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 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.

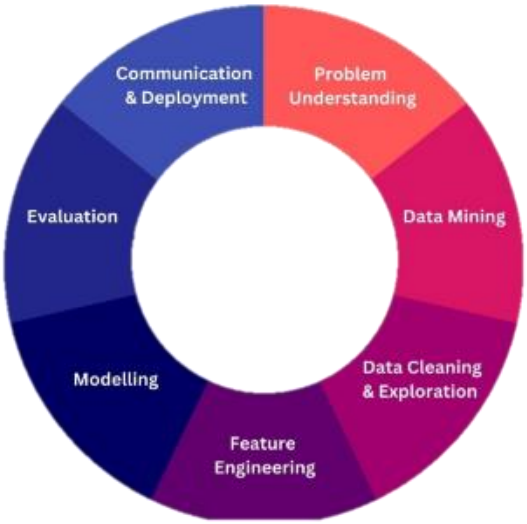

Figure 3.1 "DataMapu," The Data Science Lifecycle. [Online]. Available: Oct. 23, 2023

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, we gather the raw data 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.