

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter introduces the combination of research methods used to achieve the objectives of this study, focusing on analyzing topic-based social media posts through RNN and LSTM models. The use of deep learning techniques is the most common approach in this research field. RNN and LSTM can parse and recognize complex data from social media platforms. This chapter introduces the research framework and discusses the development of the model. It also elaborates on how to set the input indicators and discusses the reliability and usability of the framework.

3.2 Research Framework

The methodology consists of a series of systematically designed phases, each of which is essential to the progress and completeness of the study. The methodology begins by identifying the topics of social media posts. This initial phase involves a meticulous analysis of various types of tweet variables and their potential impact on the accuracy of the analysis, based on a thorough review of relevant literature and domain expertise.

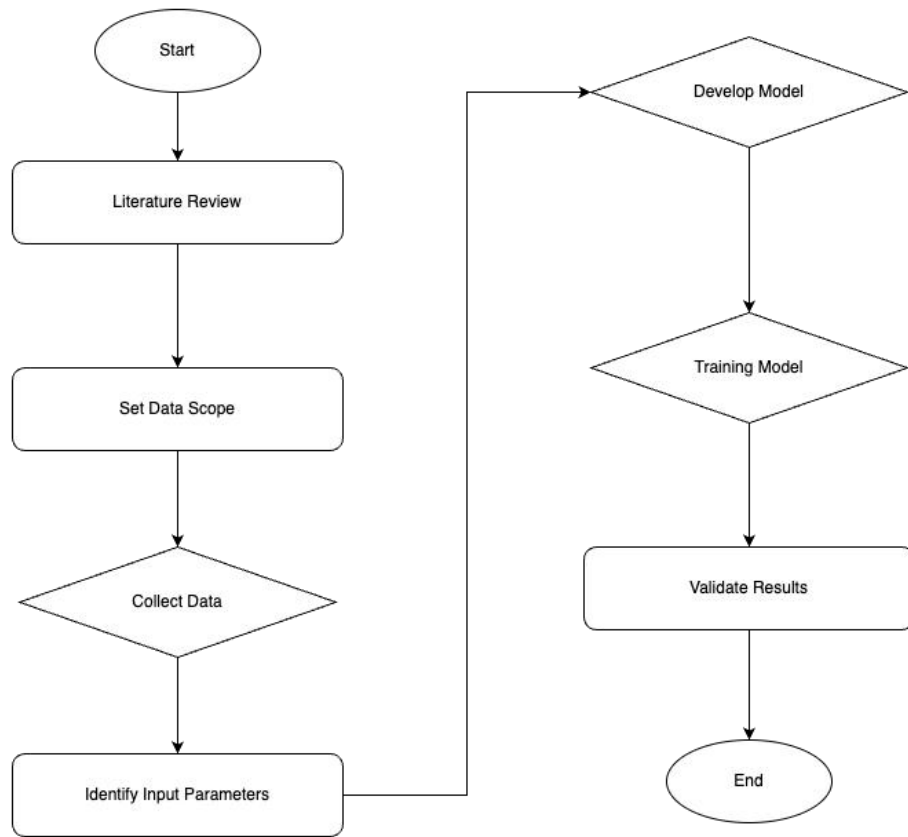


Figure 3.1 Research framework

Subsequently, the focus shifts to choosing an appropriate machine learning model. This article will use a hybrid LSTM model based on its accuracy and applicability in analyzing complex content types, as well as its proven effectiveness in predictive cases. The model will go through a testing phase of tuning and input analysis, which optimizes the model's performance to find a balance between accuracy and computational performance. Before the actual training, the input parameters will be evaluated to obtain their impact on the model's output results, and the parameters will be checked and screened for quality. After determining the model and parameters, the method enters the data collection phase. The data source is mainly the Twitter social platform. This chapter will detail the process of collecting, cleaning, and preprocessing the data to ensure that the results are consistent with the test parameters set in the previous article.

3.2.1 Data Science Lifecycle

This section provides a comprehensive description of the data science life cycle, including problem identification, data collection, data preparation, exploratory data analysis (EDA) and model development.

- (a) Problem identification defines the goals and key factors of the project. This phase determines the topics for predicting social media posts: the US election and cryptocurrency. The content data of the same topic is analyzed based on different posts to analyze the trend of user sentiment changes on the topic. The final analysis results are measured by setting trend data indicators. The setting of indicators is crucial to obtain meaningful trend analysis. When obtaining data, useless information will be filtered out through keyword collection. Although some useful data may be lost, these differences can be ignored based on a large enough sample size, and have little impact on the analysis results of the prediction model.
- (b) Data collection defines the data sources. Through the public social media platform X, the specified content is obtained through the official data interface of the platform and stored in a non-relational database. The database will use MongoDB. For some unsupported data, web crawler technology is needed to crawl. Data collection will clarify the data quality and data scope to support the subsequent preprocessing and analysis phases. By ensuring the accuracy and relevance of the data, the analysis model can work more effectively.
- (c) Data preparation defines the clean and transform the collected data, especially to remove outliers to avoid contaminating the analysis model and reducing the effectiveness of the model. At the same time, different data formats are effectively converted to allow the data analysis model to accept input

parameters and avoid model crashes. The format of the original data is formatted by the Pandas library of the Python language.

- (d) Exploratory Data Analysis (EDA) and using visualization tools can provide a more efficient view of analysis results and help researchers gain insights and identify potential data associations.
- (e) Model development is based on the structure of the Long Short-Term Memory (LSTM) neural network. The LSTM model is a type of RNN designed to solve the problem of analyzing longer sequences. It has efficient processing capabilities for complex data with complex content and very long contexts. The model is trained using an efficient framework and requires projects to input high-quality data sources for analysis.

3.3 Chapter Summary

This chapter evaluates the design and development foundation of the LSTM model and discusses how to build the project and achieve the goals. This chapter determines the scope of data collection, clarifies the data sources and quality standards, determines the data storage solution, detects the effectiveness of the model through data preprocessing, and defines the input parameters through the detection process, which is crucial to the effectiveness of the subsequent actual model training process. The project also focuses on the performance of model training and seeks a balance between accuracy and performance. The goal of the project is to obtain the emotional trend changes of users for specific events from raw social media data.