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 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.
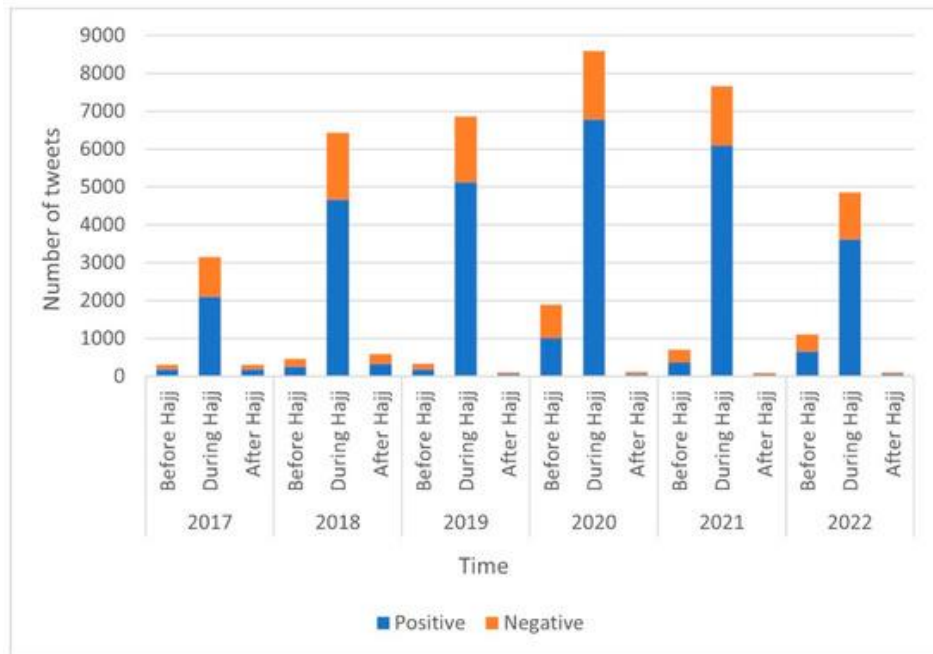
4.2.2 Tweet Frequency Over Time

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

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.
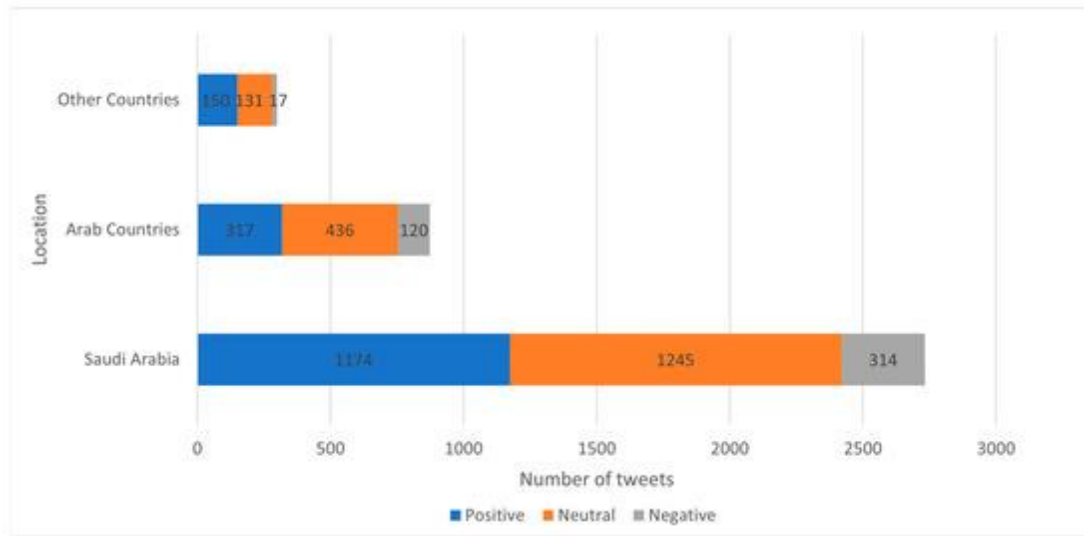
4.2.3   Tweet Distribution by Location

Figure 4.2   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 4.3   Word cloud of top terms in positive tweets.

Figure 4.4   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.