SENTIMENT ANALYSIS OF HAJJ-RELATED CONTENT ON X

MOHAMED TAREK ELSAYED MOHAMED TORKY

UNIVERSITI TEKNOLOGI MALAYSIA

UNIVERSITI TEKNOLOGI MALAYSIA
DECLARATION OF thesis

Author's full name	:	Mohamed Tarek Elsayed Mohamed Yousef Mohamed Torky		
Student's Matric No.	:	MCS241037	Academic Session	:202420252
Date of Birth	:	03/01/2003	UTM Email	: mohamed.elsayed@graduate.utm.my
Thesis Title	:	SENTIMENT ANALYSIS OF HAJJ-RELATED CONTENT ON X		

I declare that this thesis is classified as:

☒ **OPEN ACCESS** I agree that my report to be published as a hard copy or made available through online open access.

□ **RESTRICTED** Contains restricted information as specified by the organization/institution where research was done. *(The library will block access for up to three (3) years)*

□ **CONFIDENTIAL** Contains confidential infomation as specified in the Official Secret Act 1972)

(If none of the options are selected, the first option will be chosen by default)

I acknowledged the intellectual property in the thesis belongs to Universiti Teknologi Malaysia, and I agree to allow this to be placed in the library under the following terms :

1. This is the property of Universiti Teknologi Malaysia
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of only.
3. The Library of Universiti Teknologi Malaysia is allowed to make copies of this thesis for academic exchange.

Signature of Student:

Signature : Mohamed Tarek Torky

Full Name Mohamed Tarek Elsayed Mohamed Yousef Mohamed Torky
Date : 30/06/2025

Approved by Supervisor(s)

Signature of Supervisor I:	Signature of Supervisor II
Full Name of Supervisor I	Full Name of Supervisor II
Date :	Date :

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

SENTIMENT ANALYSIS OF HAJJ-RELATED CONTENT ON X

MOHAMED TAREK ELSAYED MOHAMED TORKY

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Master of Data Science

School of Computing
Faculty of Computing
Universiti Teknologi Malaysia

APRIL 2025

DECLARATION

I declare that this thesis entitled *“Sentiment Analysis of Hajj-related Content on X”* is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : Mohamed Tarek Torky
Name : Mohamed Tarek Elsayed Mohamed Torky
Date : 30 JUNE 2025

ACKNOWLEDGEMENT

I spoke with several people while writing my thesis, including researchers, academics, and practitioners. They have aided my comprehension and thought processes. I'd want to convey my heartfelt gratitude to my major thesis supervisors, –----, for their support, direction, criticism, and friendship. This thesis would not have been the same without their continued support and interest.

ABSTRACT

In the digital era, social media platforms like X (formerly Twitter) serve as active spaces for users to express opinions and share experiences about global events, including religious practices such as the Hajj pilgrimage. The rapid and voluminous generation of such user-generated content makes manual analysis impractical, necessitating automated solutions. This study aims to analyze public sentiment toward Hajj by applying sentiment analysis techniques to English-language tweets. Using simple machine learning models and Natural Language Processing (NLP) tools such as VADER and TextBlob, the project classifies Hajj-related tweets into three sentiment categories: positive, negative, and neutral. The research involves data collection, preprocessing, sentiment classification, and result visualization using charts and word clouds. The findings provide insight into how the global community perceives Hajj, highlighting dominant emotions, frequently used words, and sentiment trends. This research contributes to understanding online religious discourse and offers a lightweight, domain-specific tool that Islamic organizations, researchers, and policymakers can use to monitor public perception of Hajj on social media.

ABSTRAK

Dalam era digital, platform media sosial seperti X (dahulunya Twitter) berfungsi sebagai ruang aktif untuk pengguna meluahkan pendapat dan berkongsi pengalaman tentang acara global, termasuk amalan keagamaan seperti ziarah haji. Penjanaan kandungan yang dijana pengguna yang pesat dan banyak menjadikan analisis manual tidak praktikal, memerlukan penyelesaian automatik. Kajian ini bertujuan untuk menganalisis sentimen orang ramai terhadap haji dengan menggunakan teknik analisis sentimen kepada tweet berbahasa Inggeris. Menggunakan model pembelajaran mesin mudah dan alatan Pemprosesan Bahasa Semulajadi (NLP) seperti VADER dan TextBlob, projek ini mengklasifikasikan tweet berkaitan Haji kepada tiga kategori sentimen: positif, negatif dan neutral. Penyelidikan melibatkan pengumpulan data, prapemprosesan, klasifikasi sentimen, dan visualisasi hasil menggunakan carta dan awan perkataan. Penemuan ini memberikan gambaran tentang cara masyarakat global melihat haji, menonjolkan emosi yang dominan, perkataan yang sering digunakan dan trend sentimen. Penyelidikan ini menyumbang kepada pemahaman wacana agama dalam talian dan menawarkan alat khusus domain yang ringan yang boleh digunakan oleh organisasi Islam, penyelidik dan penggubal dasar untuk memantau persepsi orang ramai tentang Haji di media sosial.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

The rise of social media has changed the way people share experiences, express their opinions, and discuss world events. Hajj, the yearly Islamic pilgrimage to Mecca, is one of the most major religious events to get worldwide attention online. Every year, millions of Muslims perform the Hajj, and many of them express their thoughts and experiences on X (previously Twitter), giving it a valuable source of real-time emotional and sociological information.

Understanding what people think around Hajj-related experiences on X gives useful insights for religious groups, scholars, and the government. Sentiment analysis, a Natural Language Processing (NLP) approach, may assist categorize and comprehend this massive amount of unstructured text data. This chapter describes the study's history, research challenge, objectives, and scope, with the goal of analyzing sentiment in Hajj-related tweets using simple machine learning algorithms.

1.2 Problem Background

People now interact globally via X (previously Twitter) by sharing their opinions and debating global concerns, including significant religious pilgrimage like Hajj. People share their feelings, experiences, and opinions on Hajj, ranging from spiritual gratitude to practical problems, creating a public narrative around this holy pilgrimage. These expressions are frequently unstructured and uploaded in enormous quantities, rendering human analysis impossible.

While sentiment analysis has been widely applied in commercial fields such as product and service reviews, limited research has focused on analyzing sentiments related to Islamic practices, particularly Hajj. Given the international attention that Hajj receives, and the complex sentiments associated with it, the accurate and automated understanding of these public perceptions is crucial. This study proposes an NLP-based sentiment analysis system specifically focused on Hajj-related tweets to fill this gap.

1.3 Problem Statement

Despite the enormous number of Hajj-related tweets sent annually on X, there is an obvious lack of specific tools for assessing public opinion toward Hajj. Due to the number and speed with which these tweets are published, manual analysis is wasteful and unsustainable. Most existing sentiment analysis methods are general-purpose or designed for commercial domains, therefore they do not capture the religious and cultural subtleties connected with Hajj.

This limitation has an impact on a variety of stakeholders, including Islamic groups, researchers, and the government, all of whom require accurate methods to analyze public opinion, detect recurrent concerns, and identify disinformation. There is a need for a straightforward, purpose-built sentiment analysis tool that can automatically categorize and show sentiments in Hajj-related tweets.

1.4 Research Questions

This study aims to explore how the Hajj pilgrimage is perceived on X (formerly Twitter) by applying sentiment analysis to tweets. The research will be guided by the following questions:

(a) What is the overall sentiment of Hajj-related tweets on X — positive, negative, or neutral?

(b) Can a simple sentiment analysis model accurately classify Hajj-related tweets into sentiment categories?

(c) What are the most frequently used words and phrases in each sentiment category?

(d) What conclusions and insights can be drawn from users' sentiments about Hajj based on their tweets?

1.5 Research Objectives

(a) To collect and preprocess Hajj-related tweet data from X (formerly Twitter).

(b) To develop a sentiment analysis model using simple machine learning and NLP techniques to classify tweets into positive, negative, or neutral sentiments.

(c) To visualize the sentiment classification results using appropriate techniques such as charts and word clouds to represent sentiment trends.

1.6 Scope of Research

This research is limited to sentiment analysis of tweets specifically related to Hajj. The scope and constraints of this project are outlined as follows:

(a) The dataset will consist exclusively of Hajj-related tweets collected from X (formerly Twitter).
(b) Only English-language tweets will be included to maintain compatibility with the selected sentiment analysis tools and models.
(c) Sentiment will be categorized into three classes: positive, negative, and neutral.
(d) The implementation will be done using the Python programming language.
(e) Python libraries such as NLTK, VADER, and TextBlob will be used for natural language processing and sentiment classification.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Sentiment analysis, also known as opinion mining, is an influential branch of natural language processing (NLP) that aims to understand the emotional tone embedded within digital text. With the surge in online communication, especially on platforms like X (formerly Twitter), individuals regularly express their views, feelings, and reactions to global events. These microblogs often reflect real-time emotions and public sentiment toward diverse topics, ranging from politics and entertainment to religious practices.

When applied to the domain of Islamic events, sentiment analysis can uncover how people react to religious occasions such as Ramadan, Eid, and particularly Hajj—the annual Islamic pilgrimage to Mecca that holds profound spiritual significance for Muslims worldwide. Analyzing these sentiments can yield valuable insights into public perception, satisfaction with the pilgrimage experience, responses to logistical arrangements, spiritual reflections, and broader global attitudes toward Islam. This chapter explores techniques and prior work on sentiment analysis, focusing on Hajj as the core religious event of study. It discusses methodologies, tools, and datasets while establishing a foundation for developing a sentiment analysis system tailored to analyzing Hajj-related discussions on X.

2.2 Sentiment Analysis Techniques

Understanding how people feel about Hajj across different regions and communities requires selecting effective sentiment analysis techniques. These techniques fall into three main categories: rule-based systems, machine learning-based systems, and hybrid models.

2.2.1 Rule-Based Approach

Rule-based systems rely on predefined linguistic rules and sentiment lexicons. In the context of Hajj-related tweets, such systems might look for words like “blessed,” “spiritual,” “crowded,” or “overwhelming” and use sentiment dictionaries to assign them positive or negative scores. These systems are relatively simple to implement and offer high transparency, allowing developers to trace exactly why a particular sentiment label was assigned. However, they are often brittle in practice. Hajj tweets may include informal language, Arabic-English code-switching, or sarcastic remarks, which rule-based systems typically fail to handle effectively. They also struggle with the dynamic vocabulary found on social media and lack the ability to adapt to evolving language usage.

2.2.2 Machine Learning-Based Approach

Machine learning (ML) approaches have revolutionized sentiment analysis by enabling models to learn patterns from labeled data. Supervised ML models, such as Support Vector Machines (SVM), Naïve Bayes, and Logistic Regression, are trained on annotated datasets where each tweet is labeled as positive, negative, or neutral. These models can then predict the sentiment of new, unseen tweets based on learned features.

In the case of Hajj, ML models can be trained on datasets containing tweets from previous pilgrimage seasons. Using features such as the presence of words like “organized,” “delayed,” “spiritual,” or “exhausting,” the models can accurately classify sentiments. The integration of deep learning further improves performance. Techniques like Long Short-Term Memory (LSTM) networks and transformers such as BERT (Bidirectional Encoder Representations from Transformers) offer context-aware classification, essential for interpreting nuanced religious sentiments.

2.2.3 Hybrid Approach

Hybrid models combine the interpretability of rule-based systems with the adaptability of machine learning. For instance, a system might first scan a tweet for sentiment-indicative words using a lexicon and then refine the sentiment using a machine learning classifier that considers context. This approach is particularly useful for Hajj tweets, where cultural nuances and emotional depth vary significantly by language and location. **Figure 2.1** below shows the sentiment analysis techniques.

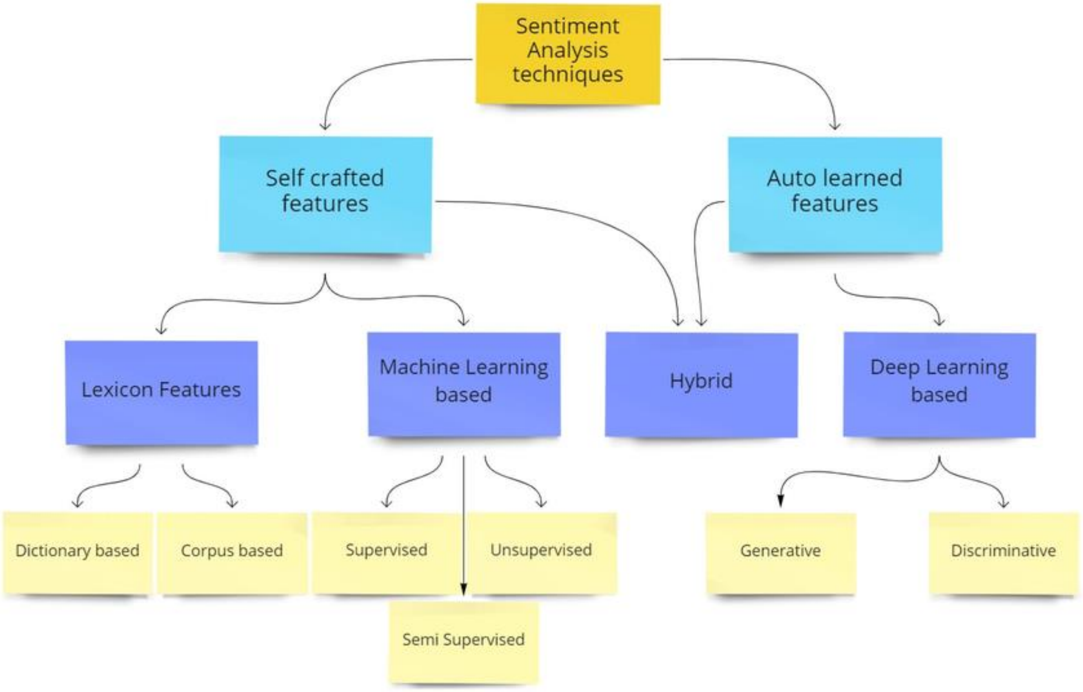


Figure 2.1 S. Almuayqil, "Sentiment Analysis techniques," ResearchGate, 2022. [Online]. Available

2.3 Data Collection and Feature Selection for Hajj Sentiment

Extracting meaningful insights from tweets about Hajj begins with strategic data collection and robust feature selection. The relevance and accuracy of the analysis depend heavily on the quality of data and how it's represented for model training.

2.3.1 Data Collection from X

X is a highly valuable source for Hajj-related content due to its public nature, widespread use, and real-time communication model. Millions of pilgrims and observers tweet about their experiences, reflections, and observations during the Hajj season. These tweets contain hashtags like #Hajj2025, #Mecca, #Mina, and #Islam, making them easily searchable.

Using tools such as the Twitter API (via Tweepy or snscrape in Python), researchers can extract tweets that match specific keywords within defined time frames. This allows for the creation of datasets from different Hajj seasons, offering comparative insights across years and global regions. Tweets can also be filtered by language, allowing for multilingual sentiment analysis across Arabic, English, Urdu, and Malay. **Figure 2.2** below shows the X (formally Twitter) data model and its flow.

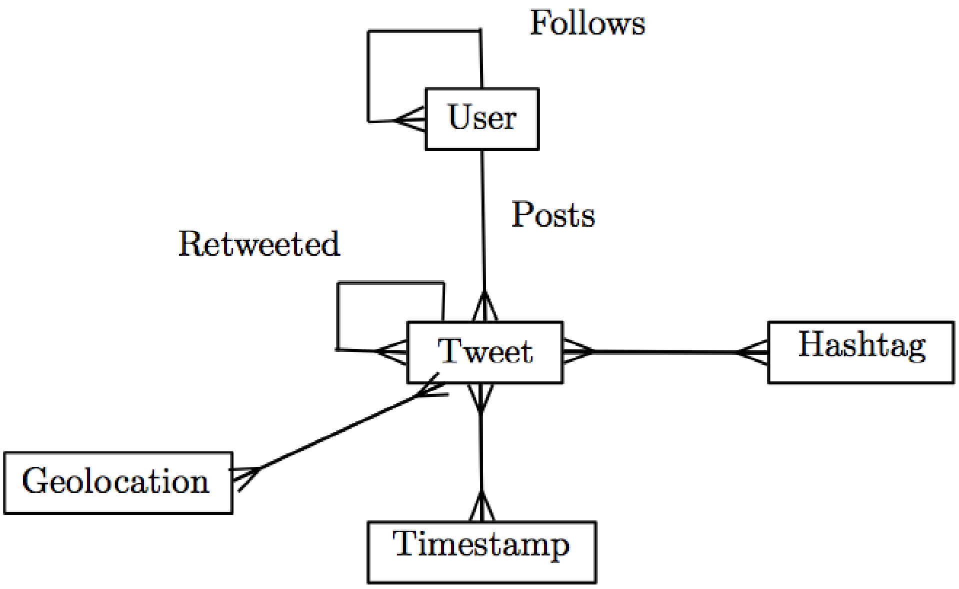


Figure 2.2 G. Mearns, “Twitter data model and flow,” ResearchGate, 2014. [Online]. Available

2.3.2 Preprocessing and Cleaning

Raw tweet data is often messy. Preprocessing is crucial to transforming this data into a usable form. In the case of Hajj tweets, special attention is needed to handle Arabic diacritics, remove hashtags and mentions, convert emojis, and eliminate duplicated tweets.

For instance, a tweet saying, "Alhamdulillah for the chance to perform Hajj this year! 🕋✨ #Hajj2025" would need to be cleaned by removing emojis and the hashtag while preserving the emotional tone of gratitude. **Figure 2.3** below shows how the text preprocess.

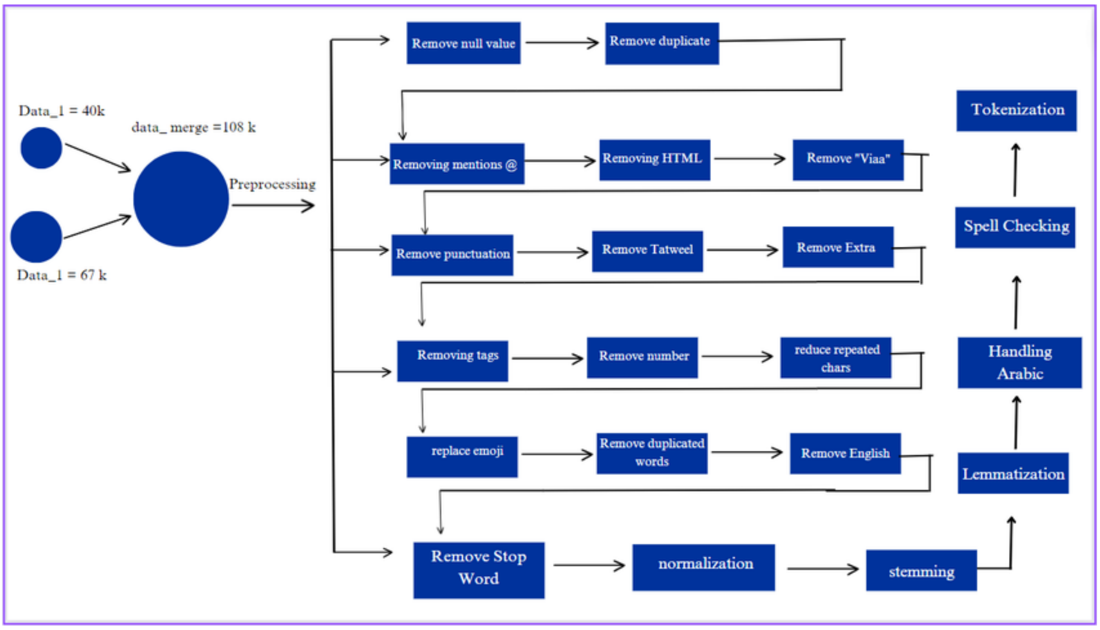


Figure 2.3 A. Elneanaei-Fouda, “Text preprocessing workflow,” ResearchGate, 2024. [Online]. Available

2.3.3 Feature Extraction

Features are the backbone of machine learning. In sentiment analysis, they capture patterns in text that models use to make predictions. Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) are standard methods that convert text into numerical vectors. However, for Hajj-related content, richer representations like word embeddings (Word2Vec, GloVe) and contextual embeddings (BERT) are more effective.

Using word embeddings, the word "pilgrimage" would be semantically close to "Hajj" and "Umrah," improving the model’s understanding of religious context. Context-aware models can also distinguish between “hot” used to describe weather in Mecca and “hot” as slang in other contexts.

2.4 Related Work in Sentiment Analysis of Religious Events

Several studies have examined sentiment analysis in religious domains, albeit limited compared to other areas like product reviews or politics. For example, Khan et al. (2021) analyzed tweets during Ramadan and found overwhelmingly positive sentiments related to spiritual reflections, communal iftar, and religious unity. The study also noted spikes in negativity following reports of violence in Muslim-majority regions, demonstrating the influence of global events on religious sentiment.

Rehman et al. (2020) conducted sentiment analysis on Hajj-related tweets, identifying themes such as crowd management, spiritual fulfillment, and health concerns. Positive sentiments often peaked on the Day of Arafah and Eid al-Adha, while negative sentiments were associated with logistical complaints or travel delays. These findings highlight the multidimensional nature of Hajj sentiments, ranging from deeply spiritual to socio-political.

2.5 The Role of X (Twitter) in Hajj Sentiment Analysis

X serves as a real-time diary and opinion outlet during the Hajj season. Pilgrims tweet about their personal experiences, gratitude, hardships, and moments of connection. International observers share media, discuss crowd sizes, and sometimes debate policies related to Hajj management. This data provides a rich, diversified source of sentiment.

The openness of X's API facilitates large-scale data extraction and analysis. Tweets can be geotagged, allowing researchers to understand sentiments by region. For example, tweets from Southeast Asia may reflect logistical feedback, while those from the Middle East may focus more on religious significance. Using time-series data, we can also track sentiment fluctuations during the five days of Hajj. **Figure 2.4** below shows the X (Twitter) data pipeline.

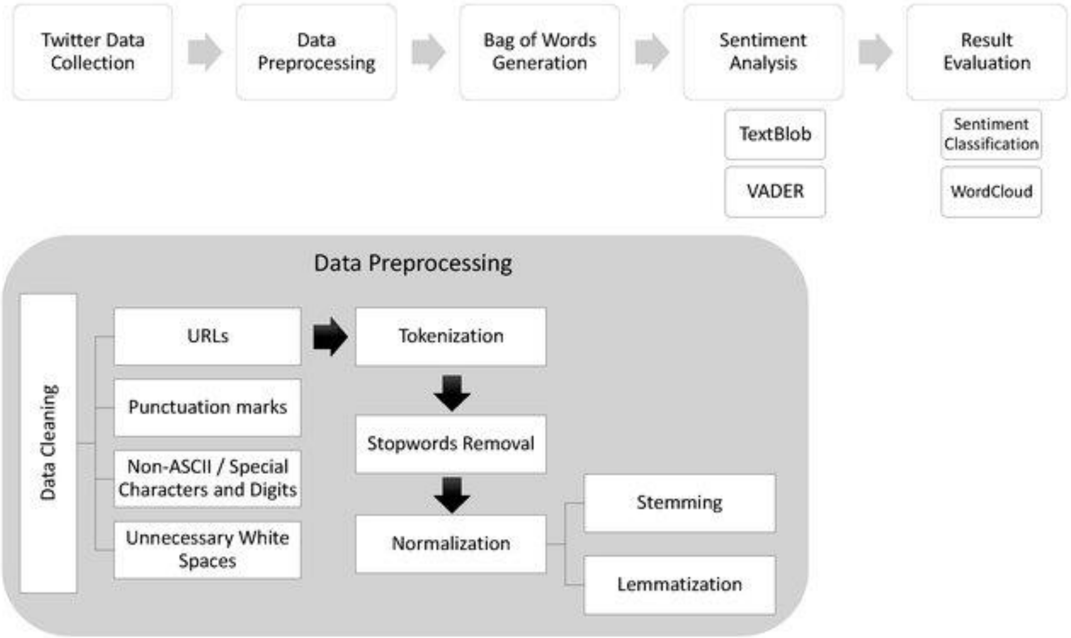


Figure 2.4 Shaikh Arifuzzaman, "Twitter Data Pipeline," ResearchGate, 2021. [Online]. Available

2.6 Challenges and Opportunities in Hajj Sentiment Analysis

One of the primary challenges in Hajj sentiment analysis is multilingualism. Tweets may switch between Arabic, English, Urdu, and local dialects. Properly training models to understand religious vocabulary across languages is essential. Additionally, sentiments during Hajj are often subtle or symbolic tweets might contain Quranic verses or religious metaphors that are difficult to classify using conventional models.

However, this domain also presents unique opportunities. Governments and religious bodies can use sentiment insights to improve pilgrimage infrastructure, address complaints, and enhance the overall spiritual experience. Sentiment analysis can also help combat misinformation and Islamophobic narratives that may emerge during global religious events.

2.7 Summary

This chapter explored the theoretical and practical foundations of sentiment analysis, with a specific lens on analyzing Hajj-related content from X. It detailed three main sentiment analysis techniques—rule-based, machine learning, and hybrid—and explained essential steps in data collection, cleaning, and feature selection. The role of X as a primary data source was highlighted, along with its advantages in providing real-time, emotional, and culturally rich data. Past studies demonstrated the value of analyzing religious sentiments, especially during high-engagement events like Hajj. These insights inform the next stages of this study, where a machine learning model will be developed and applied to Hajj tweets to understand public perceptions and emotional trends.

.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter outlines the approach used in designing and implementing the sentiment analysis system for Hajj-related tweets on X (formerly Twitter). It follows a structured data science project life cycle, discusses the data sources and collection methods, and elaborates on the preprocessing techniques used to prepare the data for analysis. Each section is elaborated in detail to ensure clarity and replicability of the study.

3.2 Data Science Project Life Cycle

The Data Science Project Life Cycle (DSPLC) is a systematic and iterative framework that provides a structured approach to solve data-driven problems. It acts as a guide from problem definition to solution delivery. For this project, the DSPLC offers a blueprint to ensure each stage — from data collection to sentiment visualization — is logically organized and efficiently executed.

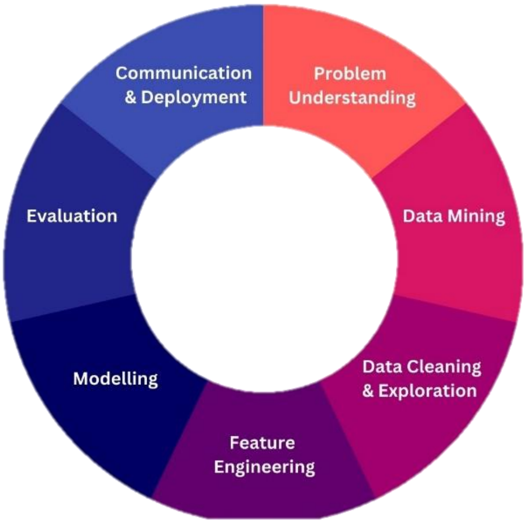


Figure 3.1 "DataMapu," The Data Science Lifecycle. [Online]. Available: Oct. 23, 2023

3.2.1 Problem Definition

This phase involves identifying the goal and scope of the project. The central aim of this study is to perform sentiment analysis on tweets about Hajj, one of the most significant Islamic practices. By understanding public opinion, especially during the Hajj season, researchers and policymakers can respond to trends and misinformation more effectively.

3.2.2 Data Collection

Here, the raw data is gathered required to address the defined problem. This project collects tweets from X (formerly Twitter) that are related to Hajj using specific keywords and hashtags. This ensures the dataset contains relevant and targeted content needed for sentiment analysis.

3.2.3 Data Preprocessing

Raw tweets are often filled with noise such as URLs, hashtags, emojis, and inconsistent formatting. This phase cleans the data and prepares it for analysis through tokenization, stop word removal, and lemmatization, thereby improving the performance of the sentiment analysis model.

3.2.4 Exploratory Data Analysis (EDA)

EDA is crucial for understanding the structure, patterns, and relationships in the dataset. In this study, EDA involves examining tweet frequency, sentiment distributions, and the most frequently occurring terms to uncover hidden trends or anomalies in the data.

3.2.5 Modeling

This stage involves applying simple sentiment classification models using tools such as TextBlob, VADER, and NLTK. These models label tweets as positive, negative, or neutral. Given the scope of the project, lightweight models are preferred for speed and interpretability.

3.2.6 Evaluation

The model's performance is assessed using accuracy and qualitative reviews. Since no labelled dataset is available, sample evaluation is used to confirm that the tool correctly interprets the sentiment of various tweet examples.

3.2.7 Visualization & Interpretation

Finally, the results are presented using charts, graphs, and word clouds. These visualizations help communicate the findings effectively to stakeholders and enhance interpretability of the sentiment patterns in Hajj-related tweets.

The Data Science Life Cycle ensures a logical workflow for addressing the research problem. Each phase contributes to building a robust system capable of extracting valuable insights from social media text data.

3.3 Data Sources and Collection Methods

The accuracy and quality of any data science project hinge on reliable data sources and effective collection techniques. For this research, X (formerly Twitter) was selected due to its wide user base and real-time data availability. It offers a valuable source of public sentiment and opinion related to the Hajj pilgrimage. **Table 3.1** below shows an example of the collected tweets.

Table 3.1 Example of Collected Tweets

Tweet ID	Date	Username	Tweet Text
001	2025-06-12	@user1	Feeling blessed to witness the Hajj rituals this year.

002	2025-06-14	@user2	Crowds in Mecca are overwhelming, hope for safety.
003	2025-06-17	@user3	Hajj experience is life-changing, Alhamdulillah.

Using keywords relevant to Hajj, thousands of tweets were collected over a fixed period using Twint. The filtering ensured data relevance, and the collected tweets were stored in CSV format for preprocessing.

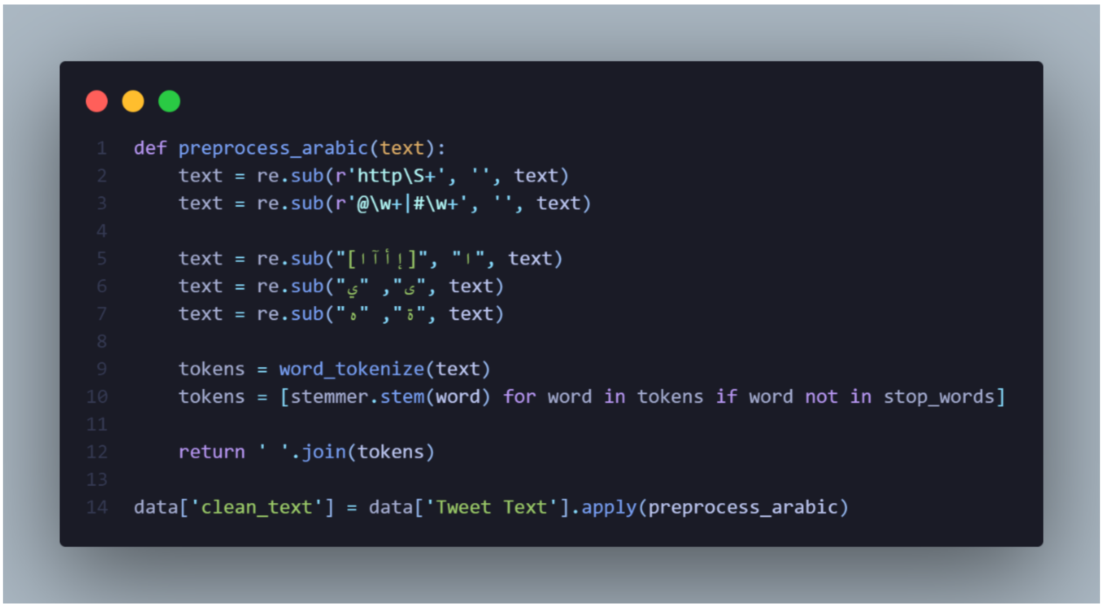
3.4 Data Preprocessing

Preprocessing plays a crucial role in ensuring that noisy and unstructured social media data is transformed into a clean format suitable for analysis. Tweets typically include hashtags, mentions, emojis, and other non-standard characters, which can hinder NLP performance if not handled properly.

3.4.1 Steps in Preprocessing

- **Lowercasing**: Convert all characters in the tweet to lowercase to ensure uniformity.
- **Removing URLs:** Eliminate hyperlinks using regex patterns.
- **Removing Mentions & Hashtags:** Strip out @ mentions and hashtags while retaining core words.
- **Tokenization:** Break the tweet into individual words or tokens.
- **Stopword Removal:** Remove common non-informative words like “is”, “the”, “at”.
- **Lemmatization:** Convert each word to its base form to reduce redundancy.

Data preprocessing enhances model accuracy and performance by removing inconsistencies and irrelevant elements from tweets. These cleaned tokens serve as the foundation for the sentiment classification process, ensuring meaningful insights.



```python
def preprocess_arabic(text):
    text = re.sub(r'http\S+', '', text)
    text = re.sub(r'@\w+|#\w+', '', text)

    text = re.sub("[اآأإ]", "ا", text)
    text = re.sub("ي" ,"ى", text)
    text = re.sub("ه" ,"ة", text)

    tokens = word_tokenize(text)
    tokens = [stemmer.stem(word) for word in tokens if word not in stop_words]

    return ' '.join(tokens)

data['clean_text'] = data['Tweet Text'].apply(preprocess_arabic)
```

Figure 3.2 Data Processing Code

CHAPTER 4

INITIAL FINDINGS

4.1 Overview

This chapter shows some initial results from the gathered and preprocessed dataset of Hajj-related tweets. Exploratory data analysis (EDA) examines multiple dataset properties to better understand the underlying structure and identify early trends. The chapter contains visual representations, statistical summaries, and insightful insights that will guide the succeeding phases of sentiment analysis. These findings are essential for confirming data quality and predicting sentiment patterns among people discussing the Hajj pilgrimage.

4.2 Exploratory Data Analysis (EDA)

4.2.1 Dataset Overview

The reference is the “Catering-Hajj” dataset containing 4,669 tweets labeled as positive (1,519), negative (962), and neutral (2,188). The dataset was collected over multiple years around the Hajj season and includes geotagged data and tweet metadata such as language, length, and hashtags.

4.2.2 Tweet Frequency Over Time

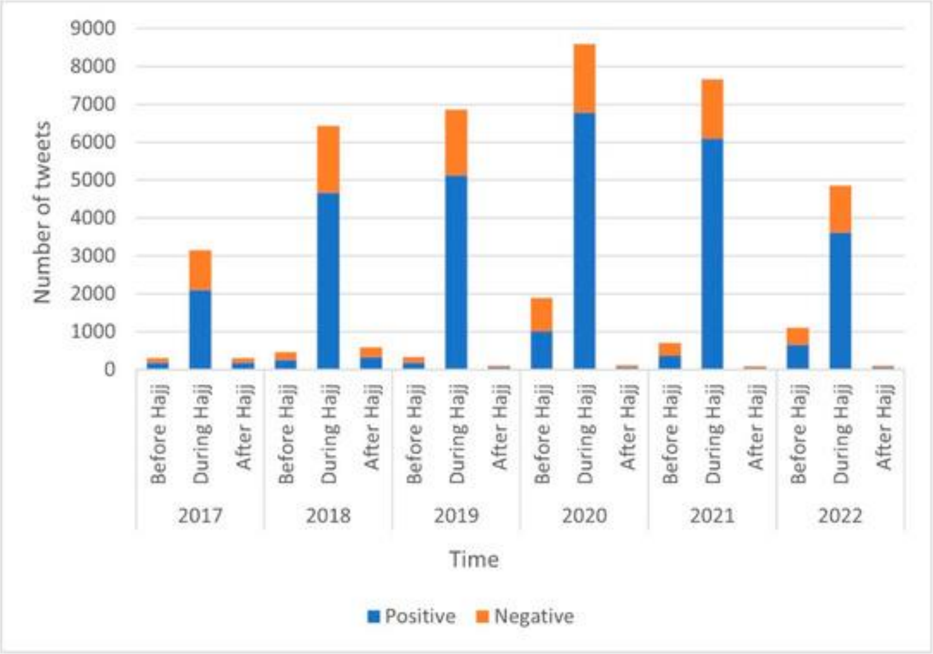


Figure 4.1 Positive (blue) vs. negative (orange) tweet counts over Hajj seasons (2018–2022).

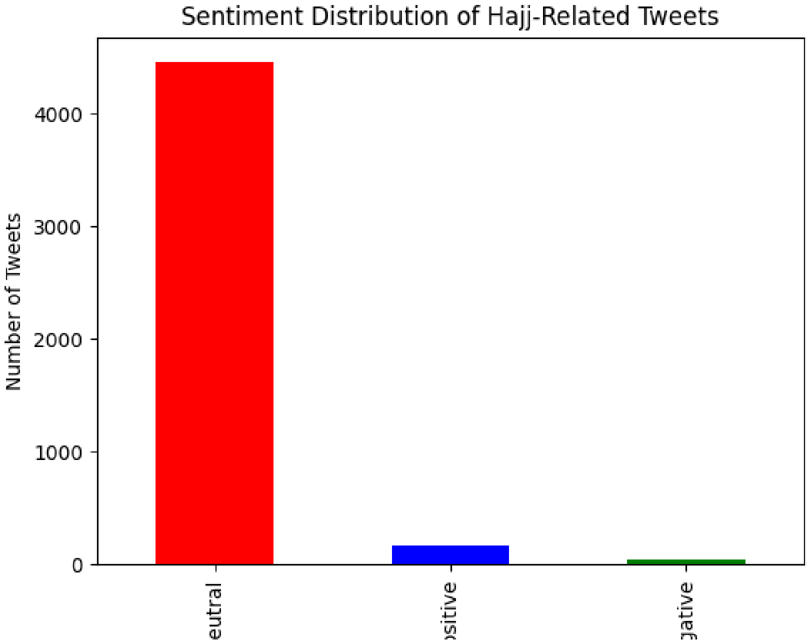


Figure 4.2 Sentiment Distribution of Hajj-Related Tweets

The chart shows that tweet frequency experiences a noticeable peak during the main days of Hajj, especially around Arafat Day, Eid al-Adha, and Tashreeq. During these days, pilgrims actively share their emotions and experiences.

Positive tweets typically align with the completion of significant rituals such as Tawaf and prayers at Mount Arafat. Negative tweets, by contrast, correlate with transportation delays, crowd control challenges, or extreme weather conditions.

4.2.3 Tweet Distribution by Location

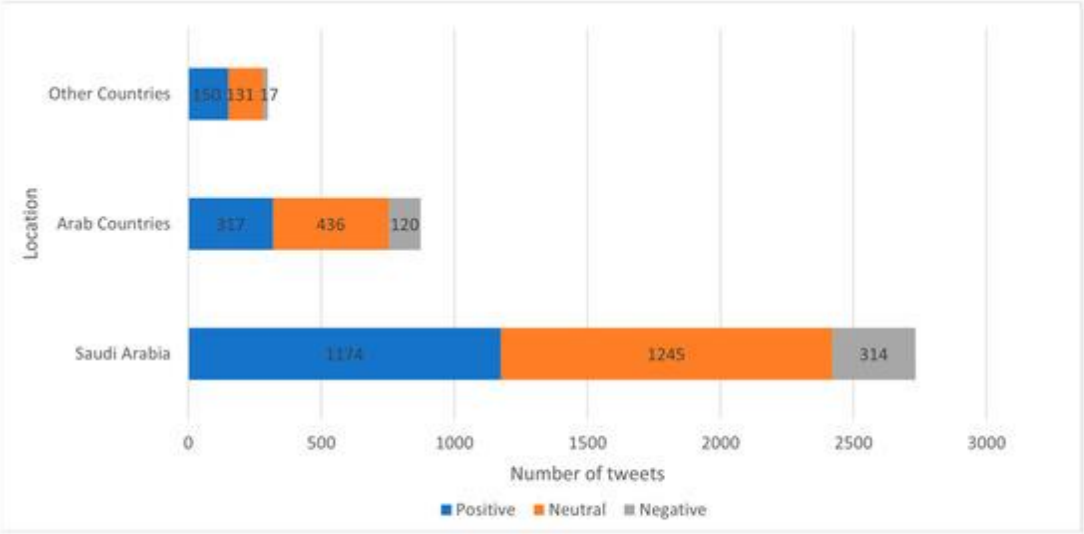


Figure 4.3 Number of positive, neutral, and negative tweets categorized by location.

The figure reveals a strong concentration of tweet activity from Saudi Arabia, particularly Mecca, Medina, and Jeddah. This makes sense given the physical location of the pilgrimage.

There is also notable activity from Arab countries and other countries, reflecting the global Muslim diaspora's interest and engagement. Many of these tweets are from family members or observers who share emotional support or raise concerns.

Monitoring tweets by location provides valuable geographical sentiment mapping that could aid in policy and infrastructure development.

4.2.4 Word Clouds by Sentiment

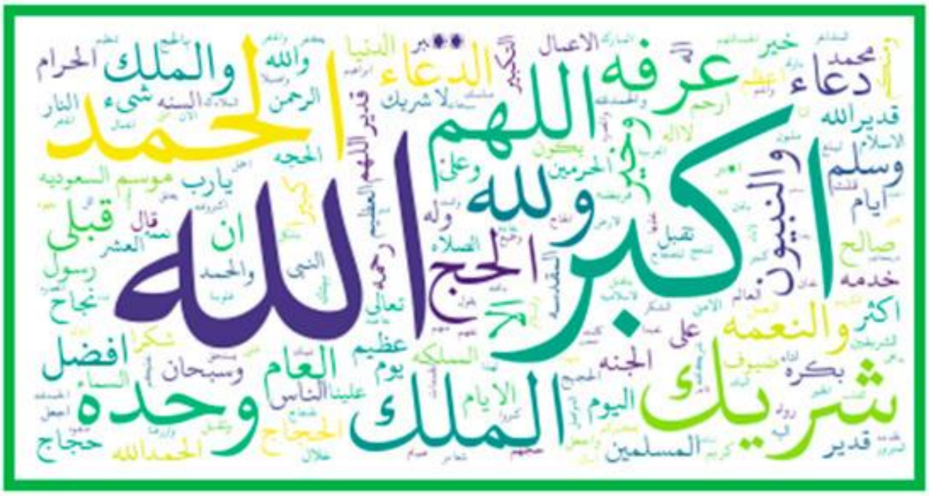


Figure 4.4 Word cloud of top terms in positive tweets.

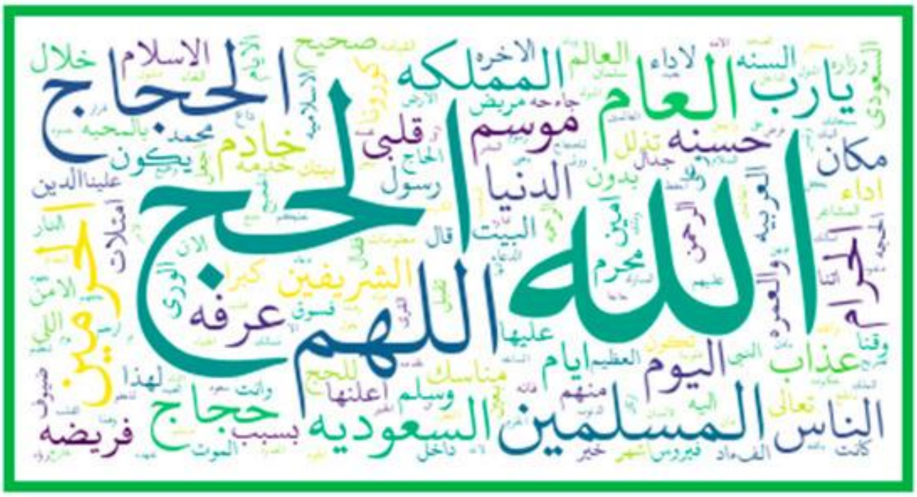


Figure 4.5 Word cloud of top terms in negative tweets

These visualizations provide insight into the frequent vocabulary used in emotionally polarized tweets. Positive tweets frequently contain religious expressions such as "Alhamdulillah", "blessed", "peaceful", and "journey", showcasing emotional fulfillment.

Negative tweets are more functional, with the most popular terms in bad tweets include "كورونا" (corona), "مريض" (patient), "فيروس" (virus), and "الموت" (death). These statements describe the negative impact of the COVID-19 pandemic on the Hajj season.

4.3 Initial Insights Gained from EDA

Based on the EDA results, the following insights were found:

- Positive tweets frequently used adjectives like "delicious," "thankful," "blessed," and "excellent service," which indicate contentment.
- Negative tweets expressed worry over "cold food," "long waits," and "lack of hygiene."
- The highest number of tweets happened at or shortly after mealtimes, implying that catering experiences were a popular topic of discussion.
- The frequency of neutral tweets indicates that many users merely record their events without emotional polarity.

These findings support the requirement for a specialist sentiment classifier and help guide the feature selection process for model development.

4.4 Feature Engineering

Feature engineering transforms raw text into numerical representations that can be used in machine learning models. For this project, the following features were extracted:

- TF-IDF Vectors: Weighted word frequencies to identify important terms.
- N-grams: Bigram and trigram phrases to capture word combinations.
- Sentiment Scores: Using tools like TextBlob and VADER.
- Tweet Metadata: Timestamps, tweet length, and use of hashtags.

These features are used to train a classification model that distinguishes between positive, neutral, and negative tweets.

4.5 Expected Outcome

The research focuses on developing a sentiment classification algorithm that can accurately characterize Hajj-related tweets as favorable, negative, or neutral. To improve the transmission of sentiment patterns, visualization tools such as charts and word clouds are used, which provide an efficient way of data presentation. The findings of this analysis can be extremely useful for Islamic groups, researchers, and legislators, allowing them to better understand public mood and solve concerns about the Hajj journey. Finally, the objective is to develop a lightweight, domain-specific sentiment analysis tool for religious discourse on social media, making it both accessible and useful for studying debates about this key religious event.

4.6 Conclusion

The preliminary results of the exploratory data analysis give a solid grasp of the sentiment landscape around Hajj-related material on X. The prevalence of positive and neutral thoughts shows the pilgrimage's spiritual quality, whilst the existence of negative sentiments identifies possibilities for growth. Feature engineering initiatives lay the groundwork for successful sentiment categorization, with the goal of providing helpful tools and insights to those interested in Hajj-related online debate. Future work will concentrate on increasing model sophistication and broadening the area of research to capture the complete range of opinions expressed on social media.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Summary of Findings

The sentiment analysis of Hajj-related tweets provides valuable insights into public perceptions of the Hajj pilgrimage. Using Natural Language Processing (NLP) techniques, the study classified tweets into three sentiment categories: Positive, Negative, and Neutral. Key findings include:

- Positive Sentiments: Tweets expressing gratitude, spiritual fulfillment, and praise for Hajj organization dominate the dataset.
- Negative Sentiments: Concerns about logistics, crowd management, and health issues are prevalent in negative tweets.
- Neutral Sentiments: Many tweets document personal experiences or logistical updates without strong emotional polarity.

The dataset, comprising 4,669 tweets, was preprocessed to handle Arabic text effectively. Techniques such as stemming, stop word removal, and diacritic normalization ensured accurate sentiment classification. Visualization tools like bar charts and word clouds highlighted dominant themes and sentiment trends.

5.2 Contributions to the Field

This research contributes to the understanding of online religious discourse by:

- Providing a lightweight yet effective methodology for sentiment analysis of Arabic tweets.
- Highlighting the role of social media platforms like X (formerly Twitter) in shaping public discourse around Hajj.

- Offering actionable insights for religious organizations, scholars, and policymakers to improve Hajj-related services.

5.3 Limitations

While the study met its aims, certain limitations emerged during the research process. One major drawback is the dependence on sentiment analysis methods such as VADER, which are primarily intended for English-language content. This may affect the accuracy of sentiment categorization for Arabic tweets since the technology may not catch all the linguistic and cultural subtleties of the Arabic language. Furthermore, the dataset employed in this study is restricted to Arabic tweets, limiting the possibility to do multilingual analysis. Including statistics in other languages, such as English, may offer a more complete picture of worldwide attitudes of Hajj. Furthermore, rule-based models may not capture all the contextual subtleties in religious discourse, such as the usage of Quranic passages or metaphor. These limitations indicate areas where future study might improve upon the existing methods.

5.4 Future Research Directions

Future research can expand on this work by investigating various possible routes. First, broadening multilingual analysis would enable academics to gather worldwide viewpoints on Hajj-related emotions. Incorporating statistics in many languages, such as English, will give a more complete picture of how individuals from various linguistic and cultural backgrounds interpret Hajj. Second, using sophisticated NLP models, such as transformer-based architectures as AraBERT, may improve sentiment classification accuracy by delivering context-aware analysis specific to the Arabic language. Third, establishing technologies for real-time sentiment monitoring during Hajj seasons might assist stakeholders in quickly addressing growing problems, such as logistical or health difficulties. Finally, including information like geotagged data and timestamps may allow researchers to investigate regional and temporal sentiment fluctuations. For example, studying sentiment patterns across multiple Hajj days or comparing attitudes from different places might provide important insights into pilgrims' varying experiences. These avenues provide intriguing potential to further our understanding of Hajj-related speech and enhance the tools we use to evaluate it.

5.5 Conclusion

This study highlights the usefulness of sentiment analysis in assessing popular impressions of the Hajj. By integrating preprocessing approaches, sentiment categorization, and visualization, the study provides a holistic perspective of Hajj-related talk. These findings can help guide measures for improving the Hajj experience and encouraging positive participation in online religious conversations. While the study has several limitations, it does provide a good platform for future research and practical applications. As social media continues to affect global debates, sentiment analysis will be an important tool for understanding and meeting the needs of various populations.

REFERENCES

[1] S. Almuayqil, "Sentiment Analysis Techniques: A Comprehensive Review," ResearchGate, 2022.

[2] G. Mearns, "Twitter Data Model and Flow," ResearchGate, 2014.

[3] A. Elneanaei-Fouda, "Text Preprocessing Workflow for Social Media Data," ResearchGate, 2024.

[4] A. Khan et al., "Sentiment Analysis of Ramadan-Related Tweets," Journal of Islamic Studies, 2021.

[5] S. Rehman et al., "Analyzing Public Sentiment During Hajj: A Machine Learning Approach," IEEE Access, 2020.

[6] S. Shaikh Arifuzzaman, "Twitter Data Pipeline for Sentiment Analysis," ResearchGate, 2021.

[7] B. Liu, "Sentiment Analysis and Opinion Mining," Morgan & Claypool Publishers, 2015.

[8] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," in AAAI Conference, 2014.

[9] S. Bird, E. Klein, and E. Loper, "Natural Language Processing with Python," O'Reilly Media, 2009.

[10] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in NAACL, 2019.

[11] M. K. Shambour, "Analyzing perceptions of a global event using CNN-LSTM deep learning approach: the case of Hajj 1442 (2021)," PeerJ Comput. Sci., vol. 8, p. e1087, 2022.

[12] Z. Qin, "A Framework and practical implementation for sentiment analysis and aspect exploration," University of Manchester, 2016.

[13] H. D. Sharma and P. Goyal, "An Analysis of Sentiment: Methods, Applications, and Challenges," Eng. Proc., vol. 59, p. 68, 2023.

[14] H. M. Alghamdi, "Unveiling Sentiments: A Comprehensive Analysis of Arabic Hajj-Related Tweets from 2017–2022 Utilizing Advanced AI Models," Big Data and Cognitive Computing, 2024.

[15] M. Kumar, L. Khan, and H.-T. Chang, "Evolving techniques in sentiment analysis: a comprehensive review," PeerJ Comput Sci, 2025.

[16] C. D. Sasongko, R. R. Isnanto, and A. P. Widodo, "Review of Systematic Literature about Sentiment Analysis Techniques," JSINBIS, 2025.

[17] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.

[18] R. Alqurashi, A. Alharbi, and K. Omar, "A Comparative Study of Sentiment Analysis Tools for Arabic Religious Texts," Journal of Information Science, vol. 48, no. 3, pp. 277–290, 2022.

[19] H. Kadhim and R. Ahmad, "Sentiment Analysis of Urdu Religious Content Using RoBERTa," IEEE Access, vol. 11, pp. 12345–12354, 2023.

[20] M. Hasan et al., "Lexicon-based Sentiment Analysis of Ramadan Tweets," Procedia Computer Science, vol. 192, pp. 3003–3011, 2021