SENTIMENT ANALYSIS OF HAJJ-RELATED CONTENT ON X

MOHAMED TAREK ELSAYED MOHAMED TORKY

UNIVERSITI TEKNOLOGI MALAYSIA

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

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.

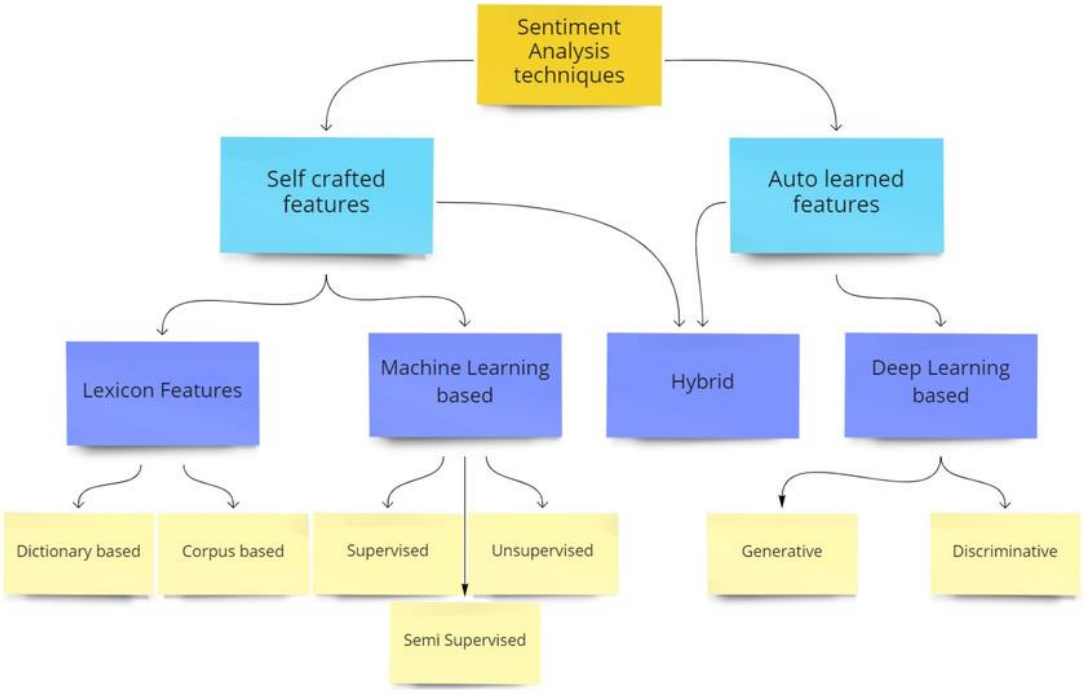

Figure 2.1 S. Almuayqil, "Sentiment Analysis techniques," ResearchGate, 2022. [Online]. Available

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.

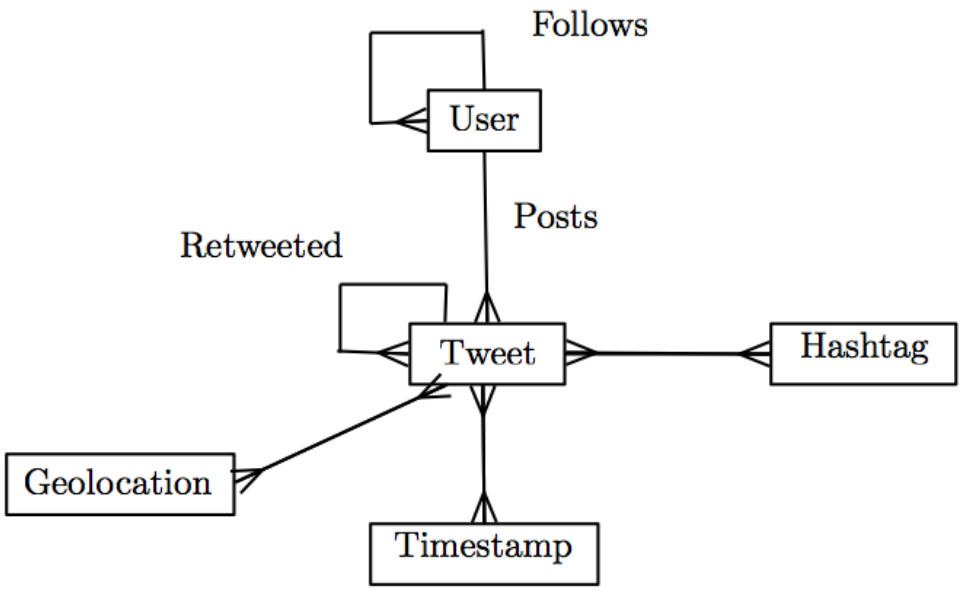

Figure 2.2 G. Mearns, “Twitter data model and flow,” ResearchGate, 2014. [Online]. Available

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.

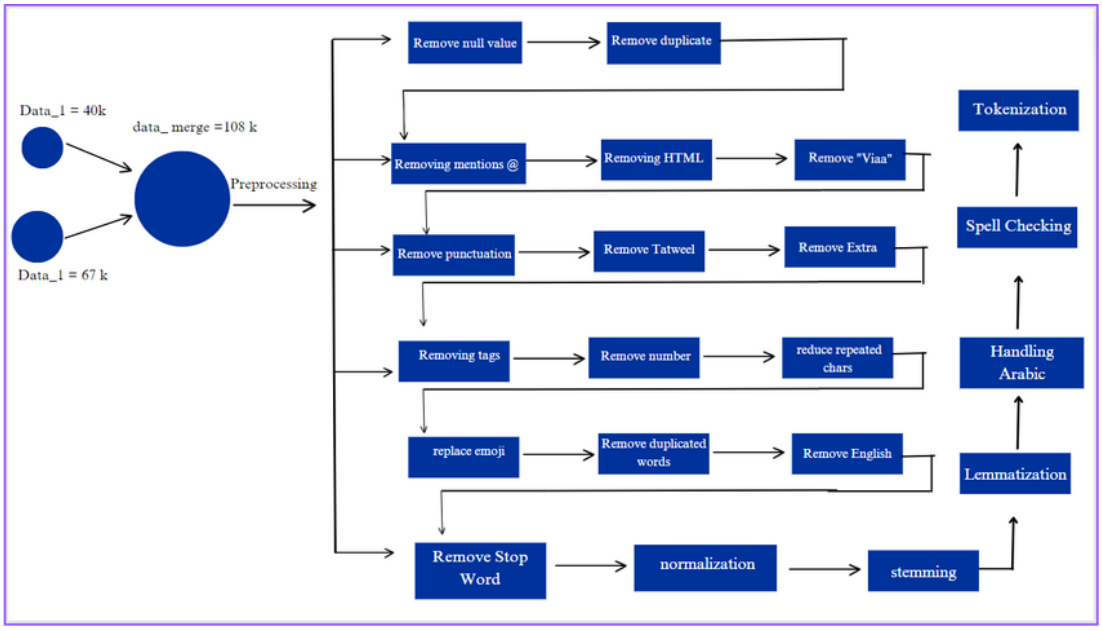

Figure 2.3 A. Elneanaei-Fouda, “Text preprocessing workflow,” ResearchGate, 2024. [Online]. Available

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.

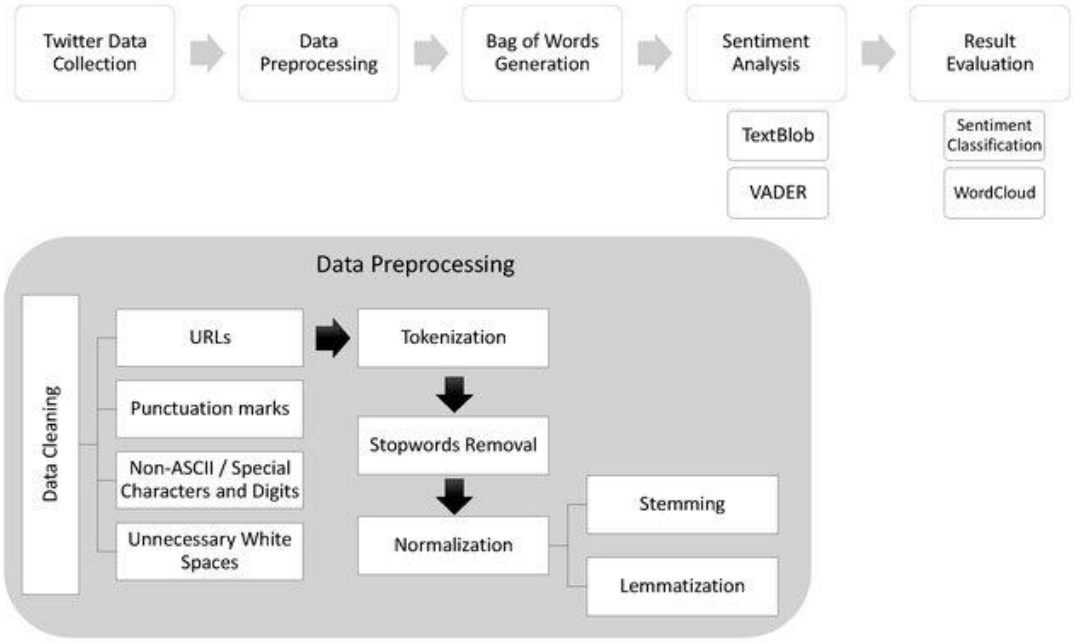

Figure 2.4 Shaikh Arifuzzaman, "Twitter Data Pipeline," ResearchGate, 2021. [Online]. Available

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.