Chapter 3

METHODOLOGY

This chapter outlines the methodology used in the study to analyze social media data, specifically focusing on Twitter. The methodology is structured according to the Data Science Life Cycle, which includes stages such as data collection, pre-processing, model building, and analysis. Each of these stages plays a critical role in ensuring that the analysis is both efficient and accurate. This methodology also employs deep learning models, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs), for analyzing social media content and deriving meaningful insights.

3.1 Data Science Life Cycle

The Data Science Life Cycle consists of several key stages that guide the process of extracting useful insights from raw data. The following steps are crucial in the context of this study:

1. Data Collection

The first step in the Data Science Life Cycle involves the collection of relevant data. For this study, we will use an agent framework to subscribe to social media data, specifically from Twitter. The agent framework will be responsible for continuously collecting data from Twitter based on specific social media topics. The key features of the data will be tweets, including text, media (images and videos), emojis, and associated metadata (e.g., user account information, timestamp, geolocation, etc.).

To collect the data, the study will use the **Twitter API** and relevant libraries in Python, such as **Tweepy**. The agent framework will run periodically to subscribe to the topics of interest and fetch data at regular intervals. This approach ensures a continuous stream of fresh data for analysis.

The primary topics selected for this study are:

- **America Election 2024**

- **Bitcoin Market**
- **Tesla Stock Price**

These topics have been identified based on their relevance to current global discussions, financial markets, and politics. The agent will collect tweets related to these topics from a variety of accounts, ensuring a diverse representation of opinions and interactions.

2. Data Pre-processing

Data pre-processing is a critical step in the data science workflow. Raw social media data often contains noise and irrelevant information, which can negatively affect the analysis. Therefore, the data needs to be cleaned and prepared before any analysis can take place.

The pre-processing steps for this study include:

- **Data Filtering:**
  The first step in data pre-processing will be filtering out irrelevant posts, such as spam, ads, and retweets. This will be achieved by setting up specific rules and filters based on keywords, hashtags, and account type. Posts that do not meet the criteria (e.g., unrelated to the selected topics or contain excessive promotional content) will be excluded.
- **Text and Emoji Extraction:**
  Social media posts often include a combination of text, images, and emojis. To perform an accurate analysis, the study will extract the textual content and emojis from each tweet. Emojis are a critical aspect of social media communication, as they often carry emotional or contextual meaning that can enhance sentiment analysis. Both the text and emojis will be saved in separate fields within the dataset.
- **Topic Classification:**
  Since the data will be collected from various sources on different topics, it is essential to classify and group the posts into their corresponding topics. The classification will be based on keywords, hashtags, and content relevance. Tweets that do not align with the topics of interest will be removed from the dataset.
- **Normalization:**
  Text data will be pre-processed by converting all text to lowercase, removing stop words (common words like "the," "and," etc.), stemming (reducing words to their root

forms), and removing punctuation. This process helps in standardizing the text and reducing dimensionality.

- **Handling Missing Data:**
  Some tweets may have incomplete data, such as missing text or missing user information. Missing data will be handled through imputation (where applicable) or removal, depending on the context of the missing values.

After pre-processing, the data will be categorized into different types: **text data**, **media data**, **emoji data**, and **unstructured data** (e.g., metadata or mixed content).

3. Build Analysis System

Once the data is cleaned and pre-processed, the next step involves setting up the analysis system. This stage includes building the necessary infrastructure to apply machine learning models for sentiment analysis, topic modeling, and predictive analytics.

- **Deep Learning Model:**
  The study will employ **Recurrent Neural Networks (RNNs)** and **Long Short-Term Memory (LSTM)** networks, both of which are effective for sequence prediction tasks like text classification and sentiment analysis. RNNs are well-suited for processing sequential data, while LSTMs, a type of RNN, are designed to overcome the issue of vanishing gradients and are better suited for analyzing longer sequences.
- **Model Architecture:**
  The LSTM network will consist of several layers, including an embedding layer (to convert text into dense vectors), one or more LSTM layers (to capture temporal dependencies in the data), and a final dense layer (to output the results of the analysis). The architecture will also include a softmax activation function for multi-class classification (e.g., sentiment analysis with positive, negative, and neutral classes).
- **Training and Validation:**
  The model will be trained on a labeled dataset (tweets that have been pre-labeled with sentiment or topic labels). The training process will involve optimizing the model's weights using backpropagation and gradient descent. The model will be validated using a separate validation dataset to prevent overfitting and ensure generalizability.
- **Database Integration:**
  To ensure that the data can be efficiently processed and retrieved, the study will

integrate the model with a **relational database** (such as MySQL or PostgreSQL) to store the collected and pre-processed data. The database will also hold the results of the analysis, which will be used for further exploration and reporting.

4. Analysis Process

Once the system is built and trained, the next step is to run the analysis. This phase will involve using the trained models to analyze the topic-based social media data collected in the earlier steps.

The analysis will focus on:

- **Sentiment Analysis:**
  Using the trained LSTM model, the sentiment of each post will be determined (positive, negative, or neutral). Sentiment analysis will help to understand the general public's opinion about the topics of interest (e.g., the 2024 U.S. election, the Bitcoin market, and Tesla's stock price).
- **Topic Modeling:**
  Topic modeling techniques, such as Latent Dirichlet Allocation (LDA), will be used to identify underlying themes and topics in the tweets. This helps in identifying the most discussed subtopics within each of the main topics.
- **Trend Analysis:**
  Temporal patterns and trends will be extracted to identify how discussions evolve over time. For example, the sentiment around the 2024 U.S. election might shift as key events unfold.

5. Final Report

The final phase of the methodology is to generate a detailed report summarizing the results of the analysis. This report will include:

- **Data Visualizations:**
  Visual representations of the data, such as bar charts, line graphs, and word clouds, will be created to illustrate the trends and findings from the analysis. These charts will be generated using Python libraries like **Matplotlib** and **Seaborn**.
- **Analysis Summary:**
  The key findings from the sentiment analysis and topic modeling will be summarized

in the report. For example, the report will highlight the general sentiment around Tesla's stock price or the Bitcoin market during specific periods.

- **Recommendations:**
  Based on the insights from the analysis, recommendations will be provided to help enterprises and governments make informed decisions. For instance, governments may use the findings to assess public sentiment on upcoming elections or policy changes, while enterprises may use the insights to gauge market sentiment or customer preferences.

3.2 Data Sources

The primary data source for this study is **Twitter**, a platform that allows users to post short messages (tweets) on various topics. The data will be collected based on three main topics identified as current "hot" topics in social media discussions:

1. **America Election 2024**
2. **Bitcoin Market**
3. **Tesla Stock Price**

The data will be collected from 50 Twitter accounts per topic, resulting in a total of 150–200 posts per topic, ensuring that the dataset is large enough to provide meaningful insights while remaining manageable for analysis.

3.3 Evaluation

The number of true positives (TP) is the number of samples which are labeled as positive by our model and are also annotated as positive by the creators of the dataset. The number of true negatives (TN) is the number of samples labeled as negative and annotated as negative. The number of false positives (FP) is the number of samples labeled as positive, but annotated as negative. Finally, the number of false negatives (FN) is the number of samples which are labeled as negative, but are annotated as positive. In the case of our binary classification task, TP and FP refer to the positive class, while TN and FN refer to the negative class.However, in the general N-class scenario, these terms can refer to any of the classes.

Equations (1)–(5) define the metrics. The ratios are multiplied by 100 to present the results in percentages.

The accuracy (Acc) is the percentage of samples which were correctly labeled by our model out of the total number of evaluated samples.

3.4 Tools and Technologies

3.4.1 Python (for data collection, pre-processing, and analysis)

3.4.2 Tweepy (Twitter API wrapper)

3.4.3 Keras and TensorFlow (for building and training deep learning models)

3.4.5 PostgreSQL (for database storage)

3.4.6 Matplotlib (for data visualization)

3.4 Chapter Summary

This chapter establishes the foundational framework for developing and evaluating Long Short-Term Memory (LSTM) models, designed to analyze trends within social media data collected from the Twitter platform. The methodology follows a structured approach, beginning with data collection, where an agent framework is used to subscribe to tweets related to selected topics such as the America Election 2024, Bitcoin Market, and Tesla Stock Price. Relevant posts, including text, media (images and videos), and emojis, are extracted and stored in a database for analysis.