2. Bag of Words (BoW)

It is a simple yet powerful method for extracting features in text analysis. It involves creating a vocabulary of unique words and counting their occurrences in a document. In the context of Arabic sentiment analysis, BoW enables the construction of a feature vector that captures the frequency of each word without considering word order or syntax [42]. Although BoW simplifies the text to a matrix of word counts, it can effectively identify sentiment-related terms in Arabic text. To extract the bag of words, this paper first defines our vocabulary, which is the set of all words found in our document set. After that, count how many times each word appears to vectorize our documents. Consequently, these vectors feed into the proposed classification model. Notice that we lose contextual information, e.g., where in the document the word appeared when we use BOW. It is like a literal bag-of-words: it only tells you what words occur in the document, not where they occurred.

3.3.3. Word2Vec Embeddings

Word2Vec is an approach based on neural networks that allows for the learning of distributed representations of words. It transforms words into continuous vector representations where words with similar meanings are mapped to neighboring points in the vector space. In the case of Arabic text, Word2Vec embeddings can effectively capture the semantic relationships between words, enabling the model to comprehend the context and nuances in sentiment expression. Arabic, known for its rich and context-specific vocabulary, utilizes Word2Vec embeddings to provide a concise representation in vector space. This representation preserves the meanings and relationships between words, allowing sentiment analysis to be conducted with a greater awareness of the context [34]. To learn the representations from Arabic words, the continuous skip-gram model is used to predict words within a certain range before and after the current word in the same sentence. While a bag-of-words model predicts a word given the neighboring context, a skip-gram model predicts the context (or neighbors) of a word given the word itself. The model is trained on skip-grams, which are n-grams that allow tokens to be skipped. The context of a word can be represented through a set of skip-gram pairs of (target_word, context_word) where context_word appears in the neighboring context of target_word. The training objective of the skip-gram model is to maximize the probability of predicting context words given the target word. For a sequence of words $w_1, w_2, \ldots w_t$, the objective can be written as the average log probability, as described in Equation (3).

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c}\log(p(w_t + j|w_t)) \quad (3)$$

where c is the size of the training context. The basic skip-gram formulation defines this probability using the softmax function. These embeddings learned through word2vec have proven to be successful on a variety of downstream natural language processing tasks.

To summarize, when it comes to conducting Arabic text analysis, the choice of a feature extraction method is contingent upon the specific objectives of the task. TF-IDF features, Bag-of-Words (BoW), and Word2Vec embeddings each offer unique benefits. TF-IDF allows for the evaluation of word and phrase significance, BoW simplifies text representation by

highlighting word frequencies, and Word2Vec embeddings capture semantic relationships among words. The choice of an appropriate method is contingent upon the complexity of the sentiment analysis task and the linguistic characteristics inherent in Arabic text. These factors play a crucial role in determining the model's capability to accurately capture the nuanced emotions portrayed by language.

3.4. Classification Models

The classification model in this study referred to a machine learning or deep learning algorithm. Its purpose was to categorize data into different classes or categories based on predetermined features or attributes. These models were trained using labeled data, where the class or category of each data point is already known. Subsequently, they were employed to predict the class or category for new and unseen data. Classification models are extensively utilized across various industries, such as finance, healthcare, marketing, and customer service. Their use enables accurate predictions and facilitates the streamlining of decision-making processes. The subsequent sections elucidate the algorithms employed in this study, which are derived from machine learning and deep learning models.

3.4.1. Machine Learning Models

In this section, various machine learning classification algorithms for sentiment analysis are discussed, specifically focusing on their applicability to Arabic text sentiment classification. The algorithms examined included Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Naive Bayes, and XGBoost.

- Logistic Regression

We employed the Logistic Regression algorithm, a cornerstone in linear classification widely utilized across diverse domains. Specifically, within sentiment analysis, we harnessed the algorithm's capabilities to effectively model the intricate relationship between text-based features and sentiment class labels encompassing positive or negative sentiments. The model incorporates a sigmoid activation function, represented as $\sigma(z)$, where z is the linear combination of input features and their corresponding weights. The probability of a given text belonging to a specific sentiment class is then determined using the logistic function.

- Random Forest

The Random Forest technique is a type of ensemble learning method that leverages the strength of multiple decision trees to make accurate predictions. It is highly regarded for its resilience and ability to handle intricate relationships between features. In the context of Arabic sentiment analysis, Random Forest proved to be quite effective in capturing non-linear patterns within the dataset and managing a vast number of features [42]. This approach becomes particularly valuable when dealing with sentiment classes that are not easily distinguished using linear models. In the experiments, we set the number of decision trees in the forest to 100 and used the Gini index to measure the quality of the split. We used bootstrap aggregation to randomly select subsets of the whole dataset and also random subsets of the features. We ran each experiment 10 times and reported the average accuracy.

- K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) algorithm is widely recognized as a straightforward yet efficient method for classification tasks. In this approach, a data point's classification is determined by the majority class of its k-nearest neighbors. When applying KNN to Arabic text sentiment analysis, it is crucial to utilize similarity metrics that account for the unique characteristics of the language. The algorithm calculates the distance between a new data point and existing data points using a chosen metric—in our case, Euclidean distance. The parameter 'k' denotes the number of nearest neighbors to consider, which we calculated to be 3 using the elbow method. The predicted class (C) is the one with the highest count

among the neighbors. This algorithm is intuitive, though sensitive to outliers, and does not require model training, as it relies on the stored dataset for predictions.

- Support Vector Machines (SVM)

It is a robust tool for accomplishing binary classification tasks. They aim to determine the most optimal hyperplane that successfully distinguishes data points belonging to distinct classes. In the field of Arabic sentiment analysis, SVMs can be effectively applied by selecting a suitable kernel function that captures the inherent non-linear characteristics of the data [42]. We've employed a polynomial kernel function. In addition, we scale this function by a coefficient of (1/number of features). We also added a regularization parameter to prevent overfitting to the training data and help the model generalize over the unseen text.

- Naive Bayes

It is a classification method that utilizes Bayes' theorem and probabilities. It is renowned for its computational efficiency and is especially valuable for text classification tasks. In the context of Arabic sentiment analysis, Naive Bayes can aid in estimating the conditional probability of sentiment classes by examining the presence of certain words or features in the text [42]. We use Gaussian probability distribution to describe the distribution of the classes and assume the independence of features. However, it is important to acknowledge that the assumption of independence between features may not always be valid for all forms of sentiment expressions in Arabic.

- XGBoost

It is a widely recognized gradient boosting algorithm that is highly regarded for its impressive performance and versatility. It excels in effectively addressing classification and regression tasks with equal proficiency. When applied to Arabic sentiment analysis, XGBoost can be employed to create a collection of decision trees that effectively capture complex relationships within the data. We set the number of estimators to be 100, the learning rate to 0.1, and the maximum depth of each tree to be 3. These numbers were selected based on experiments using our dataset.

3.4.2. Deep Learning Models

In recent times, deep learning models have gained significant recognition in the field of sentiment analysis due to their outstanding performance in various natural language processing tasks. This section will explore the utilization of Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and Bidirectional Encoder Representations from Transformers (BERT) for conducting sentiment analysis specifically on Arabic text.

- Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs), initially created for image analysis, have been successfully utilized in text classification assignments, such as sentiment analysis. CNNs possess the ability to comprehend localized patterns within data by means of convolutional layers and pooling operations. In the realm of Arabic sentiment analysis, CNNs prove instrumental in automatically extracting pertinent features and patterns from textual content. This enables them to effectively capture local word dependencies and decipher their importance in determining sentiment polarity. This approach proves to be exceptionally beneficial when dealing with the Arabic language, which is characterized by intricate linguistic structures and variations in word order [37]. We construct a CNN consisting of 2 convolutional layers, each one followed by a max pooling layer. These blocks are followed by a flattening layer, then a dense layer containing 8 neurons. The output layer is attached with one neuron to output the predicted class. The activation functions are set to be Rectified Linear Unit (ReLU) in all layers except the output layer, in which we use Sigmoid activation function to predict the probability of each sample being positive or negative.

- Long Short-Term Memory (LSTM)

LSTM networks, a type of recurrent neural network (RNN), are specifically engineered to capture extensive dependencies in sequential data. This feature makes them particularly well suited for text analysis tasks. When applied to the analysis of Arabic text, which often relies on intricate contextual cues and historical information for sentiment inference, LSTMs demonstrate noteworthy usefulness. These models excel at retaining and leveraging previous information, thereby enabling them to accurately capture nuanced emotions and sentiments expressed within the text. This is especially valuable when it comes to comprehending sentiment in the Arabic language, as it relies heavily on contextual cues and the intricate relationship among words within sentences [42]. The model we employed consisted of an embedding layer to represent the embeddings of text, followed by a bidirectional LSTM of 64 units. After that, we added a dense layer of 24 neurons followed by an output layer of 1 neuron and a sigmoid activation function to predict the probability of the output class.

- BERT Model

The BERT model, which is a transformer-based architecture that has been pre-trained, has brought about significant changes in the domain of natural language processing. Its ability to understand context bidirectionally makes it an exceptional option for Arabic sentiment analysis. By leveraging a BERT model that has undergone pre-training and fine-tuning on Arabic text, one can effectively utilize its extensive contextual comprehension to accurately interpret intricate sentiments. BERT models possess the capability to handle the subtleties and nuances present in the Arabic language, thereby making them exceedingly effective at capturing complex features of sentiment expressed in Arabic text [31]. We utilize the MINI version of the Arabic BERT model, which consists of 8 encoder layers followed by the decoder followed by the fully connected layer. The model was pre-trained on 8.2 billion Arabic words corpus and then fine-tuned on our Arabic tweets dataset. BERT undergoes pre-training through two interrelated NLP tasks, capitalizing on its bidirectional capability: predicting the next sentence and masked language modeling. When examining languages akin to human languages, BERT demonstrates heightened accuracy in grappling with uncertainty, a formidable aspect of natural language analysis. In contrast to the conventional word2vec embedding layer, which yields static, context-independent word vectors, BERT's layers generate dynamic, context-dependent word embeddings. This is achieved by considering the entire sentence as input and extracting information comprehensively. Leveraging a self-attention mechanism, BERT can concurrently process multiple tokens, necessitating specialized embedded tokens for the next sentence prediction challenge. Transformer encoders read the entire sequence of phrases at once, departing from the sequential left-to-right or right-to-left approaches. While bidirectional, it is more aptly described as non-directional. The model discerns a word's context by considering its surroundings on both sides, allowing it to derive contextual meaning based on the entirety of the input [31].

In conclusion, the selection of a classification method for sentiment analysis in Arabic text relies on the unique characteristics of the dataset and the intended task. It is recommended that various methods, pre-processing techniques, and feature representations be explored to determine the most appropriate approach for a specific Arabic sentiment analysis problem. Furthermore, domain-specific knowledge and linguistic expertise can greatly improve the performance of these methods when applied to sentiment classification in Arabic text. Overall, a comprehensive analysis and evaluation of different classification methods will lead to more accurate and reliable results in Arabic sentiment analysis tasks.

4. Results and Discussion

This section presents the outcomes of sentiment classification experiments that utilized various feature extraction and classification models. Both conventional machine learning models and advanced deep learning models were assessed. The study employed a com-

prehensive set of evaluation metrics to rigorously evaluate the performance of sentiment classification models. These metrics, such as accuracy, precision, recall, and F1-score, were quantified using specific mathematical formulas [31].

Accuracy was determined by the following formula:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ instances} \quad (4)$$

It quantifies the percentage of correctly classified instances from all the samples in a given dataset. It serves as a comprehensive measure of a model's overall effectiveness in accurately predicting and classifying data points. By considering both true positives and true negatives, accuracy provides valuable insights into the reliability and precision of the model's predictions.

Precision was quantified using the following formula:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (5)$$

It evaluates the model's performance based on its ability to effectively classify instances of positive sentiment while minimizing the occurrence of false positives.

Recall (also known as sensitivity) was calculated as:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (6)$$

This metric assesses the model's ability to accurately detect all true positive instances while minimizing the occurrence of false negatives.

The F1-score, which balances precision and recall, was computed with the formula:

$$F1 - Score = \frac{2\ Precision \times Recall}{Precision + Recall} \quad (7)$$

The F1-score offers a comprehensive evaluation of the model's performance, making it particularly valuable when working with imbalanced datasets. These metrics together provide a strong framework for assessing sentiment classification models, enabling informed decisions to be made regarding their effectiveness.

4.1. Data Analysis

The analysis of the Arabic tweet dataset for Hajj sentiment analysis has revealed interesting insights. To summarize the dataset statistics, the following figures are used for illustration. Figure 2 provides a detailed breakdown of tweet distribution over the years and during various stages of the Hajj pilgrimage, categorized by location. The data is divided into three key time periods: "Before Hajj", "During Hajj", and "After Hajj" for each year. It is evident that during the Hajj period, Saudi Arabia consistently had the highest number of tweets across all years, followed by Arab countries and then all other countries. The distribution of tweets across time periods is not uniform, indicating that there may be specific events or occurrences during the Hajj pilgrimage that trigger tweet activity.

It should be noted that the author was only able to retrieve location information for a small number of tweets. The rest of the dataset is provided without location information. Figure 2 provides insights into the overall tweet volumes over the years, categorized by time period and without considering the location. The period labeled as "During Hajj" consistently exhibits the highest tweet activity, indicating that this period serves as a focal point for discussions and expressions related to Hajj. It is also worth noting that the volume of tweets has increased over the years, particularly during the "During Hajj" period, indicating a growing interest and engagement with the topic. In the years 2020 and 2021, there was a significant increase in tweet activity about Hajj compared to other years.

This can be attributed to the strong impact of the COVID-19 pandemic on the Hajj season during those years.

Figure 3 presents a comprehensive analysis of the dataset, where the data are organized based on location and sentiment. This categorization provides a clear understanding of how sentiments are distributed across different regions. Notably, Saudi Arabia stands out with a significant number of tweets covering all sentiments, particularly showing the highest count of positive tweets (1174) and a relatively even distribution between neutral and negative sentiments. Arab countries also show notable presence in the dataset, with 317 positive, 436 neutral, and 120 negative tweets. Additionally, other countries contribute to the dataset as well, with 150 positive, 131 neutral, and 17 negative tweets. These derived insights serve as the basis for conducting a thorough sentiment analysis and delving deeper into the factors that impact tweet engagement during the Hajj pilgrimage.

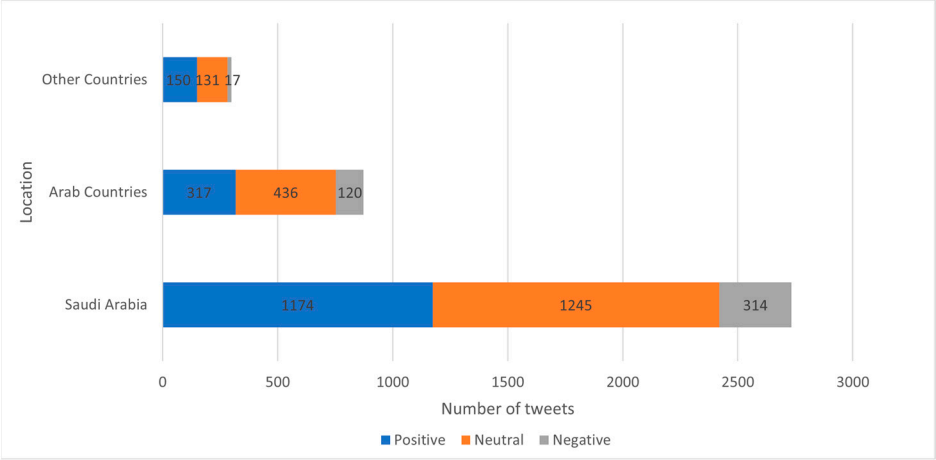


Figure 3. Number of positive, neutral, and negative tweets categorized by location.

In conjunction with the data analysis, there are two word clouds presented in Figures 4 and 5. These word clouds provide distinct viewpoints on the sentiments expressed in the dataset. A word cloud is a visual representation of text data that offers an immediate and intuitive understanding of the most commonly used words within a given set of text. The size of each word in the cloud corresponds to its frequency of occurrence in the text. The two word clouds in Figures 4 and 5 were generated in Python using the '*wordcloud*' library.

Figure 4. Word cloud showing the most frequent words in positive tweets.

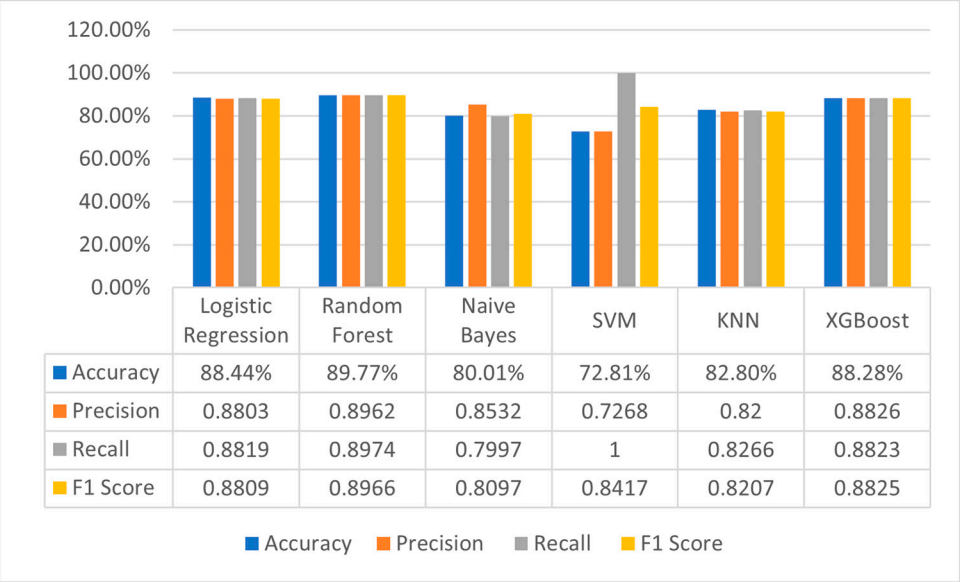
Figure 5. Word cloud showing the most frequent words in negative tweets.

In word clouds, frequently used words are usually displayed prominently, allowing for easy identification at a glance. These visually appealing representations are widely utilized to summarize and visualize textual content effectively. They enable viewers to quickly comprehend the primary themes, keywords, or sentiments within a given dataset or document. For instance, in Figure 4, the word cloud representing positive tweets prominently features words associated with prayers and expressions of gratitude that are readily noticeable. In Figure 5, which represents the common words in negative tweets, it is evident that there are words like "كورونا" (corona), "مريض" (patient), "فيروس" (virus), and "الموت" (death). These words indicate the adverse effects of the COVID-19 epidemic on the Hajj season.

4.2. Machine Learning Model Evaluation

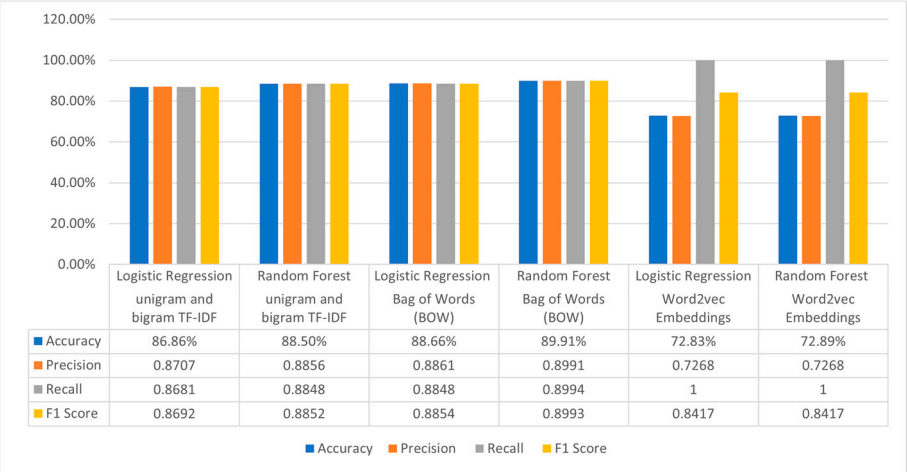
The employed machine learning models have produced valuable outcomes in analysis. These models have undergone rigorous training and testing, demonstrating their efficacy in various aspects, as depicted in Figures 6 and 7. The performance of the Logistic Regression model is dependent on the feature representation employed. For example, when utilizing Unigram TF-IDF features, an accuracy of 88.44% was achieved, along with notable precision, recall, and F1-score values of 0.8803, 0.8819, and 0.8809, respectively. By incorporating Unigram and Bigram TF-IDF features, we achieved an accuracy of 86.86% along with a well-balanced precision, recall, and F1-score. Similarly, when utilizing Bag of Words (BOW) features, the model achieved an accuracy of 88.66% and demonstrated high precision, recall, and F1-score. In contrast, the utilization of Word2vec Embeddings resulted in an accuracy of 72.83% and perfect recall (1.000), although the F1-score was 0.8417.

The Random Forest model consistently demonstrated robust performance across various feature representations. When employing Unigram TF-IDF features, this model achieved the highest accuracy among all models, reaching 89.77%. Additionally, it exhibited precision, recall, and F1-score values of 0.8962, 0.8974, and 0.8966, respectively. Even when combining Unigram and Bigram TF-IDF features, this model maintained its strength with an accuracy of 88.50% and balanced precision, recall, and F1-score. Furthermore, when using BOW features, it excelled with an accuracy of 89.91% and displayed high precision, recall, and F1-score. On the other hand, the use of Word2vec embeddings resulted in an accuracy of 72.89%, along with perfect recall (1.000) and a balanced F1-score.



	Logistic Regression	Random Forest	Naive Bayes	SVM	KNN	XGBoost
Accuracy	88.44%	89.77%	80.01%	72.81%	82.80%	88.28%
Precision	0.8803	0.8962	0.8532	0.7268	0.82	0.8826
Recall	0.8819	0.8974	0.7997	1	0.8266	0.8823
F1 Score	0.8809	0.8966	0.8097	0.8417	0.8207	0.8825

Figure 6. Results of various ML models with Unigram TF-IDF features.



	Logistic Regression unigram and bigram TF-IDF	Random Forest unigram and bigram TF-IDF	Logistic Regression Bag of Words (BOW)	Random Forest Bag of Words (BOW)	Logistic Regression Word2vec Embeddings	Random Forest Word2vec Embeddings
Accuracy	86.86%	88.50%	88.66%	89.91%	72.83%	72.89%
Precision	0.8707	0.8856	0.8861	0.8991	0.7268	0.7268
Recall	0.8681	0.8848	0.8848	0.8994	1	1
F1 Score	0.8692	0.8852	0.8854	0.8993	0.8417	0.8417

Figure 7. Results of best performing ML models (logistic regression and random forest) with various feature extractors.

The Naive Bayes model, employing Unigram TF-IDF features, attained an accuracy of 80.01%. While it exhibited noteworthy precision, it had slightly lower recall and F1-scores of 0.8532 and 0.7997, respectively.

The Support Vector Machines (SVM) model, when utilizing Unigram TF-IDF features, achieved an accuracy level of 72.81%. Similarly, the K-Nearest Neighbors (KNN) model, leveraging the same Unigram TF-IDF features, attained a higher accuracy rate of 82.80% and displayed balanced precision, recall, and F1-score values of 0.8207. Finally, when combined with Unigram TF-IDF features, the XGBoost model achieved an impressive accuracy of 88.28% and showcased well-balanced precision, recall, and F1-score metrics.

As we can see from the sensitivity curves of machine learning models in Figure 8, there is a notable variation in sensitivity, specificity, accuracy, precision, and F1-scores across different probability cutoffs. The fluctuation in these metrics provides insights into the model's performance at various decision thresholds. Notably, sensitivity increases when we increase the probability cutoff in most of the models. However, in SVM, the sensitivity is either 0 or 1, and that is because the model could not learn the data features well. Now, turning our attention to the sensitivity curves for the naive bayes model, it is evident that the sensitivity values were almost constant across most of the cutoff values.

Naive bayes demonstrates consistency across all evaluation metrics that are not affected by changing decision threshold, contributing to the overall understanding of its behavior in different scenarios.

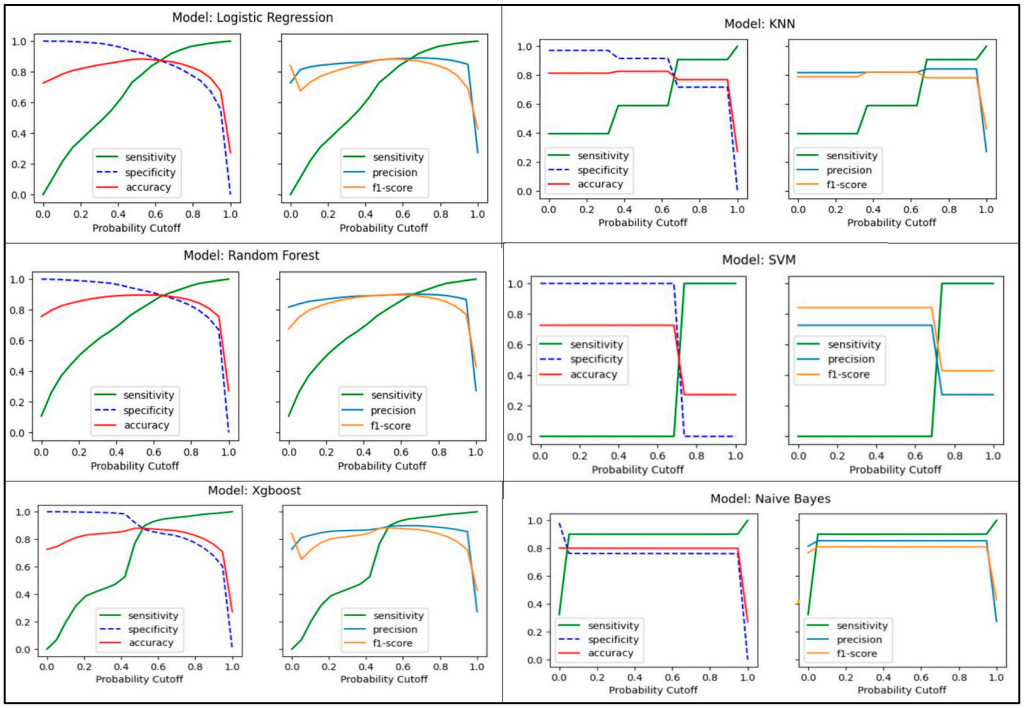


Figure 8. Sensitivity curves for ML models.

Considering the sensitivity-specificity trade-off, as we can see, in all models, the sensitivity increases when specificity decreases, and vice versa. This trade-off is crucial in determining which cutoff point to consider.

Analyzing the impact of probability cutoffs, we find that a traditional selection of 0.5 is a good choice. Notably, small changes in cutoff values can highly affect model performance, suggesting the need for careful consideration in selecting the appropriate threshold. In terms of model interpretability, logistic regression, random forest, and xGboost algorithms were the best ML models for interpreting the data. While naïve bayes, SVM and KNN were more challenging and were not considered the best choice for interpreting the data. In conclusion, the sensitivity curves of ML models offer a comprehensive understanding of their performance characteristics. By delving into the nuances of these curves, we can make informed decisions about model selection, recognizing the strengths and limitations of each in the context of our study.

In the evaluation of our models, the Receiver Operating Characteristic (ROC) curves in Figure 9 provide valuable insights into their discriminatory power, with corresponding Area Under the Curve (AUC) values quantifying overall performance. Notably, the logistic regression model demonstrated a commendable AUC of 0.84, while the random forest, naive bayes, and xGboost models exhibited even higher AUC values of 0.86, 0.83, and 0.85, respectively. In contrast, the K-Nearest Neighbors (KNN) model and Support Vector Machine (SVM) display AUC values of 0.75 and 0.5, respectively. The superior AUC values for logistic regression, random forest, naive bayes, and xGboost models may be attributed to their ability to capture complex relationships within the data and maintain a balanced trade-off between sensitivity and specificity. Meanwhile, the comparatively lower AUC for KNN may stem from its sensitivity to the local structure of the data, and the SVM's AUC of 0.5 suggests a classification performance equivalent to random chance. Our findings underscore the significance of model selection in achieving optimal AUC

and highlight the importance of algorithms capable of effectively distinguishing between classes in our dataset.

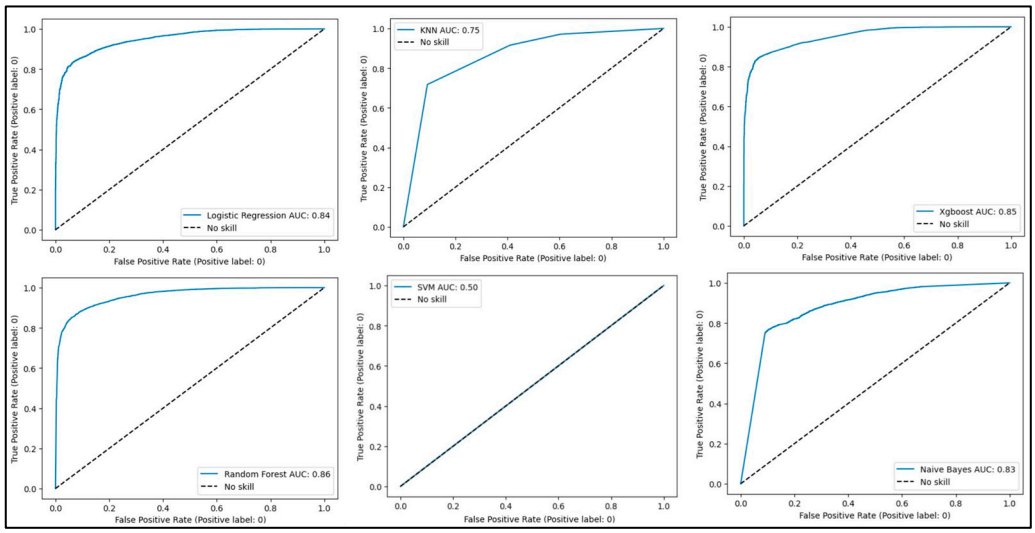
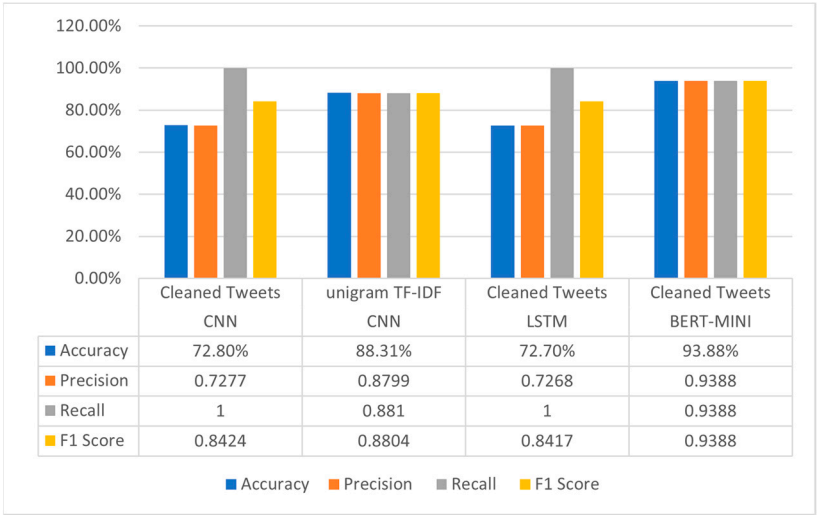


Figure 9. ROC curves for ML models.

4.3. Deep Learning Model Evaluation

Remarkable consistency in performance was observed among the deep learning models across various feature extractions. The comprehensive analysis, as depicted in Figure 10, clearly illustrates that both the Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) models achieved an impressive accuracy of approximately 72% when trained on Cleaned Tweets. Notably, the models exhibited excellent recall, with a score of 1.000. An intriguing observation is that the CNN model outperformed its counterparts by achieving a higher accuracy of 88.31% when trained on Unigram TF-IDF features. This revelation highlights the significant impact that feature extraction can have on model performance and suggests the potential efficacy of utilizing TF-IDF features for improved accuracy and precision in natural language processing tasks.



	Cleaned Tweets CNN	unigram TF-IDF CNN	Cleaned Tweets LSTM	Cleaned Tweets BERT-MINI
Accuracy	72.80%	88.31%	72.70%	93.88%
Precision	0.7277	0.8799	0.7268	0.9388
Recall	1	0.881	1	0.9388
F1 Score	0.8424	0.8804	0.8417	0.9388

Figure 10. Results of deep learning models.

When applied to cleaned tweets, the BERT-MINI model achieved an impressive accuracy rate of 93.88%, surpassing all other models in the domain. Notably, the BERT-MINI model also demonstrated perfect recall, further emphasizing its effectiveness in accurately

identifying sentiments expressed within text data. Furthermore, with an impressive F1-score of 0.9388, this state-of-the-art model has solidified its position as a formidable tool for analyzing and comprehending sentiment patterns. There is no denying the incredible potential of sentiment analysis, which has created significant opportunities for its implementation across diverse industries and sectors.

In evaluating the performance of our models—CNN, LSTM, and Mini BERT—through the sensitivity curves in Figure 11, distinct patterns emerge that shed light on their discriminatory capabilities. Notably, Mini BERT outperforms both CNN and LSTM, showing superior sensitivity across various thresholds. The sensitivity curve for Mini BERT consistently demonstrates its robust ability to capture relevant features and nuances in the data, leading to better discrimination between classes. Conversely, the sensitivity curves for CNN reveal commendable performance, although not reaching the heights achieved by Mini BERT. On the other hand, the LSTM model exhibits notably lower sensitivity, indicative of its struggle to effectively capture long-term dependencies in the data. The superior performance of Mini BERT can be attributed to its attention mechanism, allowing it to effectively leverage contextual information and semantic relationships. In contrast, CNN's ability to capture spatial hierarchies contributes to its performance. These insights underscore the critical role of model architecture in achieving optimal sensitivity and emphasize the advantages of leveraging advanced architectures like Mini BERT for tasks demanding a nuanced understanding of context and semantics.

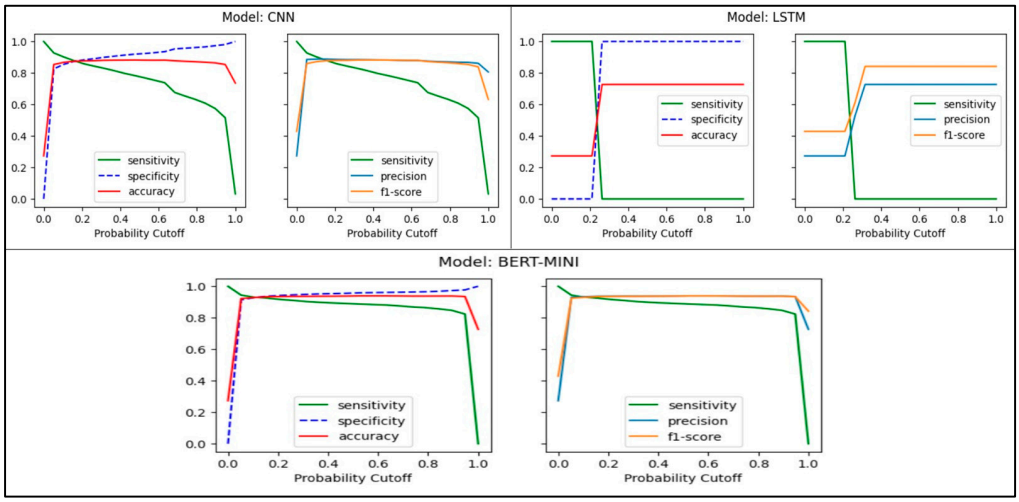


Figure 11. Sensitivity curves of DL models.

In our DL model evaluation, the ROC curves presented in Figure 12 provide insights towards understanding the performance of the models. Notably, the CNN model exhibits an impressive Area Under the Curve (AUC) of 0.94, indicating a high degree of accuracy in distinguishing between classes. In contrast, the LSTM model presents an AUC of 0.5, suggestive of a classification performance equivalent to random chance. The standout performer, however, is the Mini BERT model, showcasing an exceptional AUC of 0.981. This remarkable AUC value attests to the superior ability of Mini BERT to discern between positive and negative instances. These AUC values not only quantify the models' discriminatory power but also underscore the significance of advanced architectures, like Mini BERT, in achieving heightened performance in complex classification tasks.

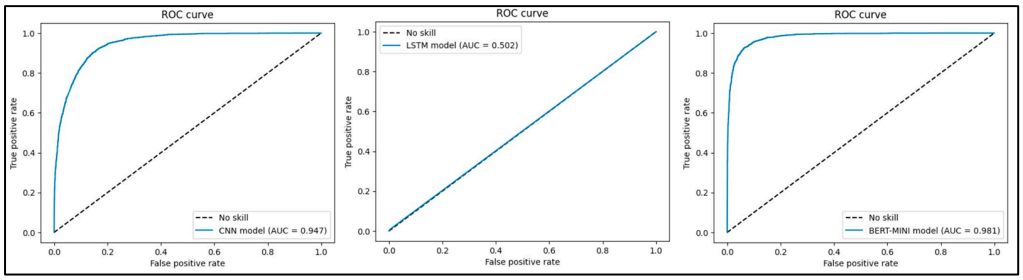


Figure 12. ROC curves for DL models.

In conclusion, the choice of a classification model plays a crucial role in sentiment classification tasks. Random Forest, particularly when combined with Unigram TF-IDF features, consistently demonstrates outstanding performance across various evaluation metrics. On the other hand, within the realm of deep learning, BERT-MINI has emerged as the top-performing model in terms of performance. This highlights the importance of leveraging advanced language models in sentiment analysis.

In order to thoroughly examine the sentiment analysis findings, Figure 13 offers a comprehensive and detailed overview of the quantity of tweets categorized as positive and negative across multiple years. The collected data is thoughtfully segmented into three distinct time periods, namely "Before Hajj", "During Hajj", and "After Hajj". This meticulous categorization allows for a more nuanced understanding of the sentiment trends exhibited during different phases of the Hajj pilgrimage seasons. Particularly noteworthy is the substantial surge in both positive and negative tweets observed during the crucial "During Hajj" period. This notable increase in tweet activity implies a heightened level of engagement and expressive sentiment during this critical phase.

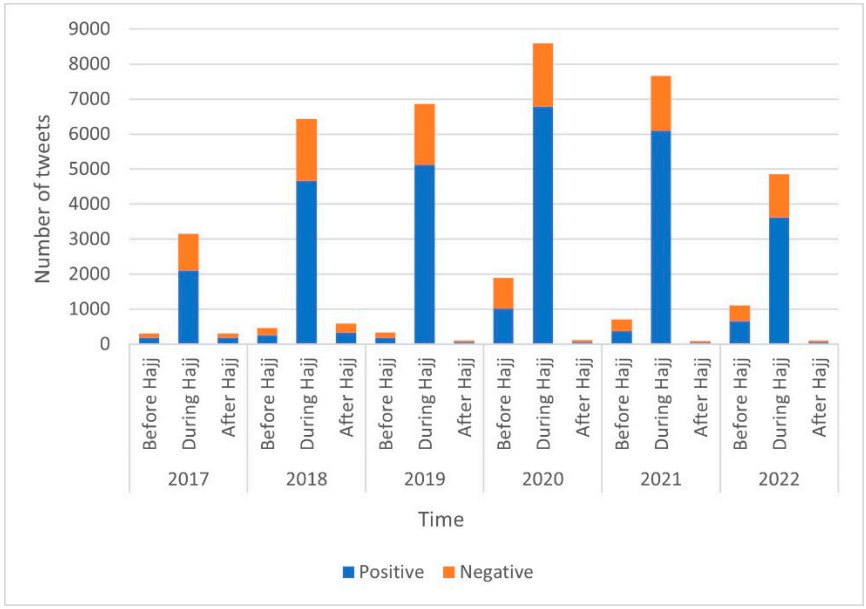


Figure 13. Sentiment Analysis along the years by lexicon-based approach.

Moreover, an analysis of the data revealed intriguing variations in sentiment distribution across different years, indicating that external factors or events may have a significant impact on the expression of sentiment within the context of Hajj. This dataset presents a valuable resource for understanding the dynamics of sentiment and has the potential to unveil the contributing factors behind both positive and negative sentiments during this significant religious pilgrimage. By delving into these patterns, researchers can gain deeper insights into the intricate interplay between external influences and emotional responses within this religious framework.

In interpreting the results of our sentiment analysis on Arabic tweets discussing the Hajj pilgrimage, a practical application emerged in enhancing services for pilgrims. By scrutinizing sentiment patterns over the six-year period before, during, and after each Hajj event, we gained valuable insights into the prevailing sentiments associated with the pilgrimage. Positive sentiments may indicate moments of heightened satisfaction or communal joy, while negative sentiments could suggest potential areas of concern or dissatisfaction. Leveraging this information, service providers and organizers can tailor their offerings and support systems to align with the emotional needs of pilgrims at different stages of the Hajj journey. For instance, identifying positive sentiments during specific periods may guide the implementation of enhancements to infrastructure, contributing to an overall positive pilgrimage experience. Similarly, promptly addressing negative sentiments can allow for targeted improvements in areas that may impact pilgrim satisfaction, ultimately contributing to the optimization of services during the Hajj pilgrimage.

4.4. Comparison with Similar Studies

Our sentiment analysis results on Arabic tweets discussing the Hajj pilgrimage showcase the efficacy of our approach, particularly with the Arabic MINI BERT model, which achieved an impressive accuracy of 93.8%. Notably, our results outperform those reported in [42], where the highest accuracy obtained was 92.13% using LSTM. Additionally, when compared to [39] and its focus on Arabic sentiment analysis, our findings surpass the reported accuracies for the LSTM and Bi-LSTM models. Specifically, our random forest model achieved 89.7%, outperforming the corresponding method in [42] (79.19%). These outcomes underscore the robustness of our sentiment analysis methodology, showing superior accuracy rates that can be crucial for understanding the nuanced sentiments expressed in Arabic tweets related to the Hajj pilgrimage. The utilization of the Arabic MINI BERT model, in particular, demonstrates its effectiveness in capturing the complexities of Arabic text, providing a valuable contribution to sentiment analysis within the context of religious and cultural events such as the Hajj.

5. Conclusions

This research ventured into sentiment analysis on Twitter texts discussing Hajj over a six-year span, aiming to unravel the intricacies of public sentiment surrounding this significant event. Leveraging an Arabic tweet dataset obtained from Twitter, carefully curated with Hajj-centric keywords, our study underwent rigorous pre-processing and feature extraction to prepare the text for analysis. Employing a spectrum of machine learning and deep learning techniques, we observed nuanced performance variations among the models.

The choice of a classification model plays a crucial role in sentiment classification tasks. Random Forest, particularly when combined with Unigram TF-IDF features, consistently demonstrates outstanding performance across various evaluation metrics. On the other hand, within the realm of deep learning, BERT-MINI has emerged as the top-performing model in terms of performance. This highlights the importance of considering both traditional machine learning approaches and deep learning techniques when selecting a classification model for sentiment analysis tasks. Ultimately, the decision should be based on the specific requirements of the project and the available resources. By carefully evaluating and comparing different models, researchers and practitioners can make informed choices that maximize accuracy and efficiency in sentiment classification.

Beyond model performance, our sentiment analysis journey across different phases of the Hajj pilgrimage revealed dynamic trends. The "During Hajj" period exhibited a significant surge in both positive and negative tweets, indicating heightened emotional expression and engagement during this critical time. Negative tweets often featured terms such as "كورونا" (corona), "مريض" (patient), "فيروس" (virus), "عذاب" (punishment), "محرم" (forbidden), and "الموت" (death). Some of these tweets implied the negative impact of the COVID-19 pandemic on the Hajj season. This analysis unveils evolving sentiment

distributions across Hajj years, implying the influence of external factors on the expression of sentiments within this unique context.

These findings have profound implications for understanding the nuanced sentiments surrounding Hajj, shedding light on the multifaceted nature of public discourse during significant events. The observed surge in emotional expression during the Hajj period suggests the importance of context-aware sentiment analysis, considering the emotional intensity linked to specific phases of events.

The practical applications of our research extend beyond academia. Event management teams and public relations practitioners can leverage sentiment-aware strategies for effective communication and engagement during major events like the Hajj. Additionally, social sentiment monitoring tools informed by our findings could provide valuable insights for authorities, helping them tailor responses and interventions based on the evolving sentiments of the public.

As we chart future directions, exploring adaptive sentiment models that account for evolving language nuances becomes pivotal. Real-time sentiment monitoring tools could be developed to offer timely insights during major events, providing stakeholders with actionable information. Further research could delve into refining sentiment analysis frameworks for diverse cultural and linguistic contexts, ensuring the generalizability of findings across various regions and events. A potential avenue for future work involves delving deeper into the nuanced distinction between true and fake positive and negative tweets and dealing with implicit and explicit aspect extraction. The next phase of the research will focus on refining the sentiment analysis methodology to accurately identify instances where expressions may convey insincerity or sarcasm.

Funding: This research received no external funding.

Institutional Review Board Statement: This article does not contain any studies with human participants performed by any of the authors.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within this article.

Acknowledgments: The authors would like to thank the Deanship of Scientific Research at Umm Al-Qura University.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Shafi, S.; Dar, O.; Khan, M.; Khan, M.; Azhar, E.I.; McCloskey, B.; Zumla, A.; Petersen, E. The annual Hajj pilgrimage—Minimizing the risk of ill health in pilgrims from Europe and opportunity for driving the best prevention and health promotion guidelines. *Int. J. Infect. Dis.* **2016**, *47*, 79–82. [CrossRef] [PubMed]
2. Jokhdar, H.; Khan, A.; Asiri, S.; Motair, W.; Assiri, A.; Alabdulaali, M. COVID-19 Mitigation Plans during Hajj 2020: A Success Story of Zero Cases. *Health Secur.* **2021**, *19*, 133–139. [CrossRef] [PubMed]
3. Alqahtany, A. Affordable housing in Saudi Arabia's vision 2030: New developments and new challenges. *Int. J. Hous. Mark. Anal.* **2021**, *14*, 243–256. [CrossRef]
4. Aly, S.; Gutub, A. Intelligent Recognition System for Identifying Items and Web-Portal System for Missing-and-Found Items. *NED Univ. J. Res.* **2018**, *966*, 17–24.
5. Al Mojamed, M. Smart mina: Lorawan technology for smart fire detection application for hajj pilgrimage. *Comput. Syst. Sci. Eng.* **2022**, *40*, 259–272. [CrossRef]
6. Rahman, A.; Hamid, N.A.W.A.; Rahiman, A.R.; Zafar, B. Towards accelerated agent-based crowd simulation for Hajj and Umrah. In Proceedings of the 2015 International Symposium on Agents, Multi-Agent Systems and Robotics (ISAMSR), Putrajaya, Malaysia, 18–19 August 2015; pp. 65–70. [CrossRef]
7. Bhuiyan, M.R.; Abdullah, J.; Hashim, N.; Al Farid, F.; Samsudin, M.A.; Abdullah, N.; Uddin, J. Hajj pilgrimage video analytics using CNN. *Bull. Electr. Eng. Inform.* **2021**, *10*, 2598–2606. [CrossRef]
8. Bati, G. Using Big Data Tools to Analyze Tweets Related to Hajj Sentimentally. In Proceedings of the Hajj Forum 2016—The 15 Scientific Hajj Research Forum, Madinah, Saudi Arabia. 2016, pp. 177–184. Available online: https://www.researchgate.net/publication/292146913_Using_Big_Data_Tools_to_Analyze_Tweets_Related_to_Hajj_Sentimentally (accessed on 12 November 2023).

9. Ottom, M.A.; Nahar, K.M.O. Social Media Sentiment Analysis: The Hajj Tweets Case Study. *J. Comput. Sci.* **2021**, *17*, 265–274. [CrossRef]
10. Shambour, M.K. Analyzing perceptions of a global event using CNN-LSTM deep learning approach: The case of Hajj 1442 (2021). *PeerJ Comput. Sci.* **2022**, *8*, e1087. [CrossRef]
11. Showail, A.J. Solving Hajj and Umrah Challenges Using Information and Communication Technology: A Survey. *IEEE Access* **2022**, *10*, 75404–75427. [CrossRef]
12. Aldhubaib, H.A. Impressions of the Community of Makkah on the Hajj in the Light of COVID-19 Pandemic: Quantitative and AI-based Sentiment Analyses. *J. King Abdulaziz Univ. Eng. Sci.* **2022**, *32*, 41–57. Available online: https://journals.kau.edu.sa/index.php/JENGSCI (accessed on 1 May 2023). [CrossRef]
13. Alanazi, N.; Khan, E.; Gutub, A. Involving Spaces of Unicode Standard within Irreversible Arabic Text Steganography for Practical Implementations. *Arab. J. Sci. Eng.* **2021**, *46*, 8869–8885. [CrossRef]
14. Almehmadi, E.; Gutub, A. Novel Arabic e-Text Watermarking Supporting Partial Dishonesty Based on Counting-Based Secret Sharing. *Arab. J. Sci. Eng.* **2022**, *47*, 2585–2609. [CrossRef]
15. Alqurashi, T. Arabic Sentiment Analysis for Twitter Data: A Systematic Literature Review. *Eng. Technol. Appl. Sci. Res.* **2023**, *13*, 10292–10300. [CrossRef]
16. Oueslati, O.; Cambria, E.; HajHmida, M.B.; Ounelli, H. A review of sentiment analysis research in Arabic language. *Futur. Gener. Comput. Syst.* **2020**, *112*, 408–430. [CrossRef]
17. Fadel, I.; ÖZ, C. A Sentiment Analysis Model for Terrorist Attacks Reviews on Twitter. *Sak. Univ. J. Sci.* **2020**, *24*, 1294–1302. [CrossRef]
18. Onan, A. Sentiment analysis on massive open online course evaluations: A text mining and deep learning approach. *Comput. Appl. Eng. Educ.* **2021**, *29*, 572–589. [CrossRef]
19. Asif, M.; Ishtiaq, A.; Ahmad, H.; Aljuaid, H.; Shah, J. Sentiment analysis of extremism in social media from textual information. *Telemat. Inform.* **2020**, *48*, 101345. [CrossRef]
20. Toçoğlu, M.A.; Onan, A. Sentiment Analysis on Students' Evaluation of Higher Educational Institutions. *Adv. Intell. Syst. Comput.* **2021**, *1197*, 1693–1700. [CrossRef]
21. Aljabri, M.; Chrouf, S.M.; Alzahrani, N.A.; Alghamdi, L.; Alfehaid, R.; Alqarawi, R.; Alhuthayfi, J.; Alduhailan, N. Sentiment analysis of arabic tweets regarding distance learning in saudi arabia during the COVID-19 pandemic. *Sensors* **2021**, *21*, 5431. [CrossRef]
22. Albahli, S. Twitter sentiment analysis: An Arabic text mining approach based on COVID-19. *Front. Public Health* **2022**, *10*, 966779. [CrossRef]
23. Rodríguez-Ibánez, M.; Casánez-Ventura, A.; Castejón-Mateos, F.; Cuenca-Jiménez, P.M. A review on sentiment analysis from social media platforms. *Expert Syst. Appl.* **2023**, *223*, 119862. [CrossRef]
24. Sunitha, D.; Patra, R.K.; Babu, N.V.; Suresh, A.; Gupta, S.C. Twitter sentiment analysis using ensemble based deep learning model towards COVID-19 in India and European countries. *Pattern Recognit. Lett.* **2022**, *158*, 164–170. [CrossRef]
25. Lomborg, S.; Bechmann, A. Using APIs for Data Collection on Social Media. *Inf. Soc.* **2014**, *30*, 256–265. [CrossRef]
26. Ben-Abdallah, E.; Boukadi, K. The effect of Facebook behaviors on the prediction of review helpfulness. *J. Data Min. Digit. Humanit.* **2022**, *2022*. [CrossRef]
27. Wilson, R.E.; Gosling, S.D.; Graham, L.T. A Review of Facebook Research in the Social Sciences. *Perspect. Psychol. Sci.* **2012**, *7*, 203–220. [CrossRef] [PubMed]
28. Breuer, J.; Kmetty, Z.; Haim, M.; Stier, S. User-centric approaches for collecting Facebook data in the 'post-API age': Experiences from two studies and recommendations for future research. *Inf. Commun. Soc.* **2022**, *26*, 2649–2668. [CrossRef]
29. McCrow-Young, A. Approaching Instagram data: Reflections on accessing, archiving and anonymising visual social media. *Commun. Res. Pract.* **2021**, *7*, 21–34. [CrossRef]
30. Bainotti, L.; Caliandro, A.; Gandini, A. From archive cultures to ephemeral content, and back: Studying Instagram Stories with digital methods. *New Media Soc.* **2021**, *23*, 3656–3676. [CrossRef]
31. Almaliki, M.; Almars, A.M.; Gad, I.; Atlam, E.S. ABMM: Arabic BERT-Mini Model for Hate-Speech Detection on Social Media. *Electronics* **2023**, *12*, 1048. [CrossRef]
32. Kready, J.; Shimray, S.A.; Hussain, M.N.; Agarwal, N. YouTube Data Collection Using Parallel Processing. In Proceedings of the 2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), New Orleans, LA, USA, 18–22 May 2020; pp. 1119–1122. [CrossRef]
33. Mai, L.; Le, B. Joint sentence and aspect-level sentiment analysis of product comments. *Ann. Oper. Res.* **2021**, *300*, 493–513. [CrossRef]
34. Albishre, K.M.; Albasam, A.S. Social Media Monitoring for Enhancing Hajj Pilgrimage Experience. In Proceedings of the 21th Scientific Forum of Hajj, Umrah and Madinah Visit Research—Scientific Bulletin, Makkah, Saudi Arabia, 9–10 March 2022; pp. 60–70.
35. Elgamal, M. Sentiment Analysis Methodology of Twitter Data with an application on Hajj season. *Int. J. Eng. Res. Sci.* **2016**, *2*, 82–87.

36. Zahrani, R.; Khaldi, I.; Qahtani, K. The Impact of Understanding Social Media Content on Improving Performance during the Hajj Season, a Twitter Case Study for the Hajj Season 1436 AH. In Proceedings of the 17th Scientific Forum for the Research of Hajj, Umrah and Madinah Visit. 2017, pp. 742–784. Available online: https://drive.uqu.edu.sa/_/hajj/files/multaqa/143817.pdf (accessed on 12 November 2023).
37. Gutub, A.; Shambour, M.K.; Abu-Hashem, M.A. Coronavirus Impact on Human Feelings During 2021 Hajj Season via Deep Learning Critical Twitter Analysis. *J. Eng. Res.* **2023**, *11*, 100001. [CrossRef]
38. Alamoudi, H.; Aljojo, N.; Munshi, A.; Alghoson, A.; Banjar, A.; Tashkandi, A.; Al-Tirawi, A.; Alsaleh, I. Arabic Sentiment Analysis for Student Evaluation using Machine Learning and the AraBERT Transformer. *Eng. Technol. Appl. Sci. Res.* **2023**, *13*, 11945–11952. [CrossRef]
39. Habbat, N.; Nouri, H.; Anoun, H.; Hassouni, L. Using AraGPT and ensemble deep learning model for sentiment analysis on Arabic imbalanced dataset. *ITM Web Conf.* **2023**, *52*, 02008. [CrossRef]
40. Kaseb, A.; Farouk, M. Active learning for Arabic sentiment analysis. *Alexandria Eng. J.* **2023**, *77*, 177–187. [CrossRef]
41. El Karfi, I.; El Fkihi, S. A combined Bi-LSTM-GPT Model for Arabic Sentiment Analysis. *Int. J. Intell. Syst. Appl. Eng.* **2023**, *11*, 77–84.
42. Binmahfoudh, A. Improved Deep Learning Sentiment Analysis for Arabic. *J. Theor. Appl. Inf. Technol.* **2023**, *101*. Available online: https://www.kaggle.com/c/ (accessed on 11 December 2023).
43. Louati, A.; Louati, H.; Kariri, E.; Alaskar, F.; Alotaibi, A. Sentiment Analysis of Arabic Course Reviews of a Saudi University Using Support Vector Machine. *Appl. Sci.* **2023**, *13*, 12539. [CrossRef]
44. 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 Cogn. Comput.* **2023**, *7*, 127. [CrossRef]
45. Shambour, M.K.; Gutub, A. Progress of IoT Research Technologies and Applications Serving Hajj and Umrah. *Arab. J. Sci. Eng.* **2021**, *47*, 1253–1273. [CrossRef]
46. Kiritchenko, S.; Zhu, X.; Mohammad, S.M. Sentiment Analysis of Short Informal Texts. *J. Artif. Intell. Res.* **2014**, *50*, 723–762. [CrossRef]
47. Mohammad, S.M.; Kiritchenko, S.; Zhu, X. Nrc-canada: Building the state- of-the-art in sentiment analysis of tweets. *arXiv* **2013**, arXiv:1308.6242.
48. Alayba, A.; Palade, V.; England, M.; Iqbal, R. Arabic language sentiment analysis on health services. In Proceedings of the 1st International Workshop on Arabic Script Analysis and Recognition (ASAR), Nancy, France, 3–5 April 2017; pp. 114–118.
49. Alghamdi, H.M.; Selamat, A. Topic detections in Arabic Dark websites using improved Vector Space Model. In Proceedings of the 2012 4th Conference on Data Mining and Optimization (DMO), Langkawi, Malaysia, 2–4 September 2012; pp. 6–12. [CrossRef]
50. Ayadi, R.; Maraoui, M.; Zrigui, M. Latent Topic Model for Indexing Arabic Documents. *Int. J. Inf. Retr. Res.* **2014**, *4*, 29–45. [CrossRef]
51. Shoukry, A. Arabic Sentence-Level Sentiment Analysis. The American University in Fountain. Master's Thesis, American University in Cairo, Cairo, Egypt, 2013.