Chapter 3: Research Methodology

3.1 Introduction

In chapter-3, it explores the research methodology employed and is divided into three main parts: Research Design, Problem Formulation, and Datasets, all critical to the overall study's success. Research design outlines the strategic plan implemented to achieve the research objectives. A thorough explanation of the methods and techniques used and the reasons for choosing particular sentiment analysis algorithms and machine learning models is given. This section also addresses the procedural steps taken to collect, process, and analyse data, ensuring the research is carried out systematically and scientifically. Problem formulation is about clearly defining the research problem. This section explores the key challenges in accurately predicting stock prices using sentiment analysis techniques, particularly within the Malaysian context. It outlines the research questions, hypotheses, and objectives, preparing for a focused analysis of the relationship between news sentiment and stock price changes. Datasets provide a thorough summary of the data sources used in the study. This includes a detailed description of the primary data sources, such as reputable Malaysian online news portals like New Straits Times, Bursa Malaysia, and The Edge Market. The section also covers the data collection process, highlighting the criteria for selecting news articles and the methods used to classify sentiments at a granular level. The datasets section ensures transparency in the data handling process and emphasizes the trustworthiness of the information used for analysis.

3.2 Research Design

This study will employ a mixed-methods research design, combining both approaches of quantitative and qualitative, to investigate the impact of sentiment expressed and found in financial news headlines on stock price movements in the Malaysian market. Research framework for this proposal as below:

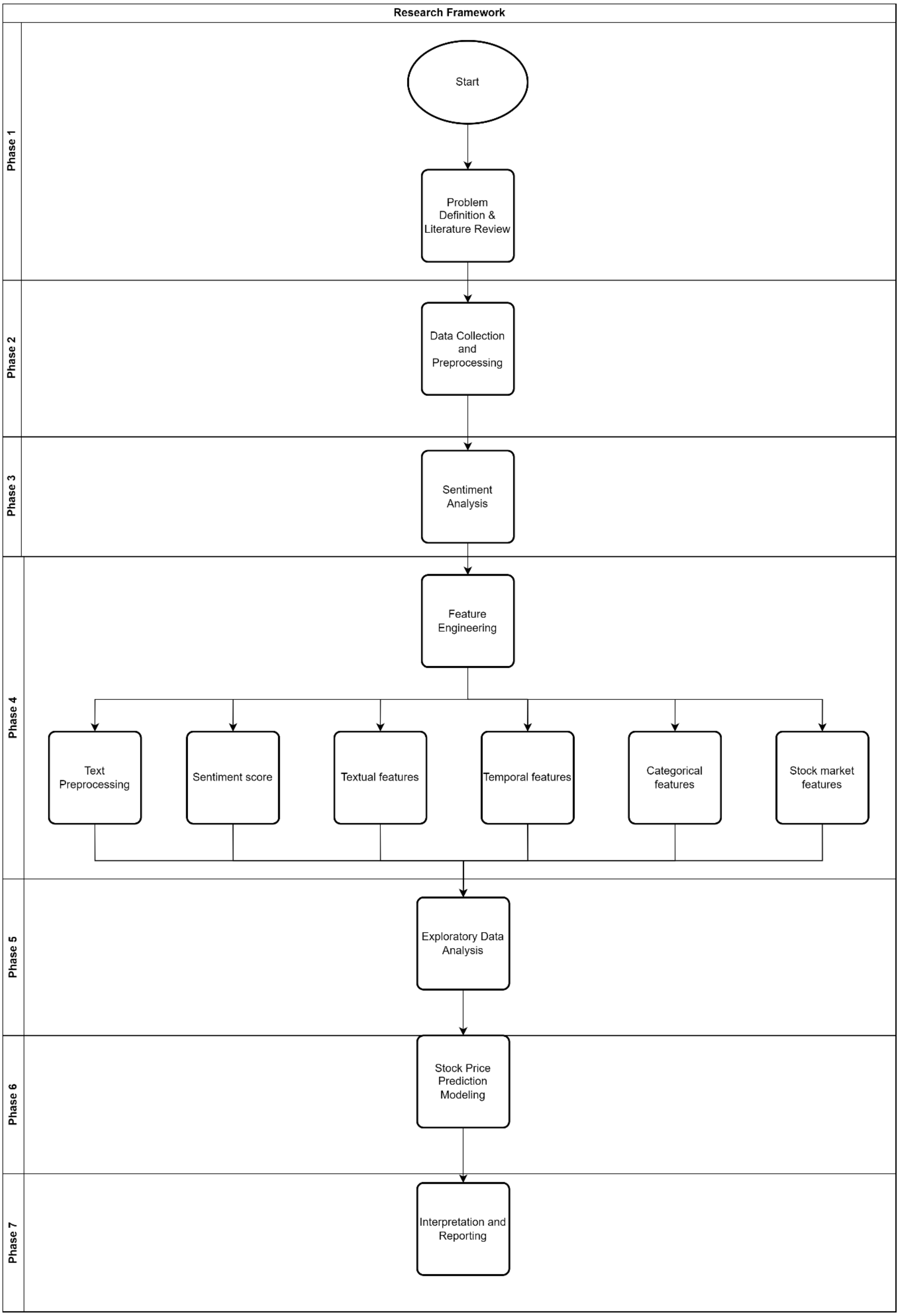


Table 3.1: Research Framework of proposal

3.2.1 Quantitative Approach

The quantitative component of the research design will involve the following key elements illustrated in Table 3.1.

Quantitative component	*Description*
Data Collection	• The main sources of data will be collected from trusted online news portals in Malaysia, such as the New Straits Times, Bursa Malaysia, and The Edge Market. • Financial news or article headlines will be collected for a period of 5 years to create a complete dataset for analysis. • Historical data on stock prices for the same period will be sourced from reputable sources such as Bursa Malaysia.
Sentiment Analysis	• Classification of sentiment will be conducted on a sentence-by-sentence basis, identifying whether each sentence in the news headlines is positive, negative, or neutral. • Traditional sentiment analysis methods like Naive Bayes and Lexicon-based approaches will be used alongside more sophisticated machine learning models like Long Short-Term Memory (LSTM) networks for sentiment analysis. • The effectiveness of these sentiment analysis strategies will be assessed and compared to identify the most effective methods for predicting stock price movements in the Malaysian market.
Stock Price Prediction Models	• The relationship between the sentiment expressed (positive, neutral, negative) in news headlines and historical stock prices will be analyzed to uncover patterns and temporal dependencies. • Traditional forecasting models, such as ARIMA, will be employed as a baseline for comparison.

	• Advanced predictive models, including LSTM and Gated Recurrent Unit (GRU) networks, will be developed and optimized to improve the precision when forecast stock price based on news sentiment. • The performance of these models will be evaluated using metrics such as root mean squared error (RMSE) and mean absolute error (MAE).

Table 3.2: Quantitative component

3.2.2 Qualitative Approach

The qualitative components of the research design are described in Table 3.2.

Qualitative component	*Description*
Interviews with Financial Experts	• Conducted semi-structured interviews with financial analysts, traders, and investment professionals to gain insights based on their perspectives on the role of sentiment analysis in stock price forecasting and decision-making. • The interviews will explore the practical challenges, limitations, and potential applications of sentiment analysis in the Malaysian stock market.
Content Analysis of News Articles	• In addition to the quantitative sentiment analysis of news headlines, a qualitative content analysis of the full text of selected news articles will be performed. • This analysis will provide a deeper understanding of the contextual factors and nuances that may impact the relationship between the movements of stock price with news sentiment.

Table 3.3: Qualitative component

3.3 Problem Formulation

Specific problems that the study aims to address are highlighted in Table 3.3.

No.	*Research Questions*	*Research Objectives*	*Proposed solutions*
1	How does specific sentiment expressed in financial news headlines impact the movement of stock prices in Malaysia?	To analyze the nuanced impact of specific sentiments expressed in financial news headlines on stock price movements within the Malaysian stock market context.	• Conduct sentiment classification at the sentence level, categorizing each sentence in the news headlines as either positive, negative, or neutral. • Employ both traditional sentiment analysis algorithms (e.g., Naive Bayes, Lexicon-based) and advanced machine learning models (e.g., LSTM networks) to capture the sentiment expressed in the news headlines at the granular level. • Investigate the relationship between the classified sentiment and the corresponding stock price movements to uncover the nuanced impact of specific sentiments.
2	What are the main challenges in order to accurately predict the stock prices in the	The aim is to recognize and assess the main obstacles in accurately forecasting stock prices in the Malaysian market	• Develop and train traditional sentiment analysis algorithms, such as Naive Bayes and Lexicon-based approaches, to

	Malaysian market using sentiment analysis techniques, and how to optimized the advanced models like LSTM networks to address these challenges?	through sentiment analysis methods, and improve advanced models such as LSTM networks to boost prediction accuracy by tackling these obstacles.	predict stock price movements based on news sentiment. • Construct advanced machine learning models, particularly LSTM networks, to forecast stock prices using the sentiment data extracted from news headlines. • Compare the performance of the traditional and advanced models using evaluation metrics, such as mean absolute error (MAE) and root mean squared error (RMSE). The aim is to identify the most effective techniques for the Malaysian market.
3	How do various sentiment analysis techniques like Hybrid Naive Bayes and Opinion Lexicon-based methods affect the prediction of stock price changes in Malaysia, and how can these methods be evaluated and enhanced for more accurate forecasts?	To analyze the effects of various sentiment analysis algorithms, like Hybrid Naive Bayes and Opinion Lexicon-based methods, on forecasting stock price changes in Malaysia, and enhancing these algorithms to enhance prediction accuracy.	• Analyze the time-series relationship between the sentiment expressed and extracted from news headlines and the corresponding related historical stock prices. • Identify patterns and temporal dependencies that influence stock price movements over time, leveraging techniques like time-series analysis and cross-correlation. • Incorporate the temporal insights into the development and optimization of the stock price prediction models,

			including LSTM and GRU networks, to enhance the accuracy of forecasts.

Table 3.4: Problem formulation

3.4 Datasets

The success of any stock price prediction model largely depends on the quality and relevance of the data used for training and evaluation. In the context of this project, the data collection process involves gathering the necessary information to support the analysis and modelling tasks. The description of these datasets is explained in Table 3.4.

Dataset	Description	Data source
Textual Data	Full text of the news articles	News websites such as Malaysiakini, The New York Times, The Washington Post, BBC, CNN provide APIs or make their article content available for download. Financial Reports and Press Releases.
Sentiment Analysis Scores	Numerical score representing the sentiment of the article content, typically derived from a machine learning model, Sentiment score calculated using a lexicon-based approach	Sentiment scores could be generated using machine learning models trained on labeled datasets, Lexicon sentiment scores might come from predefined sentiment lexicons such as AFINN, VADER (Valence Aware Dictionary and Sentiment Reasoner), or the NRC Emotion Lexicon, which assign

		sentiment values to words and phrases. Commercial and open-source sentiment analysis tools and APIs, such as those provided by Google Cloud Natural Language API, IBM Watson, or Python libraries like TextBlob and NLTK, could also be used to derive these scores
Metadata	Names of the authors of the articles, dates and times when the articles were published.	Web Scraping or news aggregators

Table 3.5: Datasets

3.4.1 News Article Data

The main data for this research will be financial news headlines gathered from trusted Malaysian online news websites such as Malaysiakini, New Straits Times, Bursa Malaysia, and The Edge Market. These news sources were selected based on their prominence and credibility in the Malaysian financial landscape. The news headlines will be collected over 5 years, from January 2018 to June 2024, to ensure a robust and representative dataset for analysis. This timeframe was chosen to capture the potential impact of various economic and market events on the relationship among news sentiment and stock price movements.

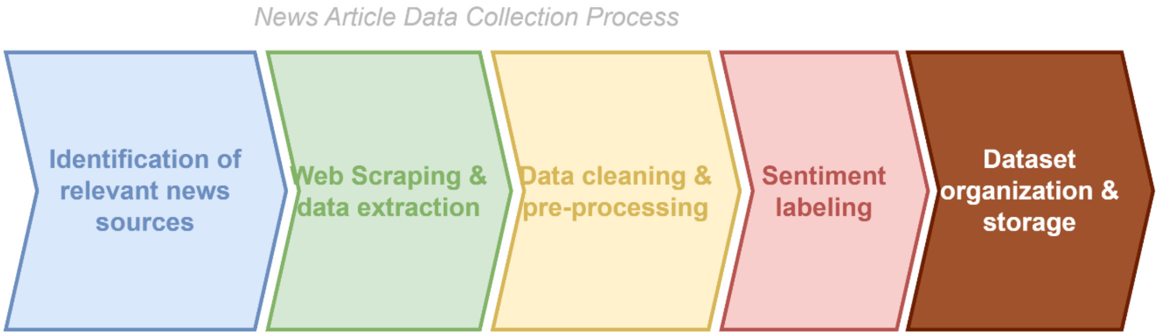


Figure 3.1: News article data collection process

Process steps	*Description*
Relevant news sources identification	The trustable Malaysian online news websites that regularly cover financial and stock market news will be identified. The sources will be chosen depending on their standing, audience, and emphasis on the Malaysian financial market.
Web scraping & data extraction	Automated web scraping techniques will be used to extract financial news headlines from chosen online news portals. This procedure will require creating scripts or using web scraping tools to gather headlines and related metadata (such as publication date and article URL) systematically over 5 years timeframe.
Data cleaning & preprocessing	The news headlines that have been gathered will be subjected to a thorough data cleaning and preprocessing step. This will involve actions like eliminating duplicate entries, addressing missing data, and standardizing the format and structure of the headlines to ensure consistency across the dataset.
Sentiment labeling	Each news headline will be manually reviewed and categorize each news headline based on its sentiment (positive, negative, or neutral). This manual labeling process will act as the foundation for the following sentiment analysis and model training.
Dataset organization & storage	The cleaned and labeled news headline dataset will be organized and stored in a structured format, such as a CSV file or a relational database, making it easier to manage and analyze the data effectively.

Table 3.6: Data collection process steps

Web scrapping step	*Description*
1. Define the URL Range	Articles were scraped from the news section of the Malaysiakini website, specifically targeting URLs within a specified range. The range covered articles from https://www.malaysiakini.com/news/405000 to https://www.malaysiakini.com/news/710000.
2. URL Construction and Looping	A loop was set up to iterate through each URL in the defined range. For each URL, the newspaper3k library was used to download and parse the article.
3. Extraction of Data	The following information was extracted from each article: - Title: The news articles' headline. - Author: Name of author of the article. - Published Date: The date on which the article was published. - Content: The main body of text of the article.
4. Filtering Criteria	Articles were filtered based on their publication dates to include only those published between 2018 and 2024. This ensures that the analysis focuses on recent and relevant articles for past 6 years.
5. Error Handling and Data Storage	Log will skip any URLs that failed to return a valid response and extracted data was stored in a panda DataFrame and subsequently saved to a CSV file with UTF-8 encoding to handle text data, including Chinese characters, accurately.

Table 3.7: Web scrapping steps

3.4.2 Stock Price Data

In addition to the financial news headlines, the study will also include historical stock price information for the Malaysian stock market. This dataset will be obtained from the yfinance which contains essential fields such as Date, Open, High, Low, Close, Volume, and Adjusted Close prices.

Attribute	Meaning
Date	the trading date.
Open	the stock's opening price on the given date.
High	the highest price of the stock on the given date.
Low	the lowest price of the stock on the given date.
Close	the closing price of the stock on the given date.
Volume	the number of shares traded on the given date.
Adjusted Close prices	the stock's closing price adjusted for corporate actions.

Table 3.8: Attribute and meaning

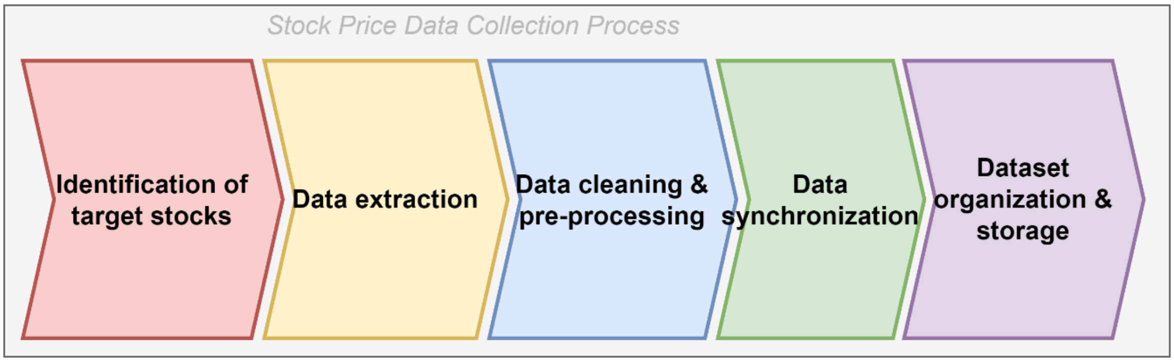


Figure 3.2: Stock price data collection process

3.4.3 Lexicon-based Approach

To aid the sentiment analysis aspect of the research, sentiment lexicons will be used. These lexicons are collections of words and their corresponding sentiment ratings. These dictionaries will act as a basis for the conventional algorithms used in sentiment analysis, such as the Lexicon-based method. The study will investigate the effectiveness of different sentiment lexicons, such as general-purpose and finance-specific lexicons, to identify the most suitable resources for the Malaysian financial context.

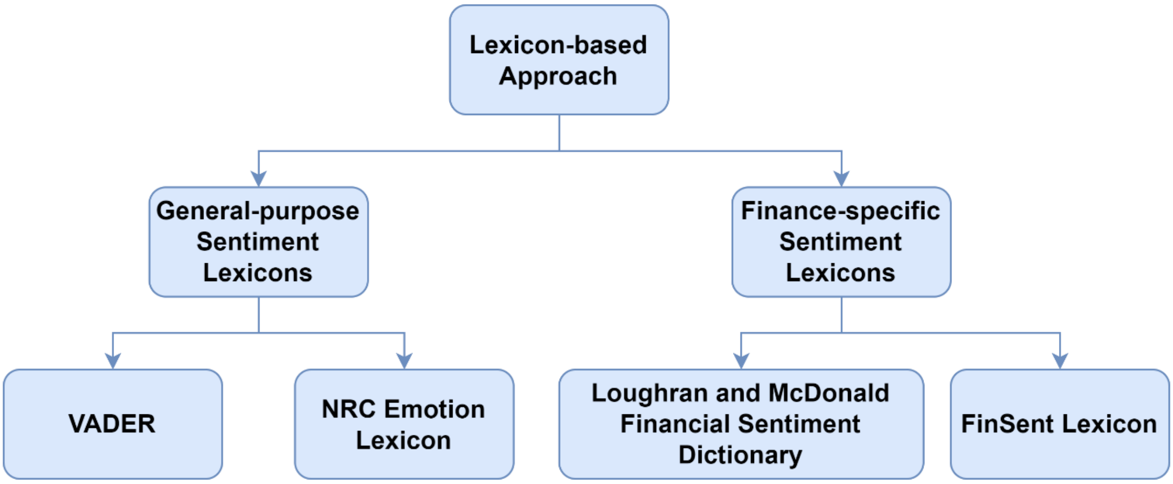


Figure 3.3: Lexicon-based approach

Lexicon-based approach		Description
General-purpose Sentiment Lexicons	VADER (Valence Aware Dictionary and Sentiment Reasoner)	VADER is a sentiment analysis tool that focuses on emotions conveyed in social media, using lexicons and rules. It offers an extensive compilation of words along with their corresponding sentiment scores.
	NRC Emotion Lexicon	The NRC Emotion Lexicon is a popular sentiment lexicon that links words with eight fundamental emotions (anger, surprise, anticipation, fear, trust, sadness, disgust, and joy) and two sentiment polarities (positive and negative).
Finance-specific Sentiment Lexicons	Loughran and McDonald Financial Sentiment Dictionary	This specialized dictionary was created for the financial domain and includes a comprehensive list of words with their corresponding sentiment ratings in the context of financial reporting and news.
	FinSent Lexicon	The FinSent Lexicon is a sentiment lexicon specifically designed for analyzing sentiment in the financial domain, using financial text data for its creation and validation.

Table 3.9: Lexicon-based approach and its description

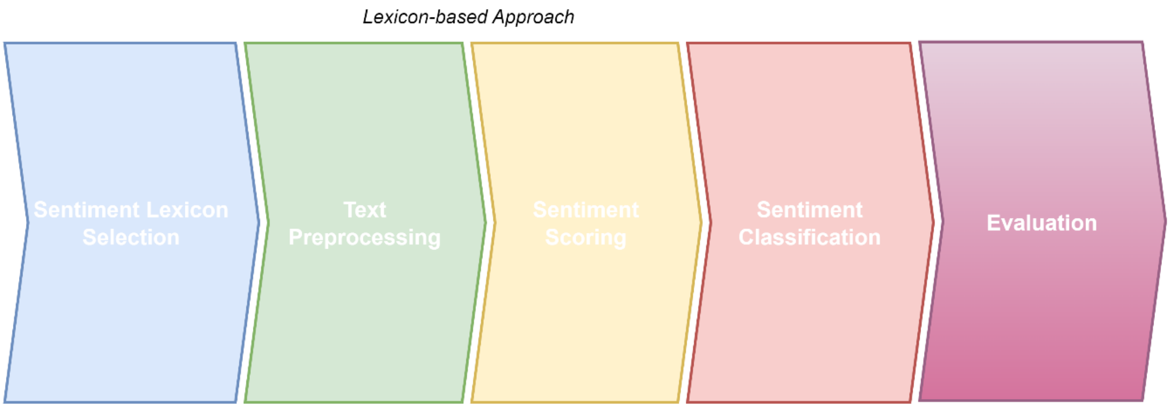


Figure 3.4: Lexicon-Based Approach steps

Lexicon-based sentiment analysis steps	Description
1.Sentiment Lexicon Selection	Several sentiment lexicons will be evaluated, such as the VADER (Valence Aware Dictionary and Sentiment Reasoner) lexicon. It is specifically designed for social media text, and the NRC Emotion Lexicon, which provides associations between words and eight basic emotions.
2.Data Pre-processing	The news article content will undergo standard text preprocessing steps, including: 1. Cleaning text– Remove unnecessary content like HTML tags, special characters, punctuations, and digits from text. 2. Standardization in lower case – Standardize text in the same lower case as the computer differentiates between lower case and upper case. 3. Tokenization – Convert sentences into words. 4. Stopword removal – Words that provide no meaningful information such as ‘this’, ‘a’, ‘there’, and ‘an’. 5. Lemmatization or stemming - to simplify words by stripping off affixes and returning them to their base form.
3.Sentiment Scoring	For each news article, the sentiment score will be calculated by the sum of sentiment scores of the individual words in the text, based on their association with positive or negative sentiment in the selected

	lexicon(s). For example, sum of sentiment of 1 is positive, while 0 is negative.
4.Sentiment Classification	The news articles will be classified into positive, neutral, or negative sentiment categories based on the calculated sentiment scores. This can be done by setting appropriate thresholds or using a rule-based approach.
5.Evaluation	The performance of the lexicon-based sentiment analysis will be evaluated using appropriate metrics, such as precision, recall, accuracy, and F1-score. This is able to provide insights into the effectiveness of the chosen lexicons and the overall reliability of the sentiment classification.

Table 3.10: Lexicon-based sentiment analysis steps (Srivastava et al., 2022)

3.4.4 Machine Learning Models

Training a model in machine learning for sentiment analysis involves categorizing text into sentiment categories like positive, neutral, or negative using a dataset with labels. This approach can be more accurate and flexible than the lexicon-based approach, as it can learn to capture more complex patterns and relationships in the data. However, it requires a larger and more labelled dataset for training and can be more computationally intensive.

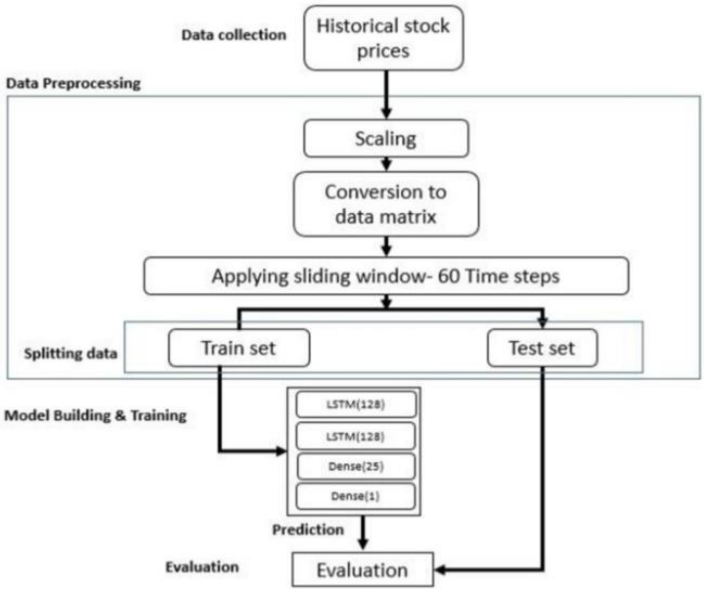
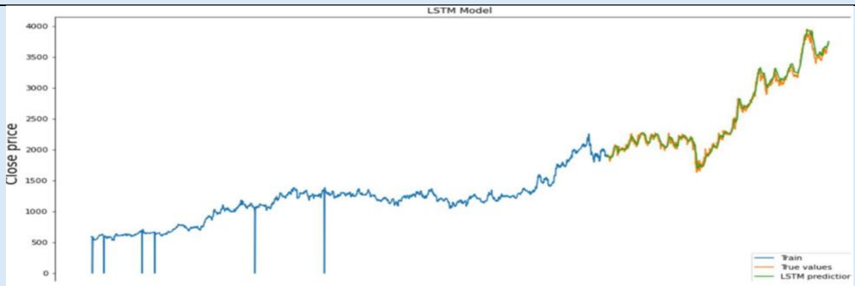


Figure 3.5: Machine learning-based sentiment analysis framework (Gangthade, 2024)

Process	Description
1.Development Phase 1	Collection of data in .csv text file.
2.Root Mean Square Error	Using formula “root mean square error (rmse) = np.sqrt(np.mean(predictions – y_test)**2))” to get value of root mean square error.
3.Plot predicted data	Plot the predicted data to examine how close is it to the actual values.
4.Example of plotted graph	 Fig: Prediction Graph

<table>
<tr>
<td>5.Example of close price and prediction</td>
<td>
<table>
<tr><th></th><th>Close</th><th>predictions</th></tr>
<tr><th>Date</th><th></th><th></th></tr>
<tr><td>2019-01-02</td><td>1923.300049</td><td>1903.230591</td></tr>
<tr><td>2019-01-03</td><td>1899.949951</td><td>1906.295898</td></tr>
<tr><td>2019-01-04</td><td>1876.849976</td><td>1904.251465</td></tr>
<tr><td>2019-01-07</td><td>1897.900024</td><td>1897.174561</td></tr>
<tr><td>2019-01-08</td><td>1893.550049</td><td>1896.034912</td></tr>
<tr><td>...</td><td>...</td><td>...</td></tr>
<tr><td>2021-12-27</td><td>3696.100098</td><td>3691.024414</td></tr>
<tr><td>2021-12-28</td><td>3706.550049</td><td>3709.432617</td></tr>
<tr><td>2021-12-29</td><td>3694.699951</td><td>3726.835449</td></tr>
<tr><td>2021-12-30</td><td>3733.750000</td><td>3737.803467</td></tr>
<tr><td>2021-12-31</td><td>3738.350098</td><td>3754.460938</td></tr>
</table>
741 rows × 2 columns
</td>
</tr>
</table>

Figure 3.6: Process steps, graph, and outputs of machine learning-based sentiment analysis (Gangthade, 2024)

3.4.5 Sentiment Classification Refinement (Hybrid Approach)

A hybrid approach, or combination of lexicon-based and machine learning-based approaches to leverage the strengths of both methods. In this approach, the lexicon-based approach is used to provide an initial sentiment score, which is then refined and adjusted by using a machine learning model trained on labelled data. This can result in more accurate and robust sentiment analysis, particularly for more complex or ambiguous text.

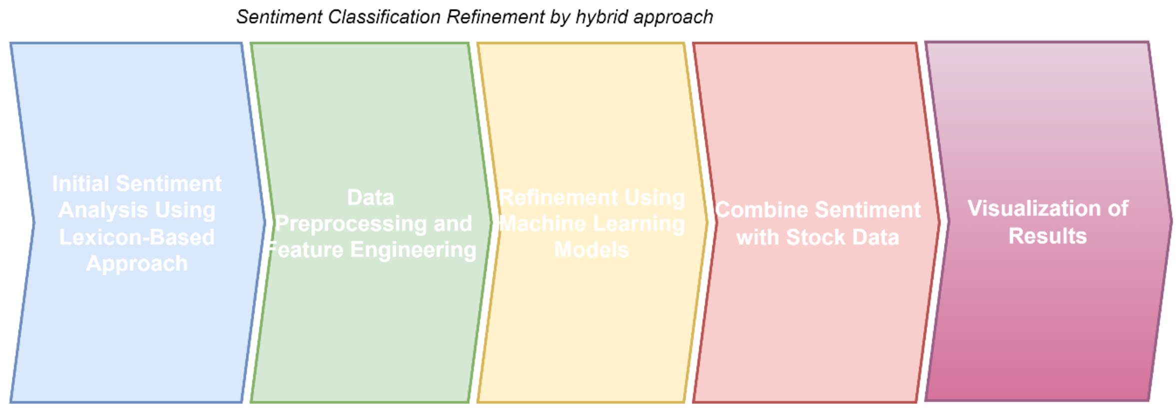


Figure 3.7: Sentiment classification refinement by hybrid approach

3.4.6 Deep Learning Techniques

Recent advancements in deep learning have led to the development of more advanced sentiment analysis techniques. Among them, include the use of neural networks, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), use to capture the semantic and contextual information in the text. Deep learning models can provide higher accuracy than traditional machine learning approaches, especially tasks that require a deep understanding of language and sentiment.

When selecting a sentiment analysis technique for stock price prediction, factors such as the availability and quality of labelled data, the complexity of the sentiment expressions in the text, and the computational resources available need to be consider. A combination of different techniques, or a hybrid approach, could provide the best results.