FACULTY OF COMPUTING

UNIVERSITI TEKNOLOGI MALAYSIA

MASTER OF SCIENCE (DATA SCIENCE)

SEMESTER 1 2023/2024

RESEARCH DESIGN AND ANALYSIS IN DATA SCIENCE (MCSD1043-01)

Proposal:

Sentiment Analysis of News Title on the Movement of Stock Prices in Malaysia

LECTURER:

DR. MOHD SHAHIZAN BIN OTHMAN

PREPARED BY:

NAME	STUDENT ID
YAP QI YUAN	MCS231025

SUBMIT BY:

24 July 2024

Table of Contents

Chapter 1: Introduction

1.1 Overview

In chapter-1, it provides a detailed summary of the research and highlights the key elements that make up the research framework. It begins with the research background, setting the scene and highlighting the importance of the topic in ongoing discussions, current challenges in the field, and existing strategies to resolve them. The problem statement defines the particular issue or lack that will be focused on in the research, setting the groundwork for the study. Next, the study's direction and focus are determined by outlining the research questions. This is then followed by research objectives that address various aspects of the issue in order to support the main aim of the research and its influencing factors. The research scope outlines the study's limits and specifies what will be included and excluded in the research. Lastly, the research significance underscored the study's importance and potential influence by detailing the benefits and demonstrating how it will enhance knowledge in the wider field.

1.2 Research Background

Sentiment analysis, also known as opinion mining, is the process of identifying individuals' opinions, emotions, attitudes, and feelings expressed in written form, specifically in financial news (Mishra, 2023; Shuhidan et al., 2018). This analysis is essential for predicting market trends in financial environments. Stakeholders can utilize sentiment analysis of news headlines to obtain valuable insights to inform their decision-making processes. Different strategies, like lexicon-based approaches and algorithms such as Naive Bayes, are commonly used in sentiment analysis(Cheng Kuan et al., 2019).

Research has shown that the mood portrayed in news articles can greatly influence the value of stocks, with technologies like long short-term memory (LSTM) networks presenting an opportunity to improve the precision of stock price predictions (Sidek et al., 2023). In Malaysia, the assessment of feelings in news headlines has been studied using machine learning techniques such as Hybrid Naive Bayes, Opinion Lexicon-based algorithm, and Naive Bayes (Cheng Kuan et al., 2019; Shuhidan et al., 2018).

The emotions conveyed in financial news appear to have a strong link with the movement of the stock market. This suggests that sentiment analysis could be an effective approach for predicting market patterns (McCarthy & Alaghband, 2023). Even with these

Commented [TP1]: Requires citations (reference)

improvements, difficulties remain in improving the precision of sentiment analysis models. In this research, our goal is to explore the sentiment analysis of news headlines and how it influences stock price changes in the Malaysian market. By filling the gaps in current literature and investigating new methods, we aim to contribute to the progress of sentiment analysis in financial environments.

1.3 Statement of the Problem

It is crucial to grasp the sentiment in financial news headlines in order to predict stock market movements accurately, as these headlines hold vital information that can greatly impact stock prices. Different methods of sentiment analysis, including the Opinion Lexicon-based algorithm and Naive Bayes, have been utilized to evaluate sentiment in financial news. Nevertheless, obstacles remain in accurately predicting stock prices, particularly in relation to public attitudes and the distinct features of Malaysian online news platforms. While past studies have indicated a link between news sentiment and stock market fluctuations, the impact of sentiment on stock prices could vary. Advanced predictive models, such as Long Short-Term Memory (LSTM) networks, guarantee the improvement of stock price predictions by incorporating sentiment data. These models provide an advanced method for examining emotions and how they influence market patterns.

Future studies can concentrate on improving these sophisticated predictive models to increase their precision in forecasting stock prices. Furthermore, investigating how these models can be applied to various financial markets can offer valuable insights for enhancing the overall precision of predicting stock prices. Researchers can improve predictive models, resolve challenges, and improve tools to forecast the stock market trends using sentiment analysis in financial news. Financial news headlines have a significant impact on stock market movements, but the exact effect of these sentiments on stock prices in Malaysia is not fully comprehended. Effectively predicting stock prices is challenging due to the need to combine historical data analysis with public sentiment analysis, making it difficult to improve accuracy through sentiment analysis integration. Existing studies provide a visualization regarding how the news sentiment influences the stock prices, and emphasize the necessity of advanced predictive models like Long Short-Term Memory (LSTM) networks to elevate prediction precision.

Although there is a known correlation between financial news sentiment and stock market movements, there is still limited research focused on Malaysia. Therefore, there is a need for further investigations into how news sentiment affects stock prices in the Malaysian financial market. Despite encouraging findings, persistent challenges exist in fine-tuning sentiment analysis models for optimal stock price prediction accuracy. The difficulty and complexity to accurate predict the stock prices in the Malaysian market involves a blend of historical data analysis and the integration of public sentiment. This poses the challenges to effectively incorporating the sentiment analysis to enhance prediction accuracy. This research aims to explore the impact of sentiment expressed in financial news headlines on stock price movements in the Malaysian market, evaluating traditional sentiment analysis techniques and advanced predictive models like Long Short-Term Memory (LSTM) networks to improve the precision of stock price forecasts within the Malaysian financial landscape.

1.4 Research Questions

1. How does specific sentiment expressed in financial news headlines impact the movement of stock prices in Malaysia?
2. What are the main challenges in order to accurately predict the stock prices in the Malaysian market using sentiment analysis techniques, and how to optimized the advanced models like LSTM networks to address these challenges?
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?

1.5 Objectives of the Research

1. To analyze the nuanced impact of specific sentiments expressed in financial news headlines on stock price movements within the Malaysian stock market context.
2. The aim is to recognize and assess the main obstacles in accurately forecasting stock prices in the Malaysian market through sentiment analysis methods, and improve advanced models such as LSTM networks to boost prediction accuracy by tackling these obstacles.

3. 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.

1.6   Scope of the Study

The scopes of this research are:

1. To investigate the impact of sentiment analysis from financial news headlines on stock price movements in the Malaysian stock market.
2. To examine the impact of feelings (positive, negative, and neutral) on stock prices in Malaysia at the sentence level, utilizing trusted Malaysian online news portals like the New Straits Times, Bursa Malaysia, and The Edge Market as main sources of data.
3. Use traditional sentiment analysis algorithms and advanced machine learning models, including Long Short-Term Memory networks to forecast the stock price movements based on news sentiment.
4. To conduct comprehensive data collection and analysis over 5 years to provide valuable insights for traders, investors, and financial analysts to optimize their investment strategies in Malaysia.

1.7   Significance of the Research

This research is highly important for traders and investors as it offers valuable information on financial news sentiment, helping in making informed decisions and minimizing the risk of making choices based on incomplete information. By incorporating sentiment analysis into forecasting models for stock prices, especially using advanced methods such as Long Short-Term Memory networks, the accuracy of predictions can be greatly enhanced. The research explores the correlation between historical stock prices and sentiment data and offers essential insights for refining predictive models and developing more accurate investment strategies in the stock market.

The comparison between traditional and advanced machine learning algorithms in sentiment analysis helps to determine the most efficient methods for forecasting stock price fluctuations using news sentiment. This research able to provide more knowledges and

understanding of the Malaysian financial market by investigating how news sentiment impacts stock prices in Malaysia. Financial stakeholders like traders, investors, and financial analysts , they can use sentiment analysis to improve market strategies and potentially boost the returns.

Furthermore, the research supports the development of automated trading systems by integrating sentiment analysis, so it results in more reactive and advanced trading algorithms. This research sets a standard for future research, providing a foundation for more investigation into sentiment analysis and predicting stock prices, revealing the possibilities and difficulties of using sentiment data in financial markets.

Chapter 2: Literature Review

2.1   Introduction

In chapter-2, it examines the present status of sentiment analysis research and its implementation in financial markets, with a specific emphasis on its usage in the Malaysian stock market. Also, it provides a summary of the main advancements, obstacles, and possibilities in this field. Section 2.2 discusses the importance of sentiment analysis in financial markets, highlighting the use of sophisticated models like FinBERT and eXplainable Lexicons (XLex) to improve effectiveness and understandability. Section 2.3 explores the relationship between sentiment analysis and stock price changes, showing how investor sentiment influences stock performance. Section 2.4 looks into the sentiment analysis's application in the Malaysian stock market, particularly amidst the COVID-19 pandemic. Section 2.5 presents predictive models for predicting stock prices, utilizing methods such as LSTM, GRU, and mixed frameworks. In conclusion, Section 2.6 highlights areas lacking or gaps in the literature and suggests research possibilities like analyzing news article headlines, experimenting with effective sentiment analysis techniques, and enhancing the categorization of neutral sentiment.

2.2   Sentiment Analysis in Financial Markets

Sentiment Analysis in financial markets can be useful for understanding and predicting market movements by analyzing textual data from the financial articles, news, and social media. Various methodologies have been investigated and developed to enhance the accuracy and efficiency of sentiment analysis in this domain. One prominent approach involves using FinBERT, a transformer model specifically tailored for financial sentiment analysis, which has been shown to be effective in predicting market movements when applied to stock news datasets. This model outperforms traditional methods like BERT, LSTM, and ARIMA, highlighting the importance of sentiment as a predictive factor (Jiang & Zeng, 2023). However, transformer models, despite their superior performance, require computational resources and extensive data, making them less suitable for real-time applications or systems with limited processing capabilities (Rizinski et al., 2024). To address these limitations, the eXplainable Lexicons (XLex) methodology combines the advantages of lexicon-based methods and transformer models, utilizing SHapley Additive exPlanations (SHAP) for enhanced

9

explainability. This approach not only improves the vocabulary coverage of financial lexicons but also offers better efficiency and interpretability compared to traditional transformer models, making it a viable tool for financial decision-making (Rizinski et al., 2024).

2.3 Sentiment Analysis and Stock Price Movements

Sentiment analysis can be emerged as a pivotal tool to predicts the stock price movements, leveraging the vast amount of textual data that available from social media, news, and reports of financial. Studies have shown that investor sentiment, particularly from platforms like Stocktwits, can significantly influence stock prices, although challenges such as the accurate classification of neutral comments persist. To address these, advanced models like FinBERT have been employed, demonstrating superior performance in sentiment analysis and improving prediction accuracy when combined with ensemble support vector machines (Liu et al., 2023). Additionally, integrating sentiment analysis with other data sources, such as historical stock prices and commodity prices, has proven effective. For instance, a classification model using Naive Bayes achieved a 60% accuracy in predicting stock movements over a three-day period by incorporating copper prices and sentiment features (Sinatrya et al., 2022).

Aslim et al. (2023) emphasize the use of Long Short-Term Memory (LSTM) models combined with lexicon-based sentiment analysis to predict stock prices, highlighting enhanced accuracy when integrating sentiment data. This approach aids investors in making informed decisions. Similarly, Praturi et al. (2023) discuss an application that leverages sentiment analysis on financial news through deep learning and cloud computing, demonstrating its efficacy in accurate stock price forecasting. Shah et al. (2019), who review the impact of Twitter sentiments on stock market movements, suggesting that social media plays a pivotal role in influencing stock prices. These research findings show that analyzing sentiments, whether from financial news or social media, is a useful method to forecast stock price changes and assist in investment strategies. Collectively, their findings underscore the critical role of sentiment analysis in stock price prediction, particularly when combined with other analytical techniques and robust models.

2.4 Sentiment Analysis in the Malaysian Stock Market

Sentiment analysis is significant in understanding and predicting stock market behavior in Malaysia, especially during the periods of economic uncertainty like COVID-19 pandemic. Studies have shown that public sentiments from online news portals, can significantly influence the stock price movements of Bursa Malaysia. For instance, a study utilizing long short-term memory (LSTM) models demonstrated that incorporating news polarity values improved the prediction accuracy of stock prices, with the root mean squared error (RMSE) values being less than one for every company stock analyzed (Sidek et al., 2023). During the COVID-19 pandemic, the sentiment expressed in the discussion of management and analysis sections of annual reports from Malaysian public listed companies revealed predominantly negative tones, reflecting concerns about the pandemic's impact on business operations. This sentiment was confirmed through both automated textual analysis and qualitative content analysis (Non & Ab Aziz, 2023). The pandemic had a significant impact on how investors felt or sentiment and the performance of the stock market. The steep downward trend in the FTSE BURSA 100 Index (T100) during the early months of the outbreak was evidenced it. The rapid increase in COVID-19 cases and deaths heightened uncertainty and it disrupted the investment decisions, as captured by the Sentiment Index (SMI) constructed using principal component analysis (PCA) (Albada & Nizar, 2022).

2.5 Advanced Predictive Models for Stock Price Forecasting

Sophisticated machine learning and deep learning techniques are used in advanced predictive models to address the unpredictable nature and complex non-linear features of stock markets. Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) models are powerful predictive models because of their talent to capture long-term dependencies and complex relationships in time series data (Momaya et al., 2023). For example, research conducted on Indian blue-chip stocks demonstrated the efficacy of LSTM and GRU models in forecasting stock prices with high accuracy by training on historical data spanning 17 years (Momaya et al., 2023). Furthermore, the Multivariate Sequential LSTM Autoencoder (MSLSTMA) model, which combines LSTM with an autoencoder component, has been found to excel in capturing dependencies between multiple variables and improving prediction accuracy through unsupervised learning (G et al., 2023). This model outperformed other techniques such as Univariate Sequential LSTM, GRU, Random Forest, and Generative

Adversarial Networks (GAN) in experiments with real stock market data (G et al., 2023). Another innovative approach involves a hybrid framework combining stacked LSTM with a Convolutional network, which has shown competitive performance in predicting stock prices, as evidenced by experiments on ADANI stock data over 14 years (Singh & Malhotra, 2023). By using deep learning architectures and hybrid frameworks on advanced models, provide robust tools for investors aiming to maximize returns through accurate stock price predictions.

2.6   Gaps in the Literature and Research Opportunities

Previous studies on the relationship between sentiment analysis and stock price movements have provided valuable insights, particularly in the Malaysian context. However, there are various gaps in the existing literature that offer possibilities for further exploration and advancement.

First, limited focus on the sentiment conveyed in news article titles, rather than the full text of the articles. While studies have analyzed the sentiment in various textual sources, such as social media posts and financial reports, the unique influence of news titles, which are often the primary information source for investors, has not been extensively examined. Investigating the relationship between the sentiment expressed in news titles and stock price changes could yield important insights into investor decision-making processes (Jiang & Zeng, 2023; Jiang et al., 2023).

Another gap in the literature is the computational and data-intensive nature of transformer models like FinBERT, which are commonly used for sentiment analysis. These models need computational resources and extensive data, limiting their suitability for real-time applications or systems that pose limited processing capabilities. Exploring more efficient and interpretable sentiment analysis methods, such as the eXplainable Lexicons (XLex) approach, could enhance financial decision-making and make sentiment analysis more accessible for a wider range of applications (Rizinski et al., 2023, 2024).

Additionally, the existing research has faced challenges in accurately classifying neutral comments in sentiment analysis of social media platforms like Stocktwits. Developing advanced models and techniques to improve the classification of neutral sentiment and it could further enhance the predictive accuracy of stock price movements (Liu et al., 2022, 2023). These gaps in the literature, as illustrated in Table 2.1, offer the opportunities for future research

to contribute to a deeper understanding of the role of sentiment analysis in the stock market of Malaysia and provide useful insights for investors, policymakers, and financial analysts.

Gap	Description
Limited focus on sentiment in news article titles	Studies have analyzed sentiment in various textual sources, but the unique influence of news titles (a primary information source for investors) has not been extensively examined. Investigating the relationship between news title sentiment and stock price changes could provide insights into investor's decision-making.
Computational and data-intensive nature of transformer models	Transformer models like FinBERT, commonly used for sentiment analysis, which need extensive data and computational resources. This limits their suitability for real-time applications or systems that pose limited processing capabilities. Exploring more efficient and interpretable sentiment analysis methods, such as eXplainable Lexicons (XLex), could enhance financial decision-making and make sentiment analysis more accessible.
Challenges in accurately classifying neutral comments	Existing research has faced difficulties in accurately classifying neutral comments in sentiment analysis of social media platforms like Stocktwits. Developing advanced models and techniques to improve neutral sentiment classification could enhance the predictive accuracy of stock price movements.
Research Opportunities	These gaps in the literature present opportunities for future research to contribute to a more depth understanding of the role of sentiment analysis in the Malaysian stock market. This could provide valuable insights for investors, policymakers, and financial analysts.

Table 2.1: Research gap analysis

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:
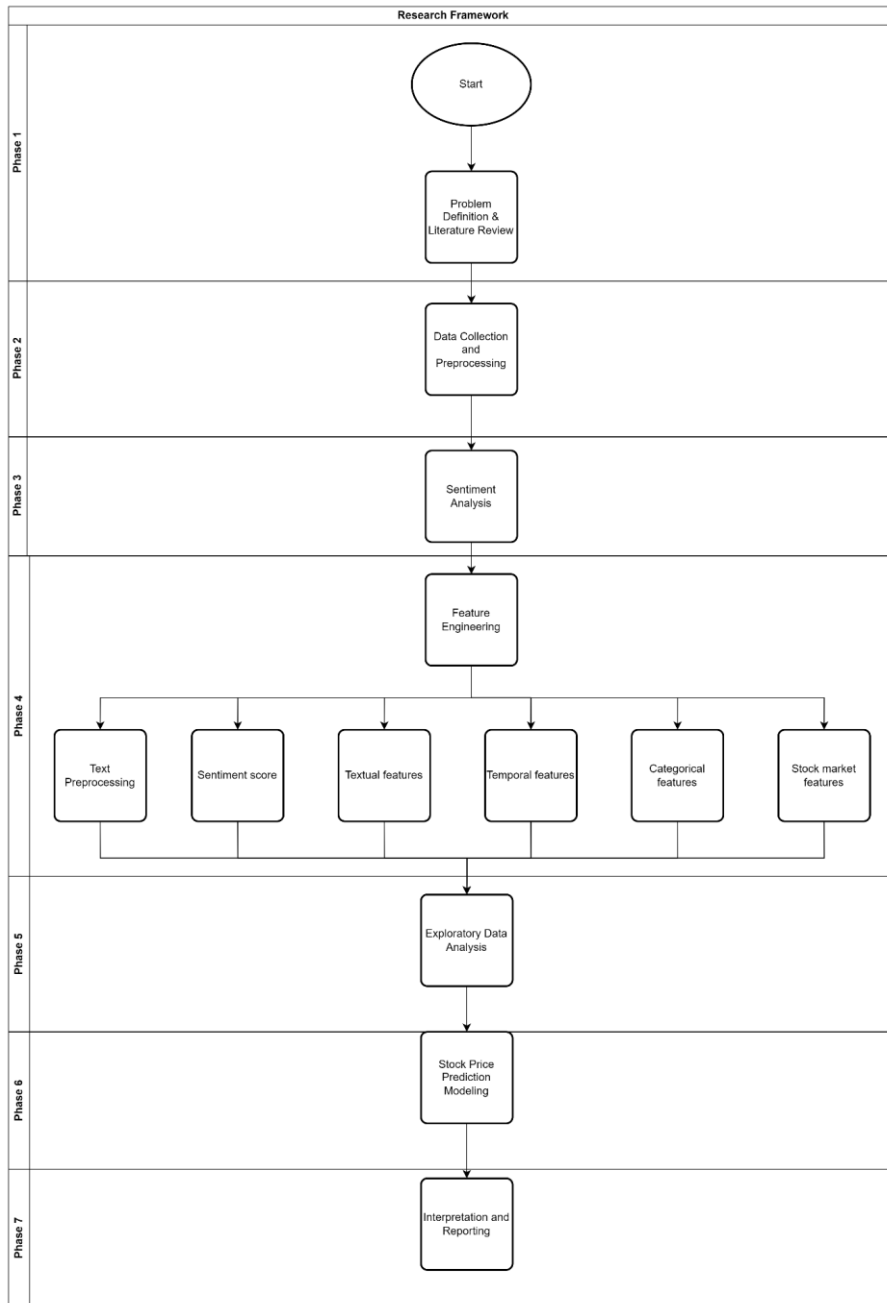
Research Framework

Phase 1 — Start → Problem Definition & Literature Review

Phase 2 — Data Collection and Preprocessing

Phase 3 — Sentiment Analysis

Phase 4 — Feature Engineering → Text Preprocessing, Sentiment score, Textual features, Temporal features, Categorical features, Stock market features

Phase 5 — Exploratory Data Analysis

Phase 6 — Stock Price Prediction Modeling

Phase 7 — Interpretation and Reporting

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	The aim is to recognize and assess the main obstacles in accurately	• Develop and train traditional sentiment analysis algorithms, such as Naive Bayes and

	predict the stock prices in the Malaysian market using sentiment analysis techniques, and how to optimized the advanced models like LSTM networks to address these challenges?	forecasting stock prices in the Malaysian market through sentiment analysis methods, and improve advanced models such as LSTM networks to boost prediction accuracy by tackling these obstacles.	Lexicon-based approaches, to 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.
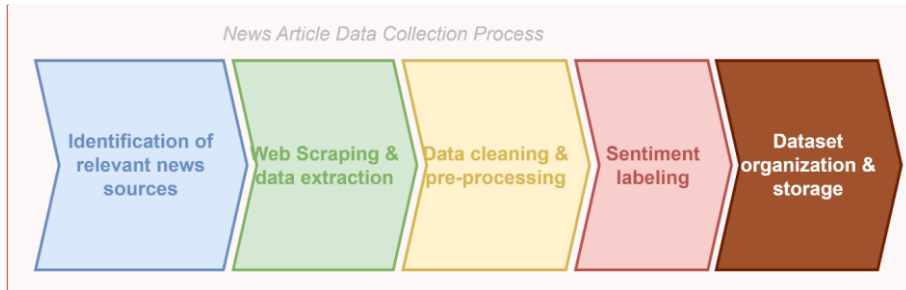
Dataset	Description	Data source
Textual Data	Full text of the news articles	News websites such as Malaysiakini, The New York Times, The Washington Post, BBC, CNN provide APIs or make their article content available for download. Financial Reports and Press Releases.
Sentiment Analysis Scores	Numerical 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.

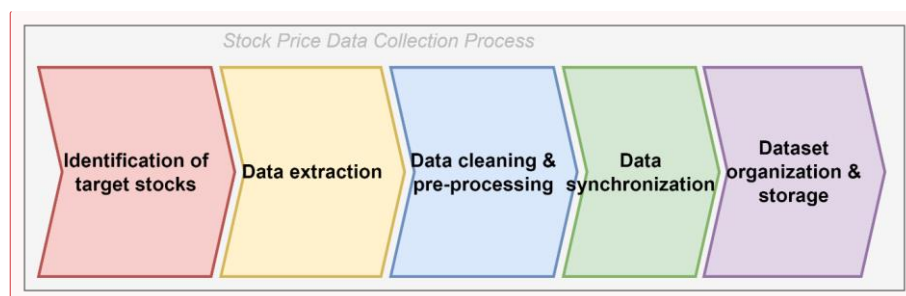Table 3.8: Attribute and meaning

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.

Commented [TP3]: You picked this from somewhere, you have to state the reference
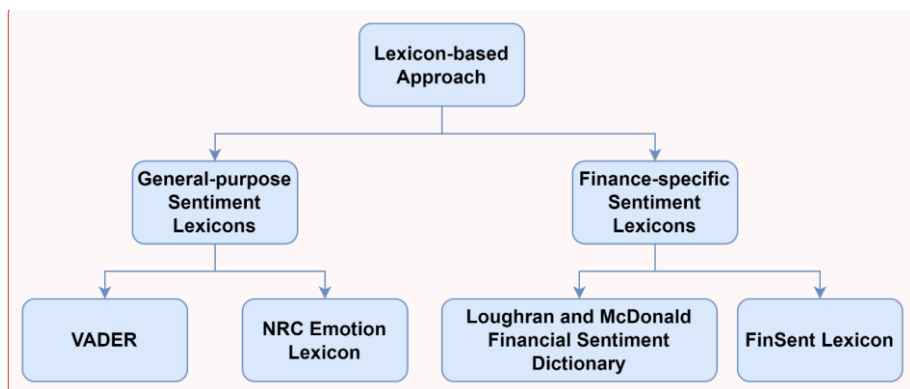
Figure 3.3: Lexicon-based approach

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.

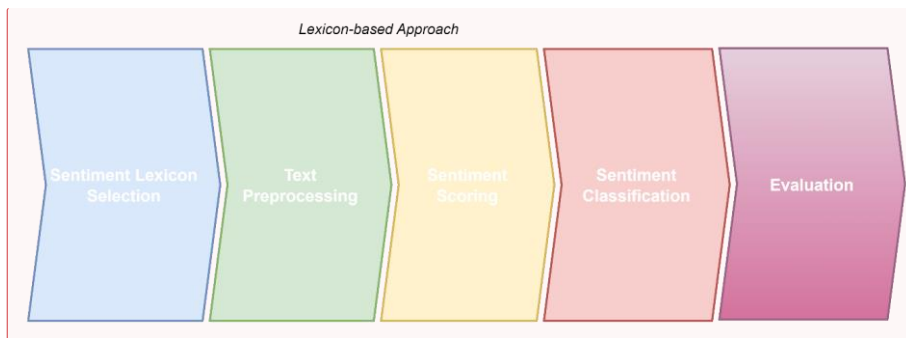Table 3.9: Lexicon-based approach and its description

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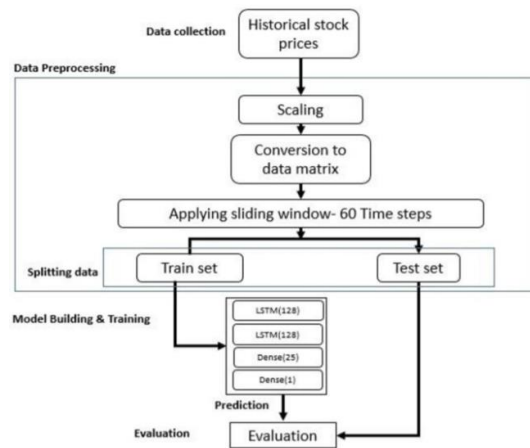
Figure 3.5: Machine learning-based sentiment analysis framework (Gangthade, 2024)

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

5.Example of close price and prediction	

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.
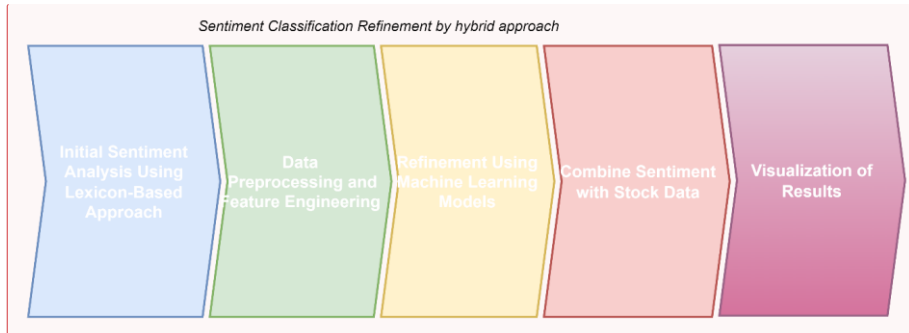
Figure 3.7: Sentiment classification refinement by hybrid approach

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.

Chapter 4: Exploratory Data Analysis/Initial Results

4.1 Introduction

In chapter-4, the exploratory data analysis (EDA) and initial results will be the main focus of the study. It is divided into three main parts: 4.2 Visualisations and descriptive statistics for data exploration, 4.3 Feature engineering, 4.4 Expected results. Section 4.2.1 using different visualization methods, the study will uncover the patterns, trends, and potential correlations between news sentiment and stock price changes in the Malaysian market. For section 4.2.2, by analyzing metrics like mean, median, standard deviation, and frequency counts, the research will provide a comprehensive understanding of the sentiment distribution and characteristics within the dataset. Section 4.3 using techniques like text preprocessing, categorical feature encoding, and the incorporation of relevant stock market features to transform the raw data into a format that suitable for the subsequent predictive modeling. In section 4.4, preliminary insights generated from the exploratory data analysis that affects stock price changes, guiding the direction and focus for more advanced modeling and evaluation.

4.2 Visualisations and descriptive statistics for data exploration

Visualizations and descriptive statistics play a crucial role in exploring and understanding the data (Qin et al., 2020). The duo will be employed in this research for effective presentation and analysis of results.

4.2.1 Visualisation tools

For this study, the first step will examine the distribution of sentiment scores using histograms. These visualizations will provide insights into the central tendency, spread, and skewness of the sentiment scores derived from both machine-learning models and lexicon-based approaches. Histograms will be used to visualize the frequency of different sentiment scores, which will allow how sentiments are distributed across the articles to be seen. Box plots will be used in this exploration to highlight the median, quartiles, and potential outliers, thus offering a clear view of the variability and central tendencies of the sentiment scores. Such visual tools, according to Ortis, et al. (2021) are essential in understanding whether the sentiment distribution is balanced or biased towards positive, negative, or neutral sentiments.

Time series analysis is another powerful approach this research will explored. Line plots of sentiment scores over time can reveal trends, patterns, and anomalies in the data, which can then be correlated with stock price movements. By resampling the data on a monthly or weekly basis, the study can observe how average sentiment scores evolve and identify periods of significant sentiment shifts. Additionally, scatter plots and heatmaps will be used for correlation analysis. Scatter plots can visually demonstrate the relationship between sentiment scores and stock prices, potentially uncovering trends or patterns that indicate a correlation between media sentiment and market behaviour. Heatmaps of correlation matrices can concisely present the direction of relationships between multiple variables (sentiment scores and various stock market indicators) and their strength. These visualizations, combined with rolling mean and standard deviation plots, will help in identifying trends and volatility in sentiment over time, providing a dynamic view of how sentiment interacts with stock market movements.

4.2.2 Descriptive statistics

Descriptive statistics complement visualization tools by providing a numerical summary of the data (Narechania, et al., 2020). Metrics such as mean, median, standard deviation, and variance offer a quantitative understanding of the central tendency and dispersion of sentiment scores which will be used in this research. Frequency counts of sentiment categories (positive, negative, neutral) add another layer of analysis that will help to quantify the overall sentiment landscape of the news articles. By combining these descriptive statistics with visual tools, the research will comprehensively explore the sentiment data, uncovering patterns and insights that inform the relationship between media sentiment and stock price movements. This integrated approach of visualization and descriptive statistics will not only enhance the understanding of the dataset but also provides a solid foundation for further predictive modelling and analysis in the context of financial markets.

4.3   Feature engineering

Feature engineering is the process that extracts and creates new variables (features) from raw data that can help improve the performance of machine learning models by using domain knowledge (Verdonck et al., 2021). It involves selecting, modifying, and transforming

data to enhance the predictive power of models. Effective feature engineering, according to Fan et al. (2019), can lead to better model accuracy and insights, as it allows the model to focus on the most relevant aspects of the data. In the context of this research, feature engineering is crucial to transform raw text data from news articles into meaningful features that can be used to predict stock price movements. In this research, feature engineering will be applied in the following:

- o **Text Preprocessing**: This involves **tokenization** which has to do with breaking down the article content into separate words or tokens; **stop words removal** (remove the common words that do not have important meaning), and **stemming and lemmatization** (reducing words to their root forms to ensure consistency e.g. "running" to "run").
- o **Sentiment Scores**: This involves **overall sentiment score** (using NLP models to assign a sentiment score to each article, indicating whether the sentiment is positive, neutral, or negative, and **lexicon-based sentiment score** (calculating sentiment based on predefined lexicons to provide an alternative measure of sentiment).
- o **Textual Features**: This involves **the term frequency-inverse document frequency (TF-IDF),** i. e. Measuring the importance of words in an article relative to a corpus of articles, helping to highlight unique and significant terms, and **N-grams (e**xtracting contiguous sequences of n words e.g., bigrams, trigrams, to capture context and common phrases).
- o **Temporal Features**: This has to do with **publication date and time** (incorporating the date and time when the article was published to analyse temporal patterns and their impact on stock prices) and **lag features** (i.e. creating lagged versions of sentiment scores to capture delayed effects of news sentiment on stock prices, e.g., sentiment score from the previous day or week).
- o **Categorical Features**: Such features include **source and author** (encoding the source and author of the article to account for potential biases or credibility variations).
- o **Stock Market Features**: This includes **previous day's stock price** (including the stock price from the previous trading day to account for existing market trends) and trading **volume** (incorporating trading volume to understand market activity and sentiment impact).
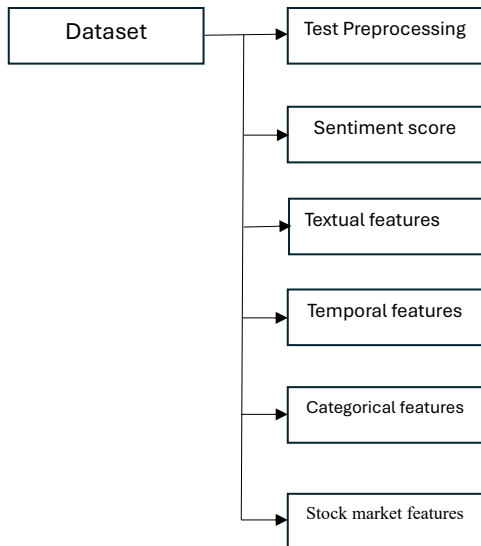
Figure 4.1: Feature engineering applications on the datasets

4.4 Expected results

S/No	Research objectives	Expected results
1	To analyze the nuanced impact of specific sentiments expressed in financial news headlines on stock price movements within the Malaysian stock market context.	• Detailed Sentiment Categories: Identification and classification of specific sentiments (optimism, pessimism, fear, confidence) expressed in financial news headlines. • Impact on Stock Price Movements: Quantitative analysis showing how these specific sentiments impact stock price movements. For example, headlines that expressed strong confidence might correlate with an increase in stock prices,

S/No	Research objectives	Expected results
		while negative headlines might lead to a decrease. • Sector-Specific Impact: Insights to check how different sectors (technology, finance, consumer goods) are affected by various sentiment types, separately. Certain sectors may react more strongly to specific sentiments than others. • Sentiment Intensity: Analysis to check how the intensity of sentiment (strongly positive/negative vs. mildly positive/negative) affects the magnitude of stock price movements.
2	The aim is to recognize and assess the main obstacles in accurately forecasting stock prices in the Malaysian market through sentiment analysis methods, and improve advanced models such as LSTM networks to boost prediction accuracy by tackling these obstacles.	• Performance Metrics: Compare the performance metrics (precision, accuracy, F1-score, recall, etc.) for traditional sentiment analysis techniques and advanced machine learning models in predicting stock price movements. • Model Effectiveness: Findings that indicate which models (e.g., Naive Bayes, Lexicon-based, LSTM networks) are more effective in capturing the sentiment from news headlines and predicting stock price movements.

S/No	Research objectives	Expected results
		- Feature Importance: Insights into which features (specific words, phrases, or sentiment scores) are most predictive of stock price movements for each model. - Time Sensitivity: Evaluation of how different models can handle the temporal aspect of news sentiment and stock price prediction. Possibly showing that LSTM networks can account for sequential data and outperform the traditional methods.
3	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.	- Temporal Patterns: Identification of temporal patterns showing the correlation of sentiment data from news headlines with historical stock price movements. For example, a lagged effect where sentiment influences stock prices after a certain time delay. - Causality and Correlation: Analysis of causality and correlation between sentiment changes and stock price movements over different time frames (either immediate, short-term, or long-term). - Volatility Analysis: Insights to check sentiment-driven news impacts stock price volatility over

S/No	Research objectives	Expected results
		time. Periods of high sentiment activity might correspond with higher volatility.
		• Event-Based Analysis: Case studies of significant events (e.g., earnings announcements, political changes) and their temporal impact on stock prices, revealing specific patterns during these events.
		• Historical Trends: Uncovering long-term trends where certain types of sentiment consistently lead or lag behind stock price movements, providing a historical perspective on sentiment's influence on the market.

Table 4.1: Research objectives and expected results

4.4.1 Machine Learning (Initial Result)

The process and steps to get initial result of machine learning by using Scikit-learn (sklearn) at Colab as shown at below:

```
!pip install openpyxl
!pip install nltk scikit-learn pandas
```

```python
import pandas as pd
import os

# Initialize a list to store DataFrames
dfs = []

# Loop through files and read each one
for i in range(1, 30):
    file_path = f'/content/filtered_sentiment_analyzed_{i}.csv'
    if os.path.exists(file_path):
        df = pd.read_csv(file_path)
        dfs.append(df)

# Concatenate all DataFrames
merged_df = pd.concat(dfs, ignore_index=True)

# Save to an Excel file
merged_excel_path = '/content/merged_sentiment_analyzed.xlsx'
merged_df.to_excel(merged_excel_path, index=False)

print(f"Merged DataFrame saved to: {merged_excel_path}")
```

```python
# Load the merged sentiment data and stock price data
sentiment_df = pd.read_excel(merged_excel_path)
stock_path = '/content/financial_services.csv'
stock_df = pd.read_csv(stock_path)

print("Sentiment DataFrame:")
print(sentiment_df.head())

print("Stock Prices DataFrame:")
print(stock_df.head())
```

```python
import string
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from sklearn.feature_extraction.text import CountVectorizer
import nltk

nltk.download('stopwords')
nltk.download('punkt')

# Preprocessing function
def preprocess_text(text):
    text = text.lower()
    text = text.translate(str.maketrans("", "", string.punctuation))
    tokens = word_tokenize(text)
    tokens = [word for word in tokens if word not in stopwords.words('english')]
    return " ".join(tokens)

# Apply preprocessing to the title column
sentiment_df['Processed Title'] = sentiment_df['Title'].apply(preprocess_text)

# Convert 'Published Date' to datetime format for merging
# Force both date columns to have the same format without timezone
sentiment_df['Published Date'] = pd.to_datetime(sentiment_df['Published Date']).dt.tz_localize(None)
stock_df['Date'] = pd.to_datetime(stock_df['Date'])

# Merge the news sentiment scores with stock prices based on the date
merged_data = pd.merge(sentiment_df[['Published Date', 'Sentiment Score']],
                       stock_df,
                       left_on='Published Date',
                       right_on='Date',
                       how='inner')

print("Merged DataFrame:")
print(merged_data.head())
```

```python
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report

# Vectorize the text data
vectorizer = CountVectorizer(max_features=1000)
X = vectorizer.fit_transform(sentiment_df['Processed Title'])

# Define target variable for binary classification
y = sentiment_df['Sentiment Score'].apply(lambda x: 1 if x > 0 else 0)  # Example binary classificatio

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Train model
model = LogisticRegression()
model.fit(X_train, y_train)

# Predict and evaluate
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
report = classification_report(y_test, y_pred, output_dict=True)

print("Accuracy:", accuracy)
print("Classification Report:")
print(classification_report(y_test, y_pred))
```

```
Accuracy: 0.7046435516767063
Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.48      0.57      8622
           1       0.70      0.86      0.77     12461

    accuracy                           0.70     21083
   macro avg       0.70      0.67      0.67     21083
weighted avg       0.70      0.70      0.69     21083
```

```python
# Convert classification report to DataFrame and save to Excel file
report_df = pd.DataFrame(report).transpose()
report_excel_path = '/content/classification_report.xlsx'
report_df.to_excel(report_excel_path, index=True)

print(f"Classification report saved to: {report_excel_path}")

Classification report saved to: /content/classification_report.xlsx
```

Figure 4.2: Initial result of machine learning by Scikit-learn

```python
!pip install newspaper3k
import pandas as pd
from newspaper import Article
import logging

def fetch_articles(start, end):
    # Create an empty list to store the article data
    data = []

    # Loop through the article range and extract the data
    for i in range(start, end + 1):
        url = f'https://www.malaysiakini.com/news/{i}'
        try:
            article = Article(url)
            article.download()
            article.parse()

            # Extract the relevant data
            title = article.title
            author = article.authors
            published_date = article.publish_date
            content = article.text

            # Filter articles between 2018 and 2024
            if published_date and 2018 <= published_date.year <= 2024:
                data.append({
                    'Title': title,
                    'Author': author,
                    'Published Date': published_date,
                    'Content': content
                })
        except Exception as e:
            # Log the error and skip the invalid URL
            logging.error(f"Error fetching article from URL: {url} - {e}")
            continue

    return data

# Set the range for article URLs
start_id = 710000
end_id = 720000

# Fetch articles within the date range and URL range
articles = fetch_articles(start_id, end_id)

# Create a DataFrame from the extracted data
df = pd.DataFrame(articles)

# Save the data to a CSV file with UTF-8 encoding
df.to_csv('malaysiakini_news_filtered_40to450000.csv', index=False, encoding='utf-8')

# Display the DataFrame
print(df)
```

Figure 4.3: Coding for Data Collection of News title

41

```
!pip install vaderSentiment

import pandas as pd
import string
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
import nltk

# To ensure have the necessary NLTK corpora
nltk.download('stopwords', quiet=True)
nltk.download('punkt', quiet=True)
```

```python
# Preprocessing function
def preprocess_text(text):
    if not isinstance(text, str):
        text = str(text)
    text = text.lower()
    text = text.translate(str.maketrans("", "", string.punctuation))
    tokens = word_tokenize(text)
    tokens = [word for word in tokens if word not in stopwords.words('english')]
    return " ".join(tokens)

# Initialize VADER sentiment analyzer
analyzer = SentimentIntensityAnalyzer()

# Function to get sentiment score using VADER
def get_sentiment_score(text):
    sentiment_dict = analyzer.polarity_scores(text)
    return sentiment_dict['compound']

# Loop through files from cleaned_1.csv to cleaned_29.csv
for i in range(1, 30):
    filename = f"cleaned_{i}.csv"
    print(f"\nProcessing file: {filename}")

    # Read file
    try:
        news_df = pd.read_csv(filename)
        print(f"Original shape: {news_df.shape}")

        # Apply preprocessing to the content column
        news_df['Processed Content'] = news_df['Content'].apply(preprocess_text)

        # Apply sentiment analysis to the processed content
        news_df['Sentiment Score'] = news_df['Processed Content'].apply(get_sentiment_score)

        # Save the results
        output_filename = f"sentiment_analyzed_{i}.csv"
        news_df.to_csv(output_filename, index=False)
        print(f"Sentiment analysis completed. Results saved as: {output_filename}")
        print(f"Final shape: {news_df.shape}")
        print(news_df[['Title', 'Sentiment Score']].head())

    except FileNotFoundError:
        print(f"File {filename} not found. Skipping...")
    except Exception as e:
        print(f"An error occurred while processing {filename}: {str(e)}")

print("\nSentiment analysis completed for all files.")
```

```
Processing file: cleaned_1.csv
Original shape: (1671, 4)
Sentiment analysis completed. Results saved as: sentiment_analyzed_1.csv
Final shape: (1671, 6)
                                        Title  Sentiment Score
0              Happy New Year from Malaysiakini         0.9451
1              500人赴双下集会，跨年喊"油价纳吉都要下"                     0.0000
2            Seorang diplomat di Kota Darul Naim         0.8555
3  Yoursay: Non-Muslims pay taxes, but can't be i...    0.9960
4  Najib praises Hadi's 'better way'; 500 rally o...    0.5267

Processing file: cleaned_2.csv
Original shape: (3039, 4)
Sentiment analysis completed. Results saved as: sentiment_analyzed_2.csv
Final shape: (3039, 6)
                                        Title  Sentiment Score
0  Say goodbye to elections should Najib win, Mah...   -0.1082
1        Maria's candidacy breathes hope for GE14         0.9985
2              "雪槟即大马未来"，阿兹敏冠英化身CEO卖政绩                 0.0000
3    Trump ready to meet N Korea's Kim Jong-un by May   0.9678
4              马哈迪促澳洲总理，趁东盟峰会向纳吉提一马案                   0.000

Processing file: cleaned_3.csv
Original shape: (2575, 4)
Sentiment analysis completed. Results saved as: sentiment_analyzed_3.csv
Final shape: (2575, 6)
                                        Title  Sentiment Score
0  Nurul Izzah: Nation may witness 'dirtiest poll...   0.0258
1              TV3记者"倒戈"坦承，报道黑函前没向安华求证                  0.0000
2  PM's 'chaos if gov't changed' remark undemocra...   0.9806
3  SAMM gives MACC documents on S'gor sand mining...   0.2023
4                        到底为什么要投废票?                     0.0000
```

Figure 4.4: Coding to get sentiment score

References

Albada, A., & Nizar, N. (2022). The Impact of the Investor Sentiment Index (SMI) on the Malaysian Stock Market during the COVID-l9 Pandamic. *International Journal of Economics and Management*, *16*(2), 225–236. https://doi.org/10.47836/ijeam.16.2.06

Aslim, M. F., Firmansyah, G., Tjahjono, B., Akbar, H., & Widodo, A. M. (2023). Utilization of LSTM (Long Short Term Memory) Based Sentiment Analysis for Stock Price Prediction. Asian Journal of Social and Humanities, 1(12), 1241-1255.

Chau, N. T., Kien, L. V., & Phong, D. T. (2023). Stock price movement prediction using text mining and sentiment analysis. *Studies in Computational Intelligence*, 167–179. https://doi.org/10.1007/978-3-031-29447-1_15

Cheng Kuan, M., Zayet, L., Akmar Ismail, T. M. A., & Mohamed Shuhidan, S. (2019). Prediction of Malaysian stock market movement using sentiment analysis. *Journal of Physics: Conference Series*, *1339*(1), 012017. https://doi.org/10.1088/1742-6596/1339/1/012017

Gangthade, R. A. (2024b). Stock price prediction using machine learning. *International Journal for Research in Applied Science and Engineering Technology*, *12*(4), 3472–3477. https://doi.org/10.22214/ijraset.2024.60725

Fan, C., Zhao, Y., Song, M., & Wang, J.Sun, Y. (2019). Deep learning-based feature engineering methods for improved building energy prediction. *Applied energy*, *240*, 35-45.

Jiang, T., & Zeng, A. (2023). *Financial sentiment analysis using FinBERT with application in predicting stock movement* (arXiv:2306.02136). arXiv. http://arxiv.org/abs/2306.02136

Liu, J.-X., Leu, J.-S., & Holst, S. (2023). Stock price movement prediction based on Stocktwits investor sentiment using FinBERT and ensemble SVM. *PeerJ Computer Science*, *9*, e1403. https://doi.org/10.7717/peerj-cs.1403

McCarthy, S., & Alaghband, G. (2023). Enhancing Financial Market Analysis and Prediction with Emotion Corpora and News Co-Occurrence Network. *Journal of Risk and Financial Management*, *16*(4), 226. https://doi.org/10.3390/jrfm16040226

Mishra, J. (2023). TWITTER SENTIMENT ANALYSIS. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, *07*(06). https://doi.org/10.55041/IJSREM24071

Momaya, H., Patel, V., & Momaya, V. (2023). Forecasting of Stock Trend and Price using Machine Intelligence LSTM and GRU Models. *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, 1–6. https://doi.org/10.1109/ViTECoN58111.2023.10157684

Narechania, A., & Stasko, J., A., Srinivasan, (2020). NL4DV: A toolkit for generating analytic specifications for data visualization from natural language queries. *IEEE Transactions on Visualization and Computer Graphics*, *27*(2), 369-379.

Non, N., & Ab Aziz, N. (2023). An exploratory study that uses textual analysis to examine the financial reporting sentiments during the COVID-19 pandemic. *Journal of Financial Reporting and Accounting*, *21*(4), 895–915. https://doi.org/10.1108/JFRA-10-2022-0364

Ortis, A., Torrisi, G. M., G., & Battiato, Farinella, S. (2021). Exploiting objective text description of images for visual sentiment analysis. *Multimedia Tools and Applications*, *80*(15), 22323-22346.

Praturi, Ramakrishnan, A., S. S. G., & Deepthi, L. R. (2023, July). Stock Price Prediction Using Sentiment Analysis on Financial News. In International Conference on Data Science and Applications (pp. 551-567). Singapore: Springer Nature Singapore.

Qin, X., Tang, N., Luo, Y., & Li, G. (2020). Making data visualization more efficient and effective: a survey. *The VLDB Journal*, *29*(1), 93-117.

Rajanikanth, J., Haritha, K., & Shankar, R. S. (2023). Forecasting stock close price using machine learning models. *ARPN Journal of Engineering and Applied Sciences*, 412–420. https://doi.org/10.59018/022361

Rizinski, M., K., Jovanovik, H., Mishev, Peshov, M., & Trajanov, D. (2024). *Sentiment Analysis in Finance: From Transformers Back to eXplainable Lexicons (XLex)* (arXiv:2306.03997). arXiv. http://arxiv.org/abs/2306.03997

Shah, D., H., & Zulkernine, Isah, F. (2019). Stock market analysis: A review and taxonomy of prediction techniques. International Journal of Financial Studies, 7(2), 26.

Shuhidan, S. R., Kazemian, S. M., Hamidi, S., Shuhidan, S. M., & Ismail, M. A. (2018). Sentiment Analysis for Financial News Headlines using Machine Learning Algorithm. In A. M. Lokman, T. Yamanaka, P. Lévy, K. Chen, & S. Koyama (Eds.), *Proceedings of the 7th International Conference on Kansei Engineering and Emotion Research 2018* (Vol. 739, pp. 64–72). Springer Singapore. https://doi.org/10.1007/978-981-10-8612-0_8

Sidek, Z., Ahmad, S. S. S., & Teo, N. H. I. (2023). Associating deep learning and the news headlines sentiment for Bursa stock price prediction. *Indonesian Journal of Electrical Engineering and Computer Science*, *31*(2), 1041. https://doi.org/10.11591/ijeecs.v31.i2.pp1041-1049

Sinatrya, N. S., I., & Budi Santoso, Budi, A. (2022). Classification of Stock Price Movement With Sentiment Analysis and Commodity Price: Case Study of Metals and Mining Sector. *2022 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 59–64. https://doi.org/10.1109/ICACSIS56558.2022.9923452

Srivastava, R., Bharti, P. K., & Verma, P. (2022). Comparative Analysis of Lexicon and Machine Learning Approach for Sentiment Analysis. *International Journal of*

Advanced Computer Science and Applications, *13*(3). https://doi.org/10.14569/IJACSA.2022.0130312

Verdonck, T., Óskarsdóttir, B., Baesens, M., & vanden Broucke, S. (2021). Special issue on feature engineering editorial. *Machine learning*, 1-12.