

**USING SENTIMENT ANALYSIS TO FORECAST STOCK SHORT-TERM TREND**

**BY**

**TAN LIN**

**A REPORT**

**SUBMITTED TO**

**Universiti Tunku Abdul Rahman**

**in partial fulfillment of the requirements**

**for the degree of**

**BACHELOR OF COMPUTER SCIENCE (HONOURS)**

**Faculty of Information and Communication Technology**

**(Kampar Campus)**

**JUNE 2024**

## REPORT STATUS DECLARATION FORM

**Title:** Using Sentiment Analysis to Forecast Stock Short-term Trend

\_\_\_\_\_

**Academic Session:** June 2024

I TAN LIN  
**(CAPITAL LETTER)**


declare that I allow this Final Year Project Report to be kept in  
Universiti Tunku Abdul Rahman Library subject to the regulations as follows:

1. The dissertation is a property of the Library.
2. The Library is allowed to make copies of this dissertation for academic purposes.

Verified by,

Tan Lin

(Author's signature)

  
\_\_\_\_\_

(Supervisor's signature)

**Address:**

Jalan Universiti,

Bandar Barat,

31900 Kampar, Perak

Kh'ng Xin Yi

Supervisor's name

**Date:** 13 September 2024

**Date:** 13/9/2024

|                                                                            |                   |                                     |                         |
|----------------------------------------------------------------------------|-------------------|-------------------------------------|-------------------------|
| <b>Universiti Tunku Abdul Rahman</b>                                       |                   |                                     |                         |
| Form Title : <b>Sample of Submission Sheet for FYP/Dissertation/Thesis</b> |                   |                                     |                         |
| Form Number: <b>FM-IAD-004</b>                                             | Rev No.: <b>0</b> | Effective Date: <b>21 JUNE 2011</b> | Page No.: <b>1 of 1</b> |

**FACULTY/INSTITUTE\* OF INFORMATION TECHNOLOGY AND COMMUNICATION**

**UNIVERSITI TUNKU ABDUL RAHMAN**

Date: September 13, 2024

**SUBMISSION OF FINAL YEAR PROJECT /DISSERTATION/THESIS**

It is hereby certified that Tan Lin (ID No: 2200188) has completed this final year project/ dissertation/ thesis\* entitled “ Using Sentiment Analysis to Forecast Stock Short-term Trend ” under the supervision of Dr. Kh'ng Xin Yi (Supervisor) from the Department of Computer Science , Faculty/Institute\* of Information Technology and Communication .

I understand that University will upload softcopy of my final year project / dissertation/ thesis\* in pdf format into UTAR Institutional Repository, which may be made accessible to UTAR community and public.

Yours truly,

Tan Lin  
(*Tan Lin*)

\*Delete whichever not applicable

## DECLARATION OF ORIGINALITY

I declare that this report entitled “**USING SENTIMENT ANALYSIS TO FORECAST STOCK SHORT-TERM TREND**” is my own work except as cited in the references. The report has not been accepted for any degree and is not being submitted concurrently in candidature for any degree or other award.

Signature :     Tan Lin    

Name :     Tan Lin    

Date :     13 September 2024

## ACKNOWLEDGEMENTS

I'm deeply grateful to my family and friends for supporting me whenever I breathe, whether on financially or mentally. Without them, I wouldn't have been able to accomplish anything.

Thanks to Dr. Kh'ng Xin Yi for her guidance, kindness, and patience. Although she initially wasn't my supervisor, she's the one who made me actually realize that responsibility in supervision is very helpful and important, a realization I couldn't come to before meeting her. It's very hard to get proper guidance in this education, but she made it. Also, she proved that it is indeed possible to have normal and inclusive relationship with supervisor in this educational environment. I couldn't imagine what I would have done if I was assigned to another supervisor just as bad. She was the best luck I had in this unfortunate experience. A million thanks to her.

Thanks to Dr. Lim Jia Qi for giving advices on technical improvement. As the moderator of this project, he made valuable suggestions that enlightened me to explore deeper and made me aware of potential mistakes.

Thanks to Dr. Saw Seow Hui for lending me a hand and providing comfort when I was suffering for this project. Although she was not responsible for this project, she still be kind to made an effort in helping me out. Her kindness inspired me to keep going.

Thanks to landscape and scenery, sky and blue sea, flowers and trees, for nurturing wonderful and lives. Thanks to pioneers for contributing innovation and knowledge, enabling later generations to learn and succeed. Thanks to every single person who extends kindness to strangers as they add immeasurable beauty to the world. Thanks to this world, accepting my existence, and this project.

Last but not least. Thanks to all the ones I love for being my anchor. Because of this strength they instilled in me, I didn't give up on my life, even when I was forced to face malices because of trusting the wrong supervisor. They are my motivation to complete this study.

## **ABSTRACT**

This research investigates the effectiveness of sentiment analysis in stock market prediction, integrating advanced computational techniques with financial analytics. Specifically, the study examines the efficacy of the Autoregressive Distributed Lag (ARDL) model combined with the GPT-4 Turbo model from OpenAI for sentiment analysis to predict the stock price movements influenced by various news sources in Malaysia. The project employs a systematic methodology to preprocess data, integrate sentiment scores, and apply the ARDL model to analyze the impact of news sentiment on stock prices. The sentiment analysis, powered by GPT-4 Turbo, provides a robust framework for interpreting the emotional tone within financial news content. Results indicate that the ARDL model, while capturing general market trends and oscillations, exhibits moderate success in forecasting, as evidenced by varying RMSE values across different news sources. This variability highlights the influential capacity of news sources and underscores the necessity for nuanced analysis techniques. The findings contribute to the broader understanding of how different news sources impact stock market movements and demonstrate the potential for enhanced predictive accuracy through the integration of advanced AI-driven tools in financial forecasting. The study's insights encourage further exploration into hybrid models that might combine traditional financial indicators with innovative sentiment analysis methodologies to improve the reliability and effectiveness of stock market predictions.

# TABLE OF CONTENTS

|                                                               |             |
|---------------------------------------------------------------|-------------|
| <b>TITLE PAGE</b>                                             | <b>i</b>    |
| <b>REPORT STATUS DECLARATION FORM</b>                         | <b>ii</b>   |
| <b>FYP THESIS SUBMISSION FORM</b>                             | <b>iii</b>  |
| <b>DECLARATION OF ORIGINALITY</b>                             | <b>iv</b>   |
| <b>ACKNOWLEDGEMENTS</b>                                       | <b>v</b>    |
| <b>ABSTRACT</b>                                               | <b>vi</b>   |
| <b>TABLE OF CONTENTS</b>                                      | <b>vii</b>  |
| <b>LIST OF FIGURES</b>                                        | <b>x</b>    |
| <b>LIST OF TABLES</b>                                         | <b>xii</b>  |
| <b>LIST OF SYMBOLS</b>                                        | <b>xiii</b> |
| <b>LIST OF ABBREVIATIONS</b>                                  | <b>xiv</b>  |
| <br>                                                          |             |
| <b>CHAPTER 1 INTRODUCTION</b>                                 | <b>1</b>    |
| 1.1 Background                                                | 1           |
| 1.2 Problem Statement                                         | 3           |
| 1.3 Research Objectives                                       | 4           |
| 1.4 Research Contributions                                    | 5           |
| 1.5 Report Organization                                       | 5           |
| <br>                                                          |             |
| <b>CHAPTER 2 LITERATURE REVIEW</b>                            | <b>6</b>    |
| 2.1 Previous works on Stock Prediction and Sentiment Analysis | 6           |
| 2.1.1 Stock Trend Prediction by News                          | 6           |
| 2.1.2 Sentiment Analysis using GPT Models                     | 9           |
| 2.2 Limitations of Previous Studies                           | 12          |
| 2.3 Proposed Solutions                                        | 13          |
| <br>                                                          |             |
| <b>CHAPTER 3 METHODOLOGY</b>                                  | <b>14</b>   |
| 3.1 Research Design                                           | 14          |
| 3.2 Data Collection                                           | 15          |

|                                        |                                 |           |
|----------------------------------------|---------------------------------|-----------|
| 3.2.1                                  | Google Custom Search JSON API   | 16        |
| 3.2.2                                  | Data Extraction                 | 18        |
| 3.2.3                                  | Elasticsearch                   | 21        |
| 3.3                                    | Data Preprocessing              | 23        |
| 3.3.1                                  | Stock Data                      | 23        |
| 3.3.2                                  | News Data                       | 24        |
| 3.4                                    | Data Integration                | 32        |
| 3.5                                    | Data Analysis                   | 33        |
| 3.5.1                                  | Augmented Dickey-Fuller         | 33        |
| 3.5.2                                  | Lags                            | 35        |
| 3.6                                    | Prediction                      | 36        |
| 3.6.1                                  | ARDL                            | 36        |
| 3.6.2                                  | Source Influence                | 38        |
| 3.7                                    | System Requirement              | 38        |
| 3.7.1                                  | Hardware Component              | 38        |
| 3.7.2                                  | Software Component              | 39        |
| 3.8                                    | Timeline                        | 40        |
| <b>CHAPTER 4 SYSTEM IMPLEMENTATION</b> |                                 | <b>41</b> |
| 4.1                                    | Yahoo Finance                   | 41        |
| 4.2                                    | Web Scraping                    | 41        |
| 4.3                                    | Elasticsearch                   | 44        |
| 4.4                                    | Generative AI                   | 44        |
| 4.5                                    | Predictive Model                | 45        |
| <b>CHAPTER 5 RESULT AND DISCUSSION</b> |                                 | <b>47</b> |
| 5.1                                    | Stock Data                      | 47        |
| 5.2                                    | News Data                       | 50        |
| 5.3                                    | Model Selection Test            | 52        |
| 5.3.1                                  | Step 1: Format Testing          | 52        |
| 5.3.2                                  | Step 2: Sentiment Trend Testing | 53        |
| 5.3.3                                  | Step 3: Stability Testing       | 55        |
| 5.4                                    | Sentiment Data                  | 55        |
| 5.5                                    | Time series analysis            | 56        |
| 5.5.1                                  | ADF Test                        | 57        |



|                                |                   |           |
|--------------------------------|-------------------|-----------|
| 5.5.2                          | Lags              | 58        |
| 5.6                            | Prediction        | 62        |
| 5.7                            | Source Comparison | 65        |
| <b>CHAPTER 6 CONCLUSION</b>    |                   | <b>67</b> |
| 6.1                            | Conclusion        | 67        |
| 6.2                            | Recommendations   | 68        |
| <b>REFERENCES</b>              |                   | <b>70</b> |
| <b>WEEKLY LOG</b>              |                   | <b>74</b> |
| <b>POSTER</b>                  |                   | <b>77</b> |
| <b>PLAGIARISM CHECK RESULT</b> |                   | <b>78</b> |
| <b>FYP2 CHECKLIST</b>          |                   | <b>80</b> |

## LIST OF FIGURES

| <b>Figure Number</b> | <b>Title</b>                                               | <b>Page</b> |
|----------------------|------------------------------------------------------------|-------------|
| Figure 3.1.1         | Research Model                                             | 15          |
| Figure 3.1.1.2       | Flow of Automated News Data Collection                     | 17          |
| Figure 3.1.1.7       | Elasticsearch Schema                                       | 22          |
| Figure 3.1.2.1       | The Transformer Model Architecture                         | 28          |
| Figure 3.1.2.3       | Testing Phase Outcomes and Actions                         | 30          |
| Figure 3.8.1         | Timeline of FYP1                                           | 40          |
| Figure 3.8.2         | Timeline of FYP1                                           | 40          |
| Figure 4.1.1         | Python Script of Stock Data Retrieval                      | 41          |
| Figure 4.2.1         | Python Script of Extract URLs                              | 42          |
| Figure 4.2.2         | Python Script of JSON Data Extraction                      | 42          |
| Figure 4.2.3         | Python Script of HTML Meta Extraction                      | 43          |
| Figure 4.2.4         | Python Script of HTML Content Extraction                   | 44          |
| Figure 4.3.1         | Python Script of Inserting Document to Elasticsearch       | 44          |
| Figure 4.4.1         | Python Script of Tasking Model                             | 45          |
| Figure 4.5.1         | Python Code Snippet of ARDL Implementation                 | 45          |
| Figure 4.5.2         | Python Code Snippet of Seasonal Differencing               | 46          |
| Figure 5.1.1         | Sample Five Rows of Stock Data                             | 47          |
| Figure 5.1.2         | Candlestick Chart of Public Bank Stock Prices 2022 to 2023 | 48          |
| Figure 5.1.3         | Candlestick Chart of CIMB Stock Prices 2022 to 2023        | 49          |
| Figure 5.2.1         | Sample Five Rows of News Data                              | 52          |
| Figure 5.3.2.1       | Comparison of Different Models Sentiment Trends            | 54          |
| Figure 5.4.1         | Number of Articles and Mean Sentiment from Each Source     | 56          |
| Figure 5.5.1         | Sample Five Rows of Aligned Data                           | 57          |
| Figure 5.5.2.1       | Autocorrelation Function for Stock Data                    | 59          |
| Figure 5.5.2.2       | Partial Autocorrelation Function for Stock Data            | 60          |
| Figure 5.5.2.3       | Autocorrelation Function for Sentiment Data                | 61          |
| Figure 5.5.2.4       | Partial Autocorrelation Function for Sentiment Data        | 61          |

|              |                                                                      |    |
|--------------|----------------------------------------------------------------------|----|
| Figure 5.6.1 | ARDL Model Forecasting Results for a Sentiment Lag of 2              | 63 |
| Figure 5.6.2 | Comparison of Initial and Forecasted Stock Prices for<br>Public Bank | 64 |
| Figure 5.6.3 | Comparison of Initial and Forecasted Stock Prices for<br>CIMB        | 64 |

## LIST OF TABLES

| <b>Table Number</b> | <b>Title</b>                                          | <b>Page</b> |
|---------------------|-------------------------------------------------------|-------------|
| Table 3.2.2.1       | Data Extraction Approaches across Selected News Sites | 20          |
| Table 3.1.2.1       | Testing Samples Summary                               | 29          |
| Table 3.1.2.2       | Selected Versions of GPT Models for Testing           | 29          |
| Table 3.7.1.1       | Specifications of laptop                              | 38          |
| Table 5.2.1         | News Sources Directory                                | 51          |
| Table 5.3.1.1       | Model Testing Result: Format                          | 52          |
| Table 5.3.2.1       | Model Testing Result: Trends Comparison               | 53          |
| Table 5.3.1         | Model Testing Result: Stability Test                  | 55          |
| Table 5.5.1.1       | ADF Test Result                                       | 58          |
| Table 5.6.1         | Comparison of RMSE for Different Sentiment Data Lags  | 63          |
| Table 5.7.1         | Prediction Performance for Each News Sources          | 66          |

## LIST OF SYMBOLS

|              |                  |
|--------------|------------------|
| $\beta$      | beta             |
| $\Omega$     | Ohm (resistance) |
| $\rho$       | rho              |
| $\alpha$     | alpha            |
| $\gamma$     | gamma            |
| $\phi_i$     | phi              |
| $\epsilon_t$ | epsilon          |
| $\hat{Y}_t$  | y hat            |

## LIST OF ABBREVIATIONS

|                |                                                         |
|----------------|---------------------------------------------------------|
| <i>NLP</i>     | Natural Language Processing                             |
| <i>AI</i>      | Artificial Intelligence                                 |
| <i>EMH</i>     | Efficient Market Hypothesis                             |
| <i>AMH</i>     | Adaptive Market Hypothesis                              |
| <i>SVM</i>     | Support Vector Machine                                  |
| <i>MPQA</i>    | Multi-perspective Question Answering                    |
| <i>CNN</i>     | Convolutional Neural Networks                           |
| <i>RNN</i>     | Recurrent Neural Networks                               |
| <i>BERT</i>    | Bidirectional Encoder Representations from Transformers |
| <i>ChatGPT</i> | Chat Generative Pre-Trained Transformer                 |
| <i>GPT</i>     | Generative Pre-Trained Transformer                      |
| <i>NST</i>     | New Straits Times                                       |
| <i>FMT</i>     | Free Malaysia Today                                     |
| <i>Bernama</i> | Berita Nasional Malaysia                                |
| <i>ARDL</i>    | Autoregressive Distributed Lag                          |
| <i>CIMB</i>    | Commerce International Merchant Bankers Berhad          |
| <i>CSE</i>     | Google Custom Search Engine                             |
| <i>HTML</i>    | Hypertext Markup Language                               |
| <i>JSON</i>    | JavaScript Object Notation                              |
| <i>SEO</i>     | Search engine optimization                              |
| <i>DOM</i>     | Document Object Model                                   |

# Chapter 1

## Introduction

This chapter presents the background and motivation of the research, the project contributions to the field, and the outline of the project.

### 1.1 Background

In the sphere of investment, a significant of people rely on stock investment for profitability or expansion of business. Stock market prediction is a vital task for investors and financial professionals who seek to optimize their portfolio performance and minimize their risks. Numerous research has shown that people are more concerned with losses compared to gains of the same magnitude [1,2,3,4]. In terms of behavioural science, loss aversion behaviour manifests under conditions of hazardous and inconclusive, such as stock investment. It is certain that an effective method for predicting stock trends will be useful for reducing such loss aversion in the realm of investment.

Several financial theories were proposed for improving investment performance. Among them, the Efficient Market Hypothesis (EMH) suggests that intense competition between investors will lead to an efficient market, which prices show volatility when new information arises [5,6]. In such an efficient market, trends of the stock impound every relevant information, and prices will show uncertainty or randomness accordingly. For example, the stock price will rise sharply when the good news is revealed. Hence, knowing more information ahead of the market is the key to profitable investing [5].

However, the complexity of the stock market makes it difficult to determine the relevant factors for forecasting prices or trends accurately. The Adaptive Markets Hypothesis (AMH) provides a different interpretation and explains the influence of individual behaviour on stock market [7,8]. According to this hypothesis, there are possibilities that investors make wrong decisions, and in the meantime, investors are constantly learning and adapting to the stock market. This leads to opportunities for arbitrage and abnormal data in stock trends. Accordingly, the overall returns tend to be delayed, and specific strategies might only be effective in specific stock markets. Under such behavioural biases, information adequacy is considered significant for arbitrage and adapting to innovation. Despite the contradiction between EMH and AMH,

both support the same argument: the information asymmetry is the key factor in predicting stock trends.

With the goal of gathering information, investors can obtain massive amounts of data from the web to identify factors influencing stock movements. Considering that individual decisions can affect stock prices, any related information—including text, images, charts, and graphs that present a comprehensive overview—constitutes factors that can be used to forecast volatility in stock prices. Previous literature has concluded that there is a correlation between stock movements and news sources, including newspapers and media [9,10,11]. The proliferation of textual documents, such as news articles, blogs, forums, reviews, and others, exerts a significant influence on public beliefs. Text data mining techniques have become effective in predicting stock trends.

The advancements in natural language processing (NLP) enable machines to understand human languages. Sentiment analysis, as a branch of NLP, is increasingly being applied to various areas to understand and investigate people's emotions or attitudes toward a specific topic or person. From past research, this technique has been proven to improve the accuracy of stock prediction [12,13,14,15]. Advanced techniques for sentiment analysis include machine learning based techniques such as Naïve Bayes and Support Vector Machine (SVM), lexicon based techniques such as Multi-perspective Question Answering (MPQA), deep learning based techniques such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), and transformer based techniques such as BERT (Bidirectional Encoder Representations from Transformers) [16,17,18,19]. Since the release of ChatGPT (Chat Generative Pre-Trained Transformer) by OpenAI, transformers have received considerable attention in various areas. Their success has inspired researchers to apply generative AI such as GPT (Generative Pre-Trained Transformer) models to evaluate their performance in sentiment analysis [20,21]. Foundational models have been shown to be effective in sentiment analytical capabilities.

Beyond its application in stock prediction, time series analysis is a powerful tool used across a wide range of domains to forecast future events based on past data. In the field of meteorology, for instance, it is essential for predicting weather patterns and climate changes, helping to prepare for and mitigate the effects of extreme weather conditions [22,23]. In the healthcare sector, time series analysis is employed to predict disease outbreaks and patient admissions, which assists in resource planning and epidemic prevention [24]. Additionally, in the realm of retail and e-commerce, this method is used to analyse consumer behaviour trends



and predict future sales, thereby optimizing inventory management and marketing strategies [25,26]. The broad utility of time series analysis in these diverse fields underscores its effectiveness as a predictive tool, reinforcing its value in stock market analysis as well.

Despite research and proposed techniques on stock movement prediction abound, existing methods still overlook various relevant factors. These include ignoring the impact of news sources, using simple or outdated sentiment analysis techniques that fail to capture the nuances and context of financial texts, and a lack of investigation into this matter specifically in Malaysia.

Therefore, this research proposes a novel method that leverages state-of-the-art GPT model for sentiment analysis of financial news and compares the influences of different news sources on stock trends. This method aims to improve the accuracy and reliability of stock market prediction by incorporating advanced GPT model and diverse data sources.

## **1.2 Problem Statement**

Existing methods for stock market prediction often overlook a critical factor: the influence of various news sources. Ignoring such nuance can lead to inaccurate predictions and missed investment opportunities. In general, the content of news from various sources tends to be similar when discussing the same specific areas and timing. However, the audience groups might vary across different news portals. According to the AMH, individuals in the stock market consistently impact stock movements. Despite assumptions about the similarity of different news sources, differences or similarities between sources in terms of impact remain to be verified, specifically in the Malaysia region. Thus, a fundamental problem in the field of stock market prediction is the lack of a comprehensive approach that considers the diverse landscape of Malaysia news sources and their varying impacts on stock prices.

Another challenge is the limitation of current sentiment analysis techniques applied in stock market prediction. As the huge potential of transformers in sentiment analysis has only recently been realized, and several state-of-the-art transformer models have just been released, existing methods seldom utilize such sentiment analysis techniques to capture the subtle nuances and contextual intricacies present in financial texts. Especially GPT model such as GPT-4 Turbo that recently issued by OpenAI, these advanced techniques might have the potential to improve the accuracy in stock prediction but remain to be tested. Hence, another issue in this domain is that ignoring the newest technologies may underestimate the ability to accurately gauge market sentiment, thus impeding the precision of stock trend predictions.

Lastly, while the Autoregressive Distributed Lag (ARDL) model is robust in dealing with non-stationary data in time series analysis, it may not inherently accommodate the rapid fluctuations and volatile nature of the Malaysian stock market when fed with variable lags and external sentiment indicators like news impact. This study will therefore test the efficacy of ARDL in predicting short-term stock movements by focusing on operationalizing this model in a volatile environment and assessing its predictive accuracy with the inclusion of news sentiment as an external regressor. This approach will provide insights into the real-world applicability of ARDL under specific economic conditions and supervisory constraints.

The primary motivation for this research is to enhance the accuracy and efficiency of stock market predictions, which is crucial for investors and financial professionals, especially in the volatile Malaysian market. By delving into the impacts of various news sources within Malaysia, this study aims to uncover which outlets significantly influence market movements, enabling investors to make more informed decisions. Furthermore, this project seeks to explore and potentially expand the application of advanced sentiment analysis technologies, like the latest transformer models including GPT-4 Turbo, to better understand and predict market dynamics. The integration of these modern analytical techniques with the ARDL model to assess short-term stock movements underlines the innovative approach of this research. Through a detailed examination of news sentiment and its integration into traditional forecasting models, this research aspires to develop a more robust and adaptable prediction method, tailored specifically for the complexities of the Malaysian financial market.

### **1.3 Research Objectives**

The primary aim of this project is to evaluate the performance and potential of GPT model for financial sentiment analysis and its application in predicting stock market movements. To achieve this overarching goal, the project is divided into the following sub-objectives:

- To design and conduct a comprehensive test to evaluate the capabilities and performance of GPT series models on sentiment analysis
- To investigate how different news sources influence stock market movements, providing insights into which sources have the most significant impact.
- To develop a simplified analysis process by comparing sentiment scores with stock trends, streamlining the use of sentiment analysis in stock prediction.

By achieving these objectives, the research project aims to offer valuable insights into the influence of news sources on short-term stock trends, enhance sentiment analysis

methodologies, and aid investors in making more informed decisions in the dynamic and complex world of financial markets.

#### **1.4 Research Contributions**

This research aims to make significant contributions to the field of finance and sentiment analysis in Malaysia. Firstly, by comparing different Malaysia news sources, this research seeks to uncover the varying levels of influence they exert on stock trends. The findings will enable investors to identify prominent news outlets that significantly impact market sentiment, providing a valuable resource for market analysts and traders in Malaysia region.

Secondly, the research intends to develop an advanced sentiment analysis methodology tailored to the financial domain, capable of discerning subtle nuances in news sentiment from different sources. This enhancement will contribute to the overall improvement of sentiment analysis techniques in finance.

Finally, through a better understanding of the influences of various news sources on stock trends, this research will empower Malaysia investors with more accurate and timely information, facilitating well-informed investment decisions. The study's insights will be valuable for portfolio managers, traders, and financial institutions seeking to optimize their strategies.

#### **1.5 Report Organization**

This research report is structured into six chapters to provide a comprehensive exploration of stock market prediction using advanced sentiment analysis. Chapter 1: Introduction sets the stage by outlining the research problem, objectives, and significance. Chapter 2: Literature Review delves into existing methodologies and the roles of news sentiment in financial markets. Chapter 3: System Design explains the methodologies used, including data collection and system setup. Chapter 4: System Implementation and Testing discusses the practical application and testing of the designed system. Chapter 5: System Outcome and Discussion presents the results, analyzing the efficacy of the ARDL model and sentiment analysis in predicting stock price movements. Finally, Chapter 6: Conclusion summarizes the findings, evaluates the research objectives, and suggests directions for future research, highlighting improvements and potential expansions of the study.

# Chapter 2

## Literature Review

### 2.1 Previous works on Stock Prediction and Sentiment Analysis

#### 2.1.1 Stock Trend Prediction by News

The study of utilizing sentiment analysis on news in the area of stock market prediction has evolved significantly over the years, garnering widespread attention from researchers and financial analysts alike. This section embarks on a comprehensive exploration of seminal works and recent advancements in sentiment analysis applied to stock market prediction. By surveying an array of studies spanning from 2015 to the present, the ever-growing sophistication of methods and the expanding body of knowledge surrounding the interplay between financial news sentiment and stock price movements are uncovered.

In 2015, Peng and Jiang proposed a model that applies word embedding and deep neural networks for stock prediction using financial news and historical price data [27]. The model consists of a conventional multi-layer perceptron with many hidden layers that learns from automatically selected keywords, polarity scores, and category tags as features representing the news articles. These features were then combined with price-related attributes to train a deep neural network classifier. Furthermore, the authors introduced a novel method to extend their predictions to additional stocks that might not be explicitly mentioned in financial news by exploiting a stock correlation graph, which transfers predictions to unseen stocks based on their price correlations. Their results indicated an error rate of 43.13% on a standard financial dataset, highlighting a significant improvement over the baseline system, which relied solely on price-related features.

In 2016, Joshi et al. presented an innovative model to forecast stock price trends through sentiment analysis of news articles [28]. Their method encompassed several steps, commencing with the collection of news articles and stock price data spanning. These news articles having gone through pre-processing, involving tokenization, stop word removal, stemming, and polarity score assignment based on an established lexicon of positive and negative words. The polarity score is a numerical value that indicates the overall sentiment of a news article, calculated by subtracting the number of negative words from the number of positive words.

Subsequently, the news articles were transmuted into document vector via the TF-IDF scheme. These vectors with polarity scores were then used to train the classifier model. The authors employed three supervised machine learning algorithms (Naïve Bayes, Random Forest, and SVM) to classify the news articles as either positive or negative, based on their TF-IDF vectors and polarity scores. They claimed that their proposed model achieved an accuracy ranging from 88% to 92% by using Random Forest, around 86% by SVM and approximately 83% by Naïve Bayes algorithm when tested with new data from Jan 2016 to April 2016. They also concluded that news sentiment and stock price movements have a strong correlation.

In 2017, Khedr et al. proposed a model that combines sentiment analysis of news articles and historical stock prices to predict the future stock market behaviour [14]. They used Naive Bayes algorithm to classify news articles into positive or negative sentiments based on n-gram features and TF-IDF weighting. They also used K-Nearest Neighbor algorithm to predict the stock trend (raise or fall) based on the news sentiment and the numeric attributes of open, high, low, and close prices. They evaluated their model on three companies and achieved prediction accuracy up to 89.80%. Their study demonstrated the importance of considering different types of news and numeric data together for improving the prediction accuracy of stock market behaviour.

In 2021, Nemes and Kiss conducted a comprehensive analysis employing four sentiment analysis tools, namely TextBlob, NLTK-VADER Lexicon, Recurrent Neural Network (RNN), and BERT, with the primary objective of examining and categorizing various business news headlines [29]. This analytical approach helped to provide insight into the value fluctuations in the stock market. Their data collection process involved the extraction of headlines from economic news source related to five companies. These headlines were subsequently matched with the corresponding daily closing prices of the associated stocks, revealing substantial insights into the influence of news sentiment on stock market dynamics. Notably, their investigation uncovered significant difference among the models concerning the impact of emotional values on stock market value changes, as demonstrated through correlation matrices. They concluded that RNN and BERT outperformed the other tools in terms of accuracy and reliability.

In June 2022, Fazlija and Harder proposed a method to use financial news sentiment for stock price direction prediction [30]. They employed the transformer model FinBERT, which is based on BERT and pre-trained on a financial text dataset. The model was fine-tuned on the same dataset, where each text is assigned a positive, negative, or neutral sentiment score by

human annotators. FinBERT then predicted the sentiment scores of news articles based on their titles, content, or both. These sentiment scores were subsequently classified by Random Forest to predict the price direction with the combination of time series data. The authors evaluated the performance of the model on a large dataset of financial news from Bloomberg and Reuters using different performance measures such as weighted precision, weighted recall, weighted F1-score, Brier score, and Matthews correlation coefficient. The paper concluded that sentiment scores based on news content are particularly useful for stock price direction prediction and that using a Random Forest classifier can improve the results compared to a simple strategy based on sentiment scores alone.

In November 2022, Cristescu et al. proposed a method to improve the accuracy of stock market prediction by using sentiment analysis on financial news headlines [33]. The authors employed the VADER model to obtain the sentiment scores of news headlines from FinViz, a platform for researching the stock market. They then incorporated these scores into linear autoregression models with and without an exogenous factor, and nonlinear autoregression models with an exogenous factor (NARX) to forecast the opening prices of various active stocks. The authors compared the goodness of fit of these models using the R squared value and the F test, and found that the polynomial autoregressions have a higher goodness of fit than the linear ones. They also found that the sentiment scores improved the performance of the linear autoregression model.

In 2023, Usmani and Shamsi introduced an innovative model known as WCN-LSTM (Weighted Categorized News-LSTM) for stock market prediction, leveraging weighted and categorized financial news [32]. This model integrated news headlines from three categories: market-related, sector-related, and stock-related news, according to the structural hierarchy of the stock market. The model learns the weights of each news category from the training data using a neural network layer. Moreover, sequential learning was embraced through the inclusion of LSTM layers, facilitating the processing of input data, which consists of stock prices, technical indicators, and news sentiment scores. The research compared the proposed model with a baseline model that uses stock prices, technical indicators, one news category and one sentiment dictionary and found that the proposed model outperformed the baseline model. The research also compared the performance of different sentiment dictionaries for each experiment and found that Harvard IV (HIV4) performed better than Loughran and McDonald (LM) and Vader for most sectors. Furthermore, the research reported the results for different

time steps and found that using the last 3 or 7 days of input data resulted in higher prediction accuracy for most stocks than using the last 10 days.

In conclusion, this section reviews various studies that apply sentiment analysis to stock market prediction using financial news. The literature review showcases the diverse and evolving technologies and methodologies used by different researchers, such as deep neural networks, word embedding, machine learning algorithms, recurrent neural networks, transformer models, autoregression models, and LSTM-based models. These studies collectively underline the versatile and evolving nature of sentiment analysis applications in predicting stock market trends, each contributing valuable insights and showcasing the continuous growth of this field.

### **2.1.2 Sentiment Analysis using GPT Models**

GPT stands for Generative Pre-trained Transformer. It is a deep learning algorithm that use unsupervised learning to generate human-like-text. It is developed by OpenAI and has been used in various applications such as language translation, chatbots, text completion and sentiment analysis. One of the earlier version language models, GPT-2, was introduced by Radford et al. in 2019 [28]. It was trained on a dataset of web pages called WebText., which contained text from 45 million website links. The paper showed that GPT-2 was capable of performing various natural language processing tasks, such as reading comprehension, summarization, translation, and question answering, without any explicit supervision or task-specific fine-tuning.

In the field of sentiment analysis, prior research has demonstrated that GPT-2 is capable of effectively performing sentiment analysis. The study conducted by Xie et al. in 2022 [34] delves into a comprehensive investigation of fine-tuning the performance and sensitivity of BERT and GPT-2 for financial sentiment analysis. This research utilized two distinct benchmarks, FiQA and Financial PhraseBank, as the basis for evaluation. The authors categorized textual data into three classes: positive, neutral, and negative sentiments. To understand the models' performance dependencies, the authors conducted meticulous experiments, tweaking various hyperparameters, including the number of frozen transformer layers, batch sizes, and learning rates. They observed their impact on the accuracy and F1-score of these models.

The study's findings unveiled significant insights into the comparative capabilities of BERT and GPT-2 for financial sentiment analysis [34]. Remarkably, GPT-2 showcased greater

resilience and stability when subjected to different hyperparameters, whereas BERT demonstrated high sensitivity, making it susceptible to minor hyperparameter changes. Furthermore, an essential revelation from the research was the criticality of the first 1-6 layers in both models, which were found to contain indispensable information for effective classification. Consequently, the study concluded that GPT-2 emerges as a more suitable choice for developers with less experience.

GPT-2 was followed by several models that used a similar algorithm and structure, but differed in the size of the model and the amount of data used to train the model. For example, GPT-3 was introduced by Brown et al. in 2020 [35] and consisted of 175 billion parameters, which was ten times larger than GPT-2 and the largest non-sparse language model at the time of its publication. The paper by Kheiri and Karimi conducted in July 2023 [21], explored the application of several GPT models for sentiment analysis, involving the extracting and interpretation of opinions, emotions, and attitudes expressed in text. There were three different strategies to leverage GPT models: prompt engineering, fine-tuning, and embedding.

According to [21], prompt engineering used a preset textual task description or prompt to guide the GPT-3.5 Turbo model to perform sentiment analysis. The prompt was designed and refined iteratively to produce the best results. The paper used GPT-3.5 Turbo to generate sentiment predictions and explanations based on the crafted prompt. Fine-tuning involved further training of pre-trained GPT models like Ada, Babbage, and Curie on specific sentiment analysis datasets, aiming to specialize them in this domain. Embedding strategy utilized the text embedding capabilities of GPT-3.5 Davinci to transform text into numerical representations that capture semantic and contextual information. The paper used GPT-3.5 Davinci to generate embeddings for each tweet. These embeddings were subsequently used to train machine learning models XGboost and Random Forest to perform classification.

In the experiment part, the paper compared the performance of these three strategies and various GPT models with existing machine learning solutions on a benchmark dataset of Twitter posts [21]. The ability of GPT models further investigated to handle linguistic nuances related to sentiment, such as emojis, sarcasm, and mixed sentiment. The paper claimed that GPT models offer a promising potential for sentiment analysis, outperforming the state-of-the-art methods by more than 22% in F1-score.

The aforementioned studies have demonstrated significant power of GPT models in sentiment analysis. However, within the GPT model family, there remains ample room for further exploration in the field of sentiment analysis. The most advanced GPT model to date is



GPT-4, which consists of 170 trillion parameters, approximately a thousand times more than GPT-3 [36]. It has achieved human-level performance on various professional and academic benchmarks. Despite GPT-4 currently only offering its text input capability through ChatGPT, a language model-based chatbot developed by OpenAI that engages in conversational interactions, along with an API with a waitlist, a recent study conducted in September 2023 has investigated its potential for sentiment analysis [20].

The authors explored the potential of GPT-4 for sentiment analysis of microblogging messages related to stock market movements [20]. They compared the performance of GPT-4 with BERT, another language model, in extracting sentiments from Stocktwits messages about Apple and Tesla in 2017. They also examined the correlation between these sentiments and the same-day price movements of the stocks. The authors developed a novel method for prompt engineering, which involved designing and refining input prompts to guide GPT-4's output. They followed a set of guidelines and features that enabled GPT-4 to capture nuanced and contextual sentiments, as well as perceived advantages or disadvantages and relevance for the analyzed companies. They presented their results in terms of probabilities for each category of sentiment, advantage or disadvantage, and relation.

Furthermore, the authors took considerable care to frame GPT-4's perspective, providing it with the role of a financial analyst entrusted with the evaluation of the potential impacts of news on either Apple or Tesla [20]. This necessitated not only the definition of a clear subject but also the precise specification of the company of interest, particularly in cases where multiple representations existed. The desired output format, structured as a JSON object with three key components—namely, "sentiment," "advantage/disadvantage," and "relevance"—further facilitated comprehensive sentiment analysis. To add a layer of clarity and transparency, the authors extracted probabilities for each sentiment, advantage, and relation category, aiding both GPT-4's understanding and the user's interpretation of result uncertainty and variability. These results were succinctly presented using lists and brackets, ensuring ease of comparison.

In their evaluation, the authors employed logistic regression to assess the predictive power of the sentiment models [20]. The results showed that GPT-4 consistently outperformed BERT in most cases, achieving high correlations between its sentiment probabilities and the stock price movements of Apple and Tesla. The highest correlation achieved by GPT-4 was 71.47% for Apple in May 2017, while the average correlation over the six months was 59.76%. These values were significantly higher than those of BERT, which had an average correlation of 54.98%, and those of a random walk model, which had an average correlation of 50%. The

paper also delves into the broader implications and limitations of employing GPT-4 for financial sentiment analysis, underlining the importance of a well-structured prompt in achieving accurate and contextually relevant sentiment analysis results. This methodology's power was particularly evident when compared to a simpler prompt, demonstrating its superiority in generating more precise and comprehensive sentiment analysis outcomes.

In conclusion, these studies applied various types of GPT models to perform sentiment analysis, and prove the GPT's remarkable capabilities in this domain. Beginning with GPT-2 model, the subsequent comparative study between GPT-2 and BERT reaffirmed GPT-2's dominance due to its stability across different hyperparameters. Furthermore, the GPT models offer a promising potential for sentiment analysis, outperforming the state-of-the-art methods in most cases and demonstrating their ability to handle linguistic nuances and contextual information. The latest iteration, GPT-4, with a staggering 170 trillion parameters, presented a new frontier in sentiment analysis, as demonstrated in a recent study examining its effectiveness in stock market sentiment analysis compared to BERT. Collectively, these findings underscore the consistent evolution and promising future of sentiment analysis facilitated by GPT models, encouraging further exploration in this dynamic field.

## **2.2 Limitations of Previous Studies**

The previous studies examining sentiment analysis for stock market prediction using financial news and utilizing GPT models to perform sentiment analysis exhibit notable limitations that merit consideration. Firstly, Peng and Jiang [27] did not incorporate the sentiment or tone of the news articles into their analysis, overlooking a significant factor that could influence stock price movements. Similarly, Joshi et al. [28] limited their analysis to news articles sourced exclusively from a single provider, the New York Times. This approach, while valuable, potentially restricted the diversity and comprehensiveness of their sentiment analysis.

Furthermore, Khedr et al. [14] relied heavily on a solitary news source, which may have constrained the applicability and diversity of their research findings. In contrast, Nemes and Kiss [29] omitted the consideration of important dynamics like news volume, frequency, or recency's impact on stock price movements. Another limitation emerges in the work of Fazlija and Harder [30], who did not delve into the influence of distinct news types, such as earnings reports or mergers and acquisitions, on stock price direction prediction. Furthermore, Cristescu et al. [31] solely used a single stock market index (S&P 500) for their prediction task, which

could limit the generalizability and robustness of their model to other markets or individual stocks.

In terms of model comparisons, Usmani and Shamsi [32] did not contrast their LSTM-based model with alternative models and did not assess its performance across various time horizons. Similarly, Xie et al. [34] utilized only two benchmarks, FiQA and Financial PhraseBank, which are relatively small and may not have adequately represented the diversity and complexity of financial texts.

Kheiri and Karimi [21] restricted their sentiment analysis to a single benchmark dataset, Twitter posts, which may not have been representative or directly relevant to stock market prediction. Finally, the study conducted in September 2023 [20] had a narrow scope, focusing exclusively on two companies, Apple and Tesla, for stock market sentiment analysis. This limited focus may have hindered the research from achieving the diversity and breadth needed to draw comprehensive conclusions.

### **2.3 Proposed Solutions**

To address the identified weaknesses and limitations of previous studies, this research adopts an advanced approach, utilizing GPT model for financial news sentiment analysis to predict stock market movements, significantly enhancing precision and context awareness compared to previous lexicon-based methods.

Besides, this project innovates by assessing the impact of diverse news sources on stock trends and identifying the most influential ones. It encompasses data collection from varied sources, providing a comprehensive analysis of their contributions to stock dynamics.

In summary, this project address limitations by utilizing advanced sentiment analysis techniques, diversifying data sources, and considering the impact of various news types. These refinements are expected to result in a more thorough and accurate analysis of the relationship between news sentiment and stock trends. Detailed methodologies will be presented in subsequent sections.

# Chapter 3

## Methodology

This section outlines the research model and propose methodologies adopted in this study.

### 3.1 Research Design

This research adopts a mixed-method research design that combines quantitative analysis of historical stock price data with sentiment analysis of news articles. The goal is to evaluate comprehensively the influence of various news sources on stock prices, offering a holistic perspective on the intricate dynamics of financial markets. Subsequent research processes will work towards achieving this objective.

Figure 3.1.1 shows the overview of the proposed method. The flow of this study begins by collecting data using web scraping techniques and API services. After data collection, the data will be pre-processed for news and stock data. Specifically, the news data will be analysed its sentiment polarity in this phase. Following the pre-processing, both data will be integrated align in daily basis. The next phase will analyse the data by time series analysis concept. Lastly, ARDL model will be adopted to forecast stock data future trend.

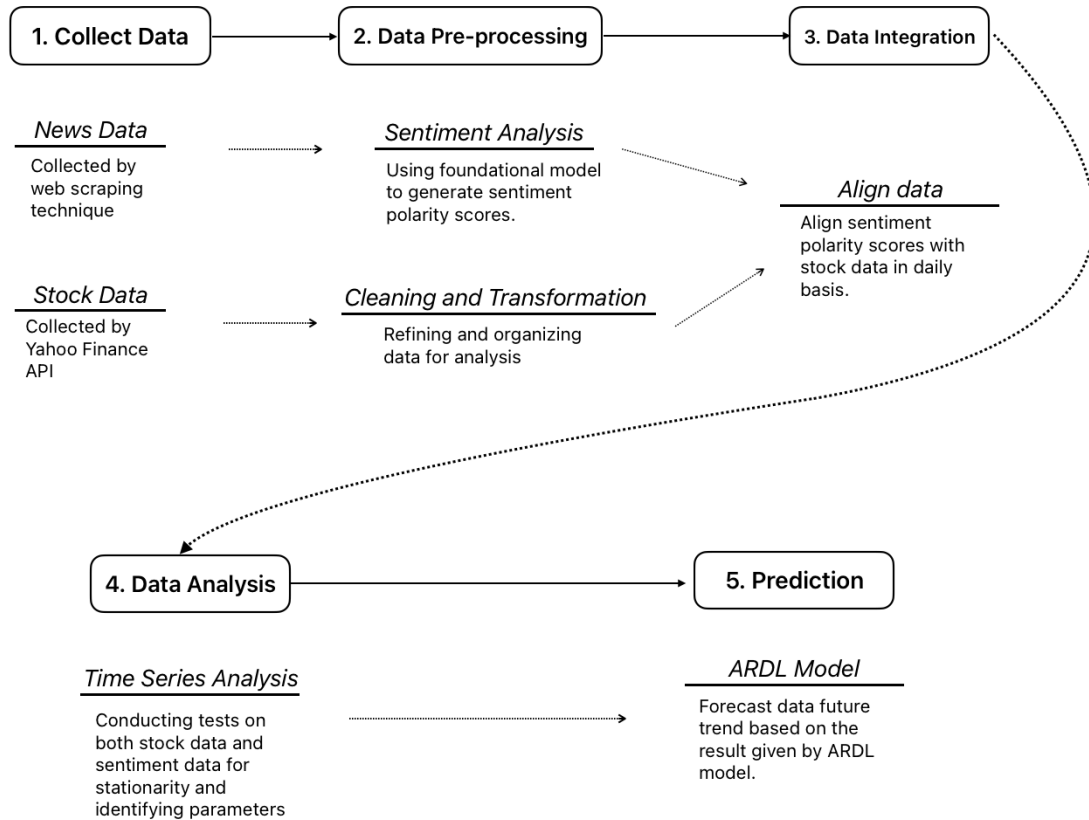


Figure 3.1.1 Research Model

The methodological details of each step are discussed below.

### 3.2 Data Collection

For this study, news data and stock data will be collected in the first phase. In the context of data types, data can be divided into two main categories: primary data and secondary data.

**Primary Data** is the direct information offered by its source without any external interpretation. It ensures high accuracy and is easily manageable. Common methods to collect primary data include questionnaires, interviews, surveys, etc. [37]. In this study, primary data consists of stock market data collected using the Yahoo Finance API. This API provides real-time and historical stock prices, trading volumes, and other market-related information for a set of predetermined stocks. The stocks chose to be studied are Public Bank Berhad and Commerce International Merchant Bankers Berhad (CIMB). The collection period for this study is planned to span two years, from 2022 to 2023, to capture recent market trends and behaviours.

On the other hand, **Secondary Data** is data that has already been processed, typically involving interpretation or aggregation. It offers benefits such as cost-effectiveness and time efficiency. Effective sources for gathering secondary data include general websites, journals, newspapers, etc. [37]. In this research, secondary data consists of news articles directly relevant to the stock companies being studied. This data will be collected from various online news sources using web scraping methods. Google API and Python libraries such as BeautifulSoup will be used to automate the extraction of news articles and associated metadata from these websites. Additionally, the collected data will be stored in a NoSQL database to facilitate scalable data management. The same two-year period is targeted for the collection of news data to ensure it parallels the stock data timeline.

### 3.2.1 Google Custom Search JSON API

Figure 3.1.1.2 illustrates the step-by-step process used for collecting news data in this research project. The first step involves extracting URLs using the Google Custom Search JSON API, which enables tailored searches across specific websites predetermined by the user. This API is a robust tool offered by Google that allows developers to create custom search engines for their websites or applications. It enables the programmatic retrieval of web search results in JSON format, providing significant customization and flexibility. Users can specify parameters such as search terms, the number of results, and particular domains to search within, allowing them to finely tailor the search experience to suit their application needs. As part of the broader Google Custom Search Engine (CSE) platform, this API offers a managed and customizable search solution that can index and search specific content on the web. In this study, the API is crucial for retrieving links to news articles relevant to specified stock companies, thus enhancing the precision and relevance of the search results.

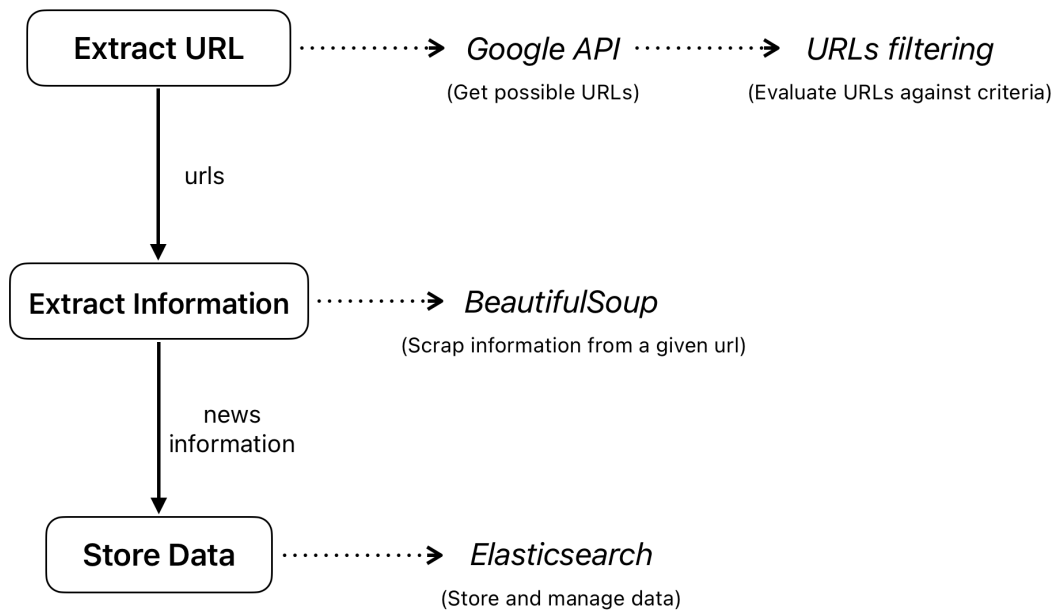


Figure 3.1.1.2 Flow of Automated News Data Collection

In addition, the utilization of Google search operators in the Custom Search API significantly enhances the precision and relevance of search results, making it a powerful tool for extracting targeted information from the web. Google search operators are special characters and commands that extend the capabilities of regular text searches [38]. When incorporated into a query, these operators enable users to narrow down search results based on specific criteria, such as exact phrases, the exclusion of certain words, the location of search terms within the content, and specific domain searches. For instance, using the “site:” operator restricts the search results to those from a particular website or domain. Similarly, the “inurl:” operator ensures that search results include pages with certain words in their URLs, adding another layer of specificity. An example of a Google search using these operators in this study is shown below:

*site:website.com "Stock Company" after:2022-01-01 before:2023-01-01*

The Google search operators will be utilized in the study, such as “site:”, “after:”, and “before:”, each fulfilling specific functions to refine search results. The “site:” operator restricts the search to a single domain, crucial for ensuring that information is sourced exclusively from a specific website. Meanwhile, the “after:” and “before:” operators define a date range, limiting search results to documents published within a specific timeframe. This is invaluable for

focusing the data collection on the most pertinent and timely content, essential for the research analyzing news published within the defined period. The use of double quotes around search terms, such as in “Stock Company,” specifies that the search engine should look for exact matches of the phrase. This precision is fundamental in narrowing down results to the most directly relevant data, thereby enhancing the efficiency and effectiveness of the data collection process.

### **3.2.2 Data Extraction**

Following the collection of URLs, the next step involves extracting specific content from these articles. This task is executed using the BeautifulSoup library, a Python tool designed for efficient HTML and XML document parsing, allowing for the detailed extraction of news information from the web pages.

Ten news sources have been selected to collect news information in this research. These sources include major Malaysian outlets such as The Star, New Straits Times, Malay Mail, Free Malaysia Today, Bernama, and The Edge Malaysia. Besides, international platforms like Yahoo News, and local publications including The Sun, Business Today, and Sinar Daily are also incorporated. This diverse selection is intended to provide a comprehensive view of the media landscape, ensuring varied perspectives and extensive coverage of relevant financial and corporate events that impact the stock market.

The methods will be used to scrape news from the ten selected sites follow a similar structured approach but are customized to the specific HTML and JSON formats. Below, the data extraction approaches that are effective for the collected news sources are discussed.

#### **3.2.2.1 JSON Data Extraction**

JSON Data Extraction method allows for a streamlined approach to obtaining data directly from webpages by parsing embedded JSON within script tags. This technique can be particularly advantageous when dealing with large volumes of data, or when data is dynamically generated, as it often provides a more direct route to the structured information without the overhead of navigating complex HTML structures. In practice, this method can be instrumental in automating the extraction of information for news aggregation services or competitive intelligence platforms where timely and accurate data retrieval is crucial. Meanwhile, leveraging JSON embedded within HTML pages ensures that the data extracted is



in a format that is readily usable for further data processing or analysis tasks. This can significantly reduce the preprocessing steps typically required when scraping content directly from HTML elements. Example code snippet is shown in Figure 4.2.2.

### **3.2.2.2 HTML Meta Extraction**

The HTML Meta Extraction approach targets meta tags within the HTML head section, which store metadata about the webpage. This metadata includes details like the article title, description, author, and publication dates. Accessing meta tags directly allows for the efficient gathering of preliminary information about articles, such as titles and authorship, commonly utilized for search engine optimization (SEO). Therefore, extracting data from meta tags represents an effective method for acquiring essential metadata. This method is particularly beneficial for applications that require a rapid assessment of web content relevancy, such as news aggregators and content management systems where speed and efficiency are paramount. Moreover, this technique supports the creation of automated systems that can monitor changes in content and update databases in real-time, ensuring that the most current and pertinent information is readily available for analysis or dissemination. Figure 4.2.3 illustrates an example Python script that utilizes this extraction method to demonstrate how meta tags can be programmatically accessed and parsed for relevant information.

### **3.2.2.3 Hybrid Extraction**

Hybrid Extraction combines the strengths of both JSON Data Extraction and HTML Meta Extraction to ensure a thorough data retrieval process. This method is particularly beneficial when dealing with complex web pages that utilize both structured JSON scripts and HTML meta for metadata. Hybrid Extraction ensures that all possible information is harnessed by first extracting readily available metadata from JSON, which is structured and easy to parse, and then supplementing it with detailed content scraped from the HTML meta. This approach is designed to maximize data completeness, providing a robust solution for scraping websites where information is distributed across different formats. It bridges the gap between structured data retrieval and the flexibility needed to handle the diverse ways content is presented on modern web platforms, making it ideal for comprehensive web scraping projects that require both broad and detailed data sets.

### 3.2.2.4 HTML Content Extraction

HTML Content Extraction is crucial for acquiring the complete text of articles when the desired information extends beyond simple metadata. This method involves navigating the Document Object Model (DOM) to identify and extract content directly from specific HTML elements such as “div”, “p”, “article”, and other relevant tags that typically house the main body of textual content. This technique requires a deep understanding of each site’s layout and the typical structures that contain article content. By employing tools like BeautifulSoup, this approach allows for the extraction of cleanly formatted text, which is essential for any comprehensive content analysis or data aggregation that depends on full articles. It’s especially useful in scenarios where metadata alone does not provide sufficient insight, or when the textual content itself is the subject of study, such as in natural language processing applications. An example of extracting a full article using this approach is shown in Figure 4.2.4 with a simple Python script.

Table 3.2.2.1 Data Extraction Approaches across Selected News Sites

| News Site                       | Applied Approaches |           |        |              | Able to get full article |
|---------------------------------|--------------------|-----------|--------|--------------|--------------------------|
|                                 | JSON Data          | HTML Meta | Hybrid | HTML Content |                          |
| <i>The Star</i>                 | ✓                  |           |        | ✓            | ✓                        |
| <i>News Straits Times</i>       | ✓                  |           |        | ✓            | ✓                        |
| <i>Yahoo News</i>               | ✓                  |           |        | ✓            | ✓                        |
| <i>Malay Mail</i>               | ✓                  |           |        | ✓            | ✓                        |
| <i>Free Malaysia Today</i>      |                    |           | ✓      | ✓            | ✓                        |
| <i>Berita Nasional Malaysia</i> |                    | ✓         |        | ✓            | ✓                        |
| <i>The Edge Malaysia</i>        |                    |           | ✓      |              | ✓                        |
| <i>The Sun</i>                  |                    | ✓         |        | ✓            | ✓                        |
| <i>Business Today</i>           |                    | ✓         |        | ✓            | ✓                        |
| <i>Sinar Daily</i>              | ✓                  |           |        |              | ✓                        |

### 3.2.3 Elasticsearch

Finally, the retrieved data is stored in Elasticsearch, a NoSQL database renowned for its rapid data retrieval capabilities and robust handling of large volumes of unstructured data. As an open-source search and analytics engine built on the Apache Lucene library, Elasticsearch excels in full-text search, real-time data processing, and rich data analysis capabilities, all facilitated through an intuitive RESTful API and JSON-based querying language. Its distributed nature allows for impressive scalability, enabling it to manage and scale large datasets seamlessly across multiple nodes.

Elasticsearch organizes data into structures called “indexes,” akin to databases in traditional relational database systems, with unique names for operations such as indexing, searching, updating, and deleting. Within these indexes, the fundamental unit of storage is a “document,” a JSON-formatted record, each identified by a unique ID. Moreover, Elasticsearch operates on “nodes,” which are individual servers that form part of a “cluster”—a collection of nodes that hold the entire data and share tasks such as indexing and searching. For scalability and resilience, data within an index can be divided into “shards,” which distribute the data across multiple nodes, with “replicas” serving as copies of these shards to prevent data loss and enhance query performance.

For this project, the integration of Elasticsearch offers significant benefits due to its flexibility in handling diverse information collected from various news sources and its capability to manage data over extended periods. This feature is crucial for ensuring that the voluminous and varied data gathered are not only processed and analyzed efficiently but are also stored securely for future analytical purposes.

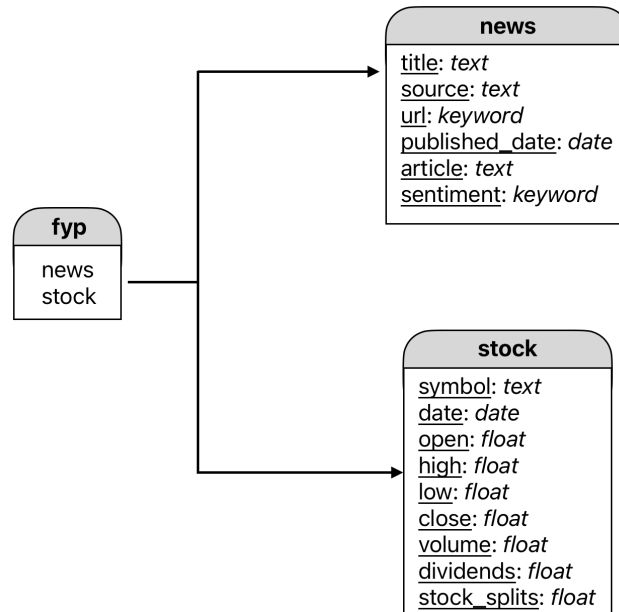


Figure 3.1.1.7 Elasticsearch Schema

The database schema designed for this research efficiently categorizes and manages extensive data sets pertaining to news articles and stock market information. Structured within the Elasticsearch NoSQL database environment, this schema comprises two primary indices: “news” and “stock”, which fall under the larger “fyp” index to consolidate data relevant to this project. The “news” index stores articles with fields such as title, source, URL, published date, article and sentiment, facilitating detailed text-based analysis and metadata retrieval. Each field is tailored to capture specific attributes of the news data, with title, article and source as text for full-text search capability, URL as keywords for precise querying, published date as a date and sentiment as a keyword to reflect the tone analysis results.

Simultaneously, the “stock” index captures financial data with fields like symbol, date, open, high, low, close, volume, dividends, and stock splits. Each of these fields serves to record comprehensive financial metrics with symbol and date ensuring accurate identification and temporal alignment of stock records, while open, high, low, close, and volume provide the quantitative measures essential for financial analysis, recorded as floats. Additional financial events such as dividends and stock splits are also tracked to account for their impact on stock performance. This dual-index schema not only ensures a structured approach to data management but also enhances the capability to perform cross-referential analyses between news sentiment and stock market reactions, thereby supporting the overarching goal of investigating the interplay between media sentiment and market dynamics.

### **3.3 Data Preprocessing**

Prior to analysing the collected data, it is essential to undergo a rigorous data preprocessing phase. Preprocessing is fundamental to enhancing the quality of data, thereby ensuring the reliability and validity of the results obtained from subsequent analyses. In this section, the methods and techniques employed to preprocess the news and stock data will be discussed.

#### **3.3.1 Stock Data**

In the preprocessing of stock data for this research, several crucial steps will undertake to ensure that the data is formatted appropriately for accurate analysis. Initially, the timestamps of the stock data will be adjusted to the “Asia/Kuala Lumpur” time zone to align with the local market context of the study. This adjustment is vital for accurate temporal analysis as it standardizes all data points to a specific geographical time zone, accurately reflecting the actual trading hours and conditions pertinent to the dataset.

Following the time zone conversion, the date and time information will be formatted into a consistent structure, “Year-Month-Day Hour:Minute:Second.” This uniformity is essential, particularly when merging with news data, ensuring that all data elements across different datasets align precisely on temporal features.

The data will then be indexed by this newly formatted datetime information. To address any missing data points due to non-trading days or other anomalies, the dataset will be resampled to a daily frequency. During this resampling process, any gaps in the data will be filled using linear interpolation, which estimates missing values by drawing a straight line between available data points. This method ensures a continuous, gap-free time series, critical for any subsequent time series analysis or predictive modelling.

Lastly, a logarithmic transformation (base 10) will be applied to the stock prices to stabilize variance and normalize the distribution. This transformation is especially useful in financial time series to moderate the effects of significant fluctuations or to bring large values into a comparable scale, thereby enhancing the analytical robustness and sensitivity to subtle dynamics in the data.

### 3.3.2 News Data

Following the preprocessing of stock data, the next critical step involves the preprocessing of news data. This phase is essential to prepare the data for subsequent analysis, particularly transforming it into numerical representations that capture the sentiment of each news article. In this research, sentiment analysis is applied to assess the emotional tone behind the content, which is crucial for understanding how news might influence market behaviours. Moreover, generative AI techniques are employed to synthesize this data into formats that can be directly correlated with market reactions.

Sentiment Analysis and Opinion Mining are primarily tasked with classifying text based on subjective information. The most common form, sentiment polarity classification or positive-negative classification, aims to determine whether the content conveys a positive or negative sentiment. This task requires identifying subjective content to classify sentiments and was a foundational aspect of sentiment analysis in early research [39]. Beyond binary classification, where outputs are labelled as positive or negative, a neutral category is often considered as the third class. This indicates either objective text that does not contain sentiments or subjective text that includes a mix of positive and negative sentiments. Furthermore, fine-grained sentiment classification involves labelling problems on different scales, such as rating levels (e.g., 1 to 5 stars), probabilistic classification (e.g., 0 to 100%), emotion classification (e.g., angry, happy), or stance classification toward a topic (support, against, or neutral).

In terms of data analysis, granularity refers to the different levels of detail in the data. According to Zong et al. [39], the granularity of sentiment analysis tasks can be categorized into four levels. Document-level sentiment analysis classifies the overall sentiment of an entire document, such as a movie review. Sentence-level sentiment analysis focuses on the sentiment expressed within a single sentence. Word-level sentiment analysis examines the smallest units of sentiment expression, such as words and phrases, to construct a sentiment lexicon. Lastly, aspect-level sentiment analysis extracts aspects associated with an opinion target in the text and defines the fine-grained sentiment polarity toward a specific aspect. In this research, the focus is on analysing news articles at the document level by assessing the overall sentiment of each article and assigning a score that represents this sentiment.

Document-level sentiment analysis is particularly effective when the sentiment across the entire document is uniform or when specific sentiment nuances within different sections of the document are not critical for the analysis. Traditional methods like lexicon-based approaches,

which rely on predefined lists of words with assigned sentiment values, often struggle with context and sarcasm in longer documents. Similarly, standard machine learning and even deep learning models may require extensive feature engineering or large amounts of labelled data to accurately capture the overall sentiment of entire documents. In contrast, generative AI, especially models trained on vast amounts of text data, can better understand and generate human-like text, effectively capturing the nuances, context, and complexities of language in full documents. This makes generative AI particularly advantageous for document-level sentiment analysis, as it can process and analyse large blocks of text holistically, adeptly recognizing subtleties and variations in sentiment that other methods might overlook.

Generative AI processes tasks based on prompts, which are inputs typically in text or image format that instruct the AI to generate the desired content. A well-defined prompt can significantly enhance the efficiency and effectiveness of generative AI; however, using more tokens results in higher costs. Therefore, minimizing prompt length without compromising task relevance is considered cost-effective, especially if the model is sufficiently task-capable. Moreover, different foundation models vary in cost, with higher-cost models typically offering superior performance within certain scopes. However, these may not always be necessary; for instance, if a mid-tier model delivers similar sentiment analysis outcomes as a high-tier model, the more economical option is deemed sufficient. This is particularly relevant for this project, which focuses on the trends and movements of sentiment polarities rather than their precise values. If a lower-cost model can capture these trends comparably to a higher-cost model, it is considered to provide sufficient performance.

Inspired by these considerations, the selection of the foundational model and the crafting of prompts are guided by two primary criteria in this research:

1. **Stability**: The same prompt should consistently elicit the same responses from the foundational model.
2. **Simplicity**: The model and prompts should aim for the lowest sufficient performance cost and the simplest effective formulation.

In addition to these primary criteria, various aspects are also important to be defined or evaluated thoroughly:

- **Scope**

Quantifying the sentiment of news articles, involves establishing a numerical scale that accurately reflects varying degrees of sentiment. The choice of range depends on the

granularity needed for the analysis and the specific requirements of the application. For instance, a -1 to +1 scale provides a straightforward interpretation where -1 is fully negative, 0 is neutral, and +1 is fully positive, suitable for general sentiment assessment. For more nuanced analysis, a 0 to 100 scale might be used, offering a finer gradation of sentiment intensity. In this study, the scope of sentiment analysis is defined by the specific objectives of understanding how news impacts stock market movements. Therefore, a -1 to +1 scale has been selected to efficiently capture the overall sentiment polarity of each article, aligning with the analytical needs to correlate these sentiments with stock performance. This scale is effective in delineating clear sentiment thresholds, which simplifies the integration and comparison of sentiment data with stock price fluctuations, thereby facilitating a more straightforward interpretation of the results.

- **Prompt**

Defining a prompt for generative AI involves crafting an input that clearly and concisely communicates the desired task or output to an AI model. An effective prompt not only specifies the nature of the task but also includes enough context to guide the AI in generating accurate and relevant responses. The prompt should be direct and unambiguous to minimize the risk of generating irrelevant or incorrect outputs. In terms of prompt engineering, one-shot prompting involves giving the model a single example to guide its generation, often helping it grasp subtle nuances of the task. Few-shot prompting expands on this by providing several examples, further refining the model's output accuracy [40]. Chain-of-thought prompting, meanwhile, encourages the model to articulate its reasoning step-by-step before arriving at a conclusion, enhancing its ability to tackle complex problems effectively [41]. To ensure simplicity and efficiency, zero-shot prompting is chosen as the initial testing approach; it requires the model to generate responses based solely on the input prompt without prior examples, utilizing minimal tokens.

- **Model**

The selected foundational models are targeted to the GPT series provided by OpenAI. Utilizing OpenAI's models ensures that the project benefits from the latest advancements in AI technology, providing a strong foundation for achieving high-quality, reliable sentiment analytical outcomes. In this research, several models from the GPT series will be tested for their stability and performance, including GPT-4, GPT-4o, GPT-4o Mini, GPT-4 Turbo, and GPT-3.5 Turbo [42]. GPT-4 offers enhanced



language understanding and text generation capabilities, making it highly effective for sophisticated sentiment analysis tasks. GPT-4o offers robust capabilities with advanced reasoning and a deeper understanding of context, suitable for complex sentiment analysis. The GPT-4o Mini, a smaller and more cost-effective variant, is ideal for quicker, less resource-intensive tasks while still maintaining considerable accuracy. GPT-4 Turbo, known for its speed and efficiency, is designed for high-performance applications requiring rapid response without compromising output quality. Lastly, GPT-3.5 Turbo combines the refined capabilities of its predecessors with enhanced processing speeds, making it highly effective for sentiment analysis. Each model will be evaluated to determine its effectiveness in consistently producing accurate sentiment assessments across a range of news contexts, ensuring the chosen model aligns optimally with the primary criteria.

The architecture of GPT models is built on the innovative transformer architecture, fundamentally altering how sequential data is processed. Introduced by Vaswani et al. [43], this architecture abandons traditional recurrent methods for a multi-layered approach centered around the self-attention mechanism. This mechanism assesses the relevance of each word in the input data, independent of their positions. Enhanced by positional encodings, it imparts a sense of sequential order to the model—a crucial feature absent in earlier architectures. Comprising multiple layers of transformer blocks, each with self-attention and feed-forward neural networks, the architecture processes data in parallel, significantly boosting efficiency and effectiveness in complex tasks like natural language understanding and generation. This paradigm not only accelerates the training process but also significantly improves the model's ability to produce contextually coherent and relevant text outputs, establishing GPT as a foundational technology in AI-driven natural language processing.

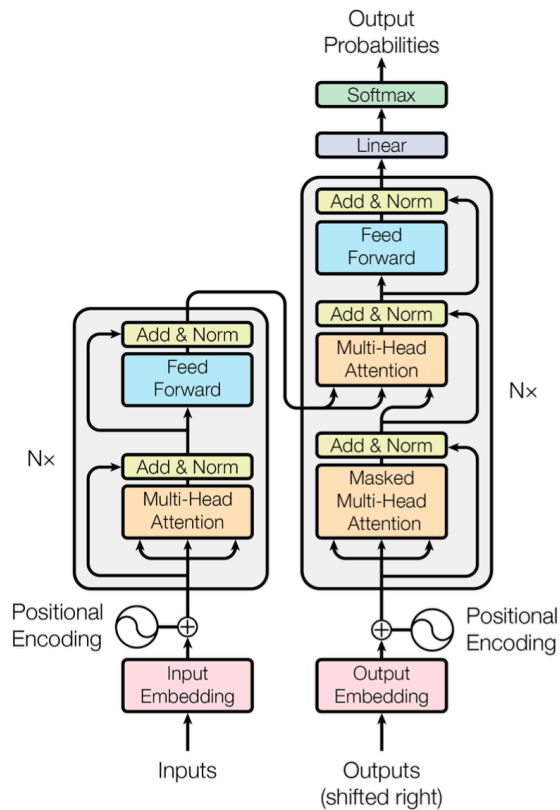


Figure 3.1.2.1 The Transformer Model Architecture [43]

The initial testing prompt use to task the GPT models is shown below. It utilizes zero-shot prompting and clearly defines the scope of the expected result.

*“You are a financial analyst analyzing the news sentiment of {Company Name}.  
{News Article}*

*Review this news article and respond to its sentiment score (-1 to 1, 2 decimal places). For example, respond ``-1.00`` when the article has a highly negative impact on the company, respond ``0.00`` when the article is neutral or irrelevant, respond ``1.00`` when the article has a highly positive impact on the company. You can have any number in between depending on its sentiment. Respond with only 1 number and do not include any other words.”*

The testing samples are randomly selected from the collected data, covering news articles about Public Bank Berhad. The following table demonstrates the sources and the number of testing samples.

Table 3.1.2.1 Testing Samples Summary

| News Site                | Number of Articles |
|--------------------------|--------------------|
| <i>The Star</i>          | 50                 |
| <i>New Straits Times</i> | 50                 |
| <b>Total</b>             | <b>100</b>         |

The testing models include GPT-4, GPT-4o, GPT-4o Mini, GPT-4 Turbo, and GPT-3.5 Turbo, as previously defined. According to [5], the cost ranking for each model from highest to lowest is as follows: GPT-4, GPT-4 Turbo, GPT-4o, GPT-3.5 Turbo, and GPT-4o Mini. For each model, the version selected prioritizes minimal pricing while ensuring it incorporates the most up-to-date training data available. This strategy balances computational efficiency with the benefits of the latest advancements in model training, optimizing both cost-effectiveness and performance. The following table outlines the versions chosen for each model.

Table 3.1.2.2 Selected Versions of GPT Models for Testing [42,44]

| GPT Model            | Selected Version       | Pricing                                                 | Training Data  |
|----------------------|------------------------|---------------------------------------------------------|----------------|
| <i>GPT-4</i>         | gpt-4-0613             | \$30.00 / 1M input tokens<br>\$60.00 / 1M output tokens | Up to Sep 2021 |
| <i>GPT-4 Turbo</i>   | gpt-4-0125-preview     | \$10.00 / 1M input tokens<br>\$30.00 / 1M output tokens | Up to Dec 2023 |
| <i>GPT-4o</i>        | gpt-4o-2024-08-06      | \$2.50 / 1M input tokens<br>\$10.00 / 1M output tokens  | Up to Oct 2023 |
| <i>GPT-3.5 Turbo</i> | gpt-3.5-turbo-0125     | \$0.50 / 1M tokens<br>\$1.50 / 1M tokens                | Up to Sep 2021 |
| <i>GPT-4o Mini</i>   | gpt-4o-mini-2024-07-18 | \$0.150 / 1M input tokens<br>\$0.600 / 1M output token  | Up to Oct 2023 |

The testing phase is divided into three parts: evaluating the generated result format, comparing sentiment trends, and checking for stability. The first step aims to evaluate whether the results can be converted into numerical form and plotted on a graph; this preparation ensures that the sentiment trends can be effectively visualized in the second step. The second step involves visualizing the sentiment data from each model in the same sequence to ensure consistency in presentation and then comparing these figures to assess their accuracy in

generating results in the effective trend. A model will be chosen based on its performance, with a priority given to simplicity. Lastly, the chosen model will be tasked with generating the same sample data using the same prompt twice more to ensure its stability. This comprehensive testing phase evaluates all selected models to determine their suitability for long-term use in sentiment analysis tasks. Below, the figure illustrates the possible outcomes of each testing step and the consequent actions that will be taken.

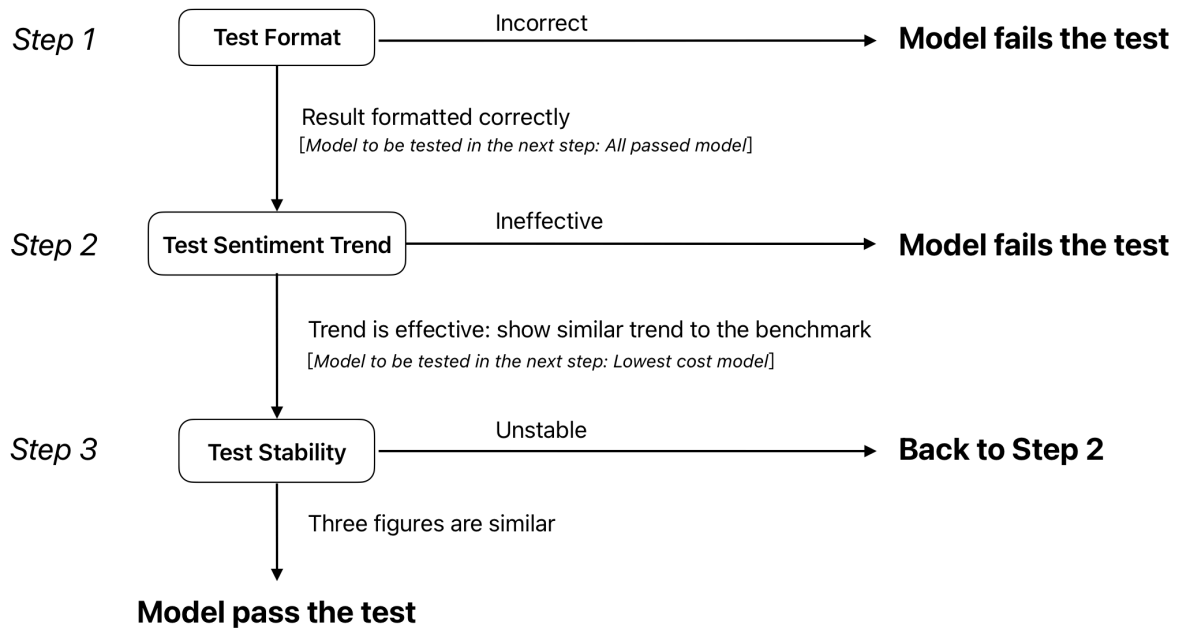


Figure 3.1.2.3 Testing Phase Outcomes and Actions

Each step of the testing phase begins with validated models. All models selected for involvement in this phase are deemed valid in Step 1. Each model that satisfies the condition of “Result formatted correctly” progresses to the Step 2. The criteria for valid models advancing to the next step require that models meet the condition without a quantity limit. If any specific model fails to meet this condition, it will not advance to the subsequent step, as the designation “Incorrect” leads to “Model fails the test.” Upon validation of models in Step 2, these models are then compared using appropriate correlation metrics. In this test, the GPT-4 model serves as the benchmark for sentiment analytical capabilities, given its superior performance. Each model is compared against GPT-4 to assess its effectiveness. The lowest cost model among those validated will be selected as the valid model for Step 3. In this final step, the chosen model is tasked twice with generating responses to the same prompt. Together

with the previously generated results, these three outcomes are compared to check for consistency. Highly similar results are deemed stable, and the model is then considered to have passed the test. However, if the model fails, the process reverts to Step 2, where another valid model will be chosen to undergo Step 3.

Besides these valid steps, several exceptions may occur during the testing phase.

1. **Exception:** No model passes Step 1
  - In this case, the prompt will be enhanced, and the testing phase will be conducted again.
2. **Exception:** GPT-4 is the only model which passes Step 1
  - An enhanced prompt will be used to re-conduct the testing phase, and the overall cost of using the valid model will be calculated to compare with the results obtained using the previous prompt with the GPT-4 model.
3. **Exception:** No model shows a similar trend to the benchmark in Step 2
  - Similar to the Exception 2, a new prompt will be used in the test, and a cost comparison will be conducted.
4. **Exception:** Model fails Step 3 and only GPT-4 remains valid in Step 2
  - This situation is identical to Exception 3, and the actions listed in Exception 2 will be taken.

The criteria to evaluate the results of each step are defined as follows:

1. **Step 1:**

Each model's response result must be displayed in numerical format with two decimal places. The scores must range between -1.00 and 1.00, where -1.00 is the minimum and 1.00 is the maximum score allowed for any response result.
2. **Step 2:**

Spearman's rank correlation will be used to compare the performance of the models to the benchmark. This method is ideal for comparing the sentiment trends produced by different GPT models as it focuses on the order of the values rather than their specific magnitudes. It assesses whether increases or decreases in sentiment scores from one model correspond to similar movements in scores from another model, regardless of the exact numerical values. The formula for Spearman's Rank Correlation Coefficient is shown below:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (1)$$

where,

- $\rho$  represents the Spearman's rank correlation coefficient.
- $d_i$  is the difference in ranks between the corresponding values of the two variables
- $n$  is the number of observations.

A Spearman's rho value ranging from 0.80 to 1.00 is considered indicative of effective results and sufficient model performance.

### 3. Step 3:

The similarity of three generated results from the same model will be evaluated using Spearman's rank correlation, as outlined in the Step 2.

Spearman's rank correlation plays a crucial role in evaluating the monotonic relationships between variables. In this research, Spearman's rank correlation is calculated using Python. The `scipy.stats` library, specifically the `spearmanr` function, is employed for this purpose. This function computes the Spearman correlation coefficient, which measures the strength and direction of association between two ranked variables. By passing the relevant data arrays to `spearmanr`, the research obtains both the correlation coefficient and the p-value, which helps in evaluating the statistical significance of the correlation observed.

After confirming a valid model, this model will be tasked with generating sentiment scores for the entire dataset. The primary preprocessing step before integrating these sentiment scores with stock data involves converting all sentiment outputs, which are initially generated in string format, to numeric values. This conversion is essential to standardize the format and ensure that the sentiment data is compatible with quantitative analysis tools.

### 3.4 Data Integration

Following the preprocessing of stock and news sentiment data, the next critical phase in this research involves integrating these two distinct data streams. This section discusses how the news sentiment data aligned with stock data indexed by timeline.

The primary step in the data integration process involves the consolidation of sentiment scores from various news sources. For each day, sentiment scores provided by different sources are aggregated to compute a mean sentiment value. This method ensures that the data represents

a comprehensive daily media sentiment, which is crucial for understanding the potential impacts on stock market movements.

Once the daily sentiment scores are consolidated, they are aligned with the corresponding stock market data. This alignment is critical as it ensures that the sentiment data corresponds accurately to the stock data by dates, facilitating a direct comparison between media sentiment and stock market responses. The aligned dataset is crucial for the subsequent analysis phase, where relationships between sentiment and stock movements are explored.

Before proceeding to analytical stages, the integrated dataset undergoes a thorough verification and cleaning process. This step ensures the dataset is free from inconsistencies or missing data points, which could affect the reliability of subsequent analyses. This process checks for alignment issues between sentiment scores and stock data, guaranteeing that the dataset is robust and ready for in-depth analysis.

The last step in data integration involves preparing the dataset for analysis. This preparation includes the final adjustments to the data format, ensuring that all data points are consistently formatted and properly indexed. This preparation sets the stage for seamless application of analytical tools and techniques in the analysis section of the study, where the relationships between news sentiment and stock performance are systematically explored.

### **3.5 Data Analysis**

This phase involves several time series analysis techniques and tests to prepare the data for application to the ARDL model.

#### **3.5.1 Augmented Dickey-Fuller**

Firstly, the Augmented Dickey-Fuller (ADF) test will be adopted. The ADF test is a widely used statistical test that checks for the presence of unit roots in a time series dataset, which is a way to test for non-stationarity. A unit root indicates that the time series is influenced by a stochastic trend, meaning it can wander away from its mean over time and not return to it, which makes predictive modelling challenging. The presence of a unit root suggests that the time series might need to be differenced to make it stationary, which is an important step in preparing data for time series forecasting models.

The typical equation form of the ADF test is [45]:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^p \phi_i \Delta y_{t-i} + \epsilon_t \quad (2)$$

where,

- $\Delta y_t$  is the first difference of the series  $y_t$ .
- $\alpha$  is a constant term.
- $\beta t$  represents a time trend.
- $\gamma$  is the coefficient on the lagged level of the series,  $y_{t-1}$ , which is the key parameter for the hypothesis test.
- $\phi_i$  are the coefficients for the lagged differences of the series.
- $\epsilon_t$  is the error term.
- $p$  is the number of lagged first differences included in the regression.

A hypothesis test will be used for testing the unit root existence. The hypotheses for the ADF test can be stated as follows:

- **Null Hypothesis**  $H_0: \gamma = 0$ , The series has a unit root.
- **Alternative Hypothesis**  $H_1: \gamma < 0$ , The series has no unit root.

P-value is the probability that the observed statistic would be at least as extreme as the one observed if the null hypothesis were true. In the case of the ADF test, the null hypothesis states that the time series has a unit root.

Decision Rule:

- If the **p-value is less than or equal to a chosen significance level** (commonly set at 0.05, or 5%), then the null hypothesis is rejected. This means there is sufficient statistical evidence to conclude that the series does not have a unit root and is stationary.
- If the **p-value is greater than the significance level**, the null hypothesis cannot be rejected. This suggests that the series likely has a unit root, indicating it is non-stationary and may require differencing to make it stationary.

Stationarity is crucial for time series models. When trends and seasonality are present in a dataset, they can distort the true relationships between variables, leading to misleading conclusions and unreliable forecasts. To address this, data can be differenced to remove these elements and thereby achieve stationarity, stabilizing the mean and variance over time. This



process enhances the predictive power of time series models by focusing analysis on the stochastic processes rather than deterministic trends

In this study, Python's statsmodels library is utilised in conducting the ADF test to assess the stationarity of time series data. Utilizing the adfuller function within this library allows for an efficient evaluation of whether a time series is stationary. This function provides the test statistic, p-value, and critical values for various confidence levels, facilitating a straightforward determination of the presence of a unit root. Employing Python and its statistical tools enables precise and automated analysis, essential for ensuring the data's suitability for subsequent forecasting and econometric modelling.

### 3.5.2 Lags

To determine the appropriate lag length for the ARDL model, the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) will be employed. The Autocorrelation Function (ACF) measures the correlation between a time series and its lagged versions. It is used to identify the internal dependency within the data, which can suggest the presence of time-based patterns such as seasonality or autoregressive behaviour.

The equation of the ACF at lag  $k$  is given by:

$$\rho_k = \frac{\sum_{t=k+1}^T (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^T (y_t - \bar{y})^2} \quad (3)$$

where:

- $\rho_k$  is the autocorrelation at lag  $k$ .
- $y_t$  is the value of the series at time  $t$ .
- $\bar{y}$  is the mean of the series.
- $T$  is the total number of observations.

ACF is primarily used to determine the order of a Moving Average (MA) component by examining where the autocorrelations essentially become insignificant (drop to zero or within a confidence band).

On the other hand, the Partial Autocorrelation Function (PACF) measures the correlation between the time series and its lag, controlling for the values of the time series at all shorter lags. It isolates the effect of each lag, which helps in identifying the order of an Autoregressive (AR) model.

The equation of PACF at lag  $k$  can be mathematically derived from a series of linear equations involving the autocorrelations of the series:

$$\alpha_{kk} = \phi_k - \sum_{j=1}^{k-1} \alpha_{k-1,j} \phi_{k-j} \quad (4)$$

where:

- $\alpha_{kk}$  is the partial autocorrelation at lag  $k$ .
- $\phi_k$  and  $\phi_{k-j}$  are the coefficients from the autoregressive model of the time series.
- $\alpha_{k-1,j}$  are the coefficients from the previous order PACF.

PACF is used to determine the number of lags to be included in an AR model, as it shows the extent of autocorrelation at each lag after removing the effect of these shorter lags.

Both ACF and PACF are equipped with hypothesis tests for determining the statistical significance of each lag. These tests typically involve checking if the autocorrelations and partial autocorrelations are significantly different from zero at a certain confidence level.

- **Null Hypothesis** ( $H_0$ ): There is no autocorrelation at lag  $k$ .
- **Alternative Hypothesis** ( $H_1$ ): There is significant autocorrelation at lag  $k$ .

The p-values from these tests help decide whether to include specific lags in your models. Critical values for decision making usually come from the chi-square or standard normal distributions, depending on the sample size and specific method used.

This study conduct ACF and PACF analyses using Python's statsmodels library to identify appropriate lag values. The plot\_acf and plot\_pacf functions from statsmodels.graphics.tsaplots are employed to visually inspect the correlation between the series and its lags. Additionally, scipy.stats.norm is used to calculate the statistical significance of the autocorrelations, further guiding the optimal model specification.

## 3.6 Prediction

### 3.6.1 ARDL

The Autoregressive Distributed Lag approach is employed to explore the dynamic relationship between stock prices and sentiment indices, providing a robust methodology for analyzing the short-term and long-term interactions within time series data. The ARDL model framework is

advantageous as it accommodates variables of different integration orders, making it suitable for datasets where variables are not necessarily stationary.

The general form of the ARDL model is expressed as:

$$Y_t = \alpha + \sum_{i=1}^p \beta_i Y_{t-i} + \sum_{j=0}^q \gamma_j X_{t-j} + \epsilon_t \quad (5)$$

where:

- $Y_t$  is the dependent variable, stock prices in this context.
- $X_t$  represents the independent variable(s), sentiment indices in this context.
- $p$  and  $q$  denote the number of lags used for the dependent and independent variables, respectively.
- $\alpha$ ,  $\beta_i$ , and  $\gamma_j$  are coefficients to be estimated.
- $\epsilon_t$  is the error term.

The ARDL model's application to this research involves quantifying how past values of both stock prices and sentiment indices influence current stock prices to forecast stock short term trend. This dual consideration helps in understanding both the immediate and delayed effects of sentiment changes on stock prices.

1. **Model Specification:** Define the number of lags based on criteria such as the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) to ensure optimal model fit.
2. **Estimation:** Fit the ARDL model using historical data to estimate the relationship dynamics.
3. **Diagnostic Checks:** Conduct tests for autocorrelation, heteroscedasticity, and model stability to ensure the robustness of the model.
4. **Forecasting:** Utilize the model to forecast future stock movements based on current and historical sentiment data.

To assess the accuracy and efficacy of the ARDL model's predictions, the Root Mean Square Error (RMSE) will be employed as a key performance metric. RMSE is a widely used measure of the differences between values predicted by a model and the values actually observed. It is particularly useful in quantitative finance as it provides a clear indication of the model's prediction error in terms of the units of the variable of interest.

The RMSE is defined as the square root of the average squared differences between the predicted values ( $\hat{Y}_t$ ) and the observed values ( $Y_t$ ):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n (\hat{Y}_t - Y_t)^2} \quad (6)$$

where:

- $n$  is the number of observations.
- $\hat{Y}_t$  is the predicted value at time  $t$ .
- $Y_t$  is the actual value at time  $t$ .

### 3.6.2 Source Influence

The ARDL accuracy, quantified through the Root Mean Squared Error (RMSE), will serve as the primary metric to evaluate the influences of different news sources on the stock market. RMSE provides a clear measure of how closely the predicted stock prices align with the actual market values, thereby reflecting the effectiveness of incorporating sentiment data from various news outlets into the ARDL model. Lower RMSE values indicate more accurate predictions, suggesting that the sentiment analysis from those news sources more effectively captures the market-moving information that influences stock prices.

## 3.7 System Requirement

### 3.7.1 Hardware Component

A computer with sufficient processing power and memory to handle the data processing and analysis tasks is essential for the proposed method. Table 3.1 shows the computer specification used in the project.

Table 3.7.1.1 Specifications of laptop

| Description      | Specifications           |
|------------------|--------------------------|
| Model            | MacBook Air              |
| Processor        | Apple M1 chip 8-core CPU |
| Operating System | macOS Sonoma             |
| Memory           | 16GB Unified Memory      |
| Storage          | 1TB SSD                  |

### 3.7.2 Software Component

A list of the key software components utilized in this research to perform various tasks ranging from data analysis to sentiment analysis and data visualization is shown below:

#### 1. **Python:**

- Primary programming language used for scripting, data manipulation, and running machine learning algorithms.
- Libraries such as Pandas for data manipulation, NumPy for numerical data operations, Matplotlib and Seaborn for data visualization, and Statsmodel for time series analysis.

#### 2. **Jupyter Notebook:**

- An open-source web application that allows the creation and sharing of documents that contain live code, equations, visualizations, and narrative text.
- Used for coding, visualizing, and presenting the data analysis and results in a comprehensible format.

#### 3. **Elasticsearch:**

- A distributed, RESTful search and analytics engine capable of solving a growing number of use cases.
- Utilized for efficiently storing, searching, and analyzing large volumes of textual data quickly and in near real-time.

#### 4. **Google Cloud API:**

- APIs provided by Google Cloud for different purposes including the Google Cloud Natural Language API which was utilized for an additional layer of sentiment analysis.
- These APIs are used for enhancing the capabilities of the models by integrating cloud-based machine learning and data analysis services.

#### 5. **GPT-4 Turbo** (part of OpenAI's models):

- Utilized for performing state-of-the-art sentiment analysis as part of the data preprocessing and analysis process.
- Integrated within the Python environment to analyze financial news and derive sentiment scores.

#### 6. **YFinance:**

- A popular Python library used to fetch historical market data from Yahoo Finance.
- Employed to obtain stock price data that is essential for correlating with the sentiment analysis results.

### 3.8 Timeline



Figure 3.8.1 Timeline of FYP1

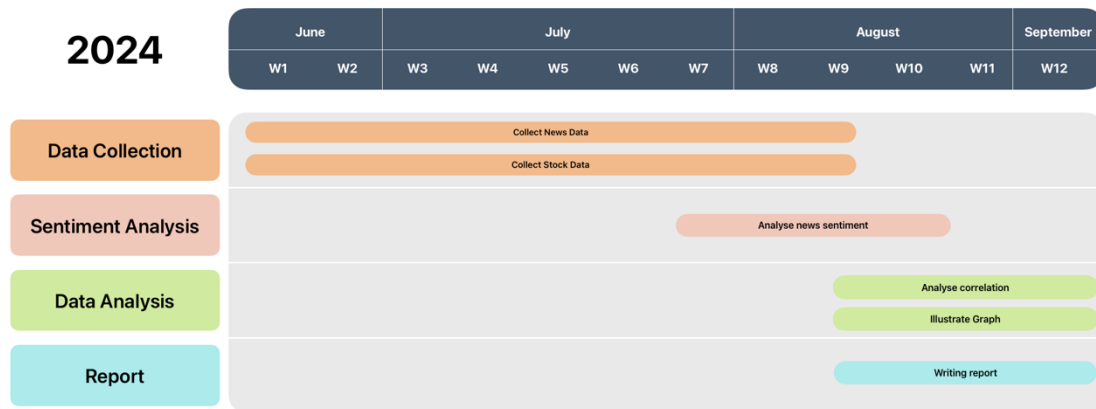


Figure 3.8.2 Timeline of FYP2

# Chapter 4

## System Implementation

### 4.1 Yahoo Finance

Figure 4.1 shows an example Python script for retrieving stock data via the yfinance library. The yfinance library provides a simple and efficient way to access financial data available on Yahoo Finance. In the Python script, the stock symbol and date range are required to fetch historical market data for a specific stock. The Ticker object is used to retrieve data based on these parameters. Besides, the period is consistently set to '1d' (one day), ensuring daily data granularity.

```
import yfinance as yf

def get_stock(symbol, start_date, end_date, period='1d'):
    stock = yf.Ticker(symbol)
    stock_data = stock.history(period=period, start=start_date, end=end_date)

    return stock_data
```

Figure 4.1.1 Python Script of Stock Data Retrieval

### 4.2 Web Scraping

Figure 4.2.1 presents a Python code snippet illustrating how to utilize the Google Custom Search JSON API for performing search queries programmatically. The script starts by importing the build function from the googleapiclient library, which is used to create a service object that interacts with Google's APIs. A service object for the Custom Search API is then initialized with the build function, specifying customsearch for the API type, v1 for the API version, and providing a developer API key. This service object is then used to make a search query by calling the cse() method followed by list(), where the query parameters q (the search keyword) and cx (the search engine ID) are specified. The search operation is executed with the execute() method, which sends the request to the API and returns the search results. This

allows for fetching specific, customized search data programmatically, which can be useful for gathering targeted news sites links across the web.

```
from googleapiclient.discovery import build

service = build("customsearch", "v1", developerKey=api_key)
res = service.cse().list(
    q="Search keyword",
    cx=cx
).execute()
```

Figure 4.2.1 Python Script of Extract URLs

Figure 4.2.2 depicts an example Python script that utilizes the JSON Data Extraction approach to scrape headline information from news articles. It begins by importing necessary libraries: BeautifulSoup for parsing HTML content, requests to fetch web pages, and json for handling JSON data. The process starts with defining the URL of the news article to be scraped. The script uses the requests library to retrieve the web page, and the HTML content of this page is then parsed into a BeautifulSoup object. This object searches for a specific script tag that contains JSON formatted data, from which the headline of the news can be extracted. This method provides a precise and structured approach to extract news information, demonstrating a powerful way to gather specific data from news websites for this study.

```
from bs4 import BeautifulSoup
import requests
import json

url = "https://www.nst.com.my/business/2023/06/924197/bursa-malay"

page = requests.get(url)
soup = BeautifulSoup(page.content, 'html.parser')
script_tag = soup.find('script', {'type': 'application/ld+json'})
script_content = script_tag.string

json_data = json.loads(script_content)
headline = json_data.get('headline', 'N/A')
headline = html.unescape(headline)
```

Figure 4.2.2 Python Script of JSON Data Extraction



Figure 4.2.3 demonstrates an example of the HTML Meta Extraction approach to scrape and format the publication date from a web page. After importing necessary modules, the script fetches a web page and then searches for a meta tag specifically containing the defined property. The content of this tag, which is the publication date, is retrieved and stripped of any leading or trailing whitespace. As the date format is required to fulfil the format needed in this study, the date is then parsed into a datetime object and subsequently formatted into a standard format. This process efficiently extracts and standardizes the publication date from the article's metadata for further use or analysis.

```
from bs4 import BeautifulSoup
import requests
from datetime import datetime, timedelta, timezone
import pytz

page = requests.get(url)
soup = BeautifulSoup(page.content, 'html.parser')

date_tag = soup.find('meta', {'property': 'article:published_time'})
date_published = date_tag.get('content', 'N/A').strip()
date_published = datetime.strptime(date_published, '%d/%m/%Y %I:%M %p')
date_published = date_published.strftime('%Y-%m-%d %H:%M:%S')
```

Figure 4.2.3 Python Script of HTML Meta Extraction

The following Python script employs the HTML Content Extraction approach to extract the full text of an article from a webpage. After importing the necessary libraries, the script fetches the page content and parses it into a BeautifulSoup object. It specifically searches for a “div” element with an itemprop attribute set to “articleBody”, which is commonly used to demarcate the main body of an article on web pages. Upon locating this “div” element, the script extracts its text content, effectively stripping away any HTML tags to leave only the plain text of the article. This method is streamlined and effective for pages that use standardized HTML5 semantic tags to structure their content.

```

from bs4 import BeautifulSoup
import requests

page = requests.get(url)
soup = BeautifulSoup(page.content, 'html.parser')

article_tag = soup.find('div', {'itemprop': 'articleBody'})
full_article = article_tag.text

```

Figure 4.2.4 Python Script of HTML Content Extraction

### 4.3 Elasticsearch

Figure 4.3.1 provides Python code snippet utilizes the Elasticsearch library to establish a connection to an Elasticsearch server and insert documents into a specified index. First, an instance of the Elasticsearch class is created, where the connection parameters such as the server host, API key for authentication, and a certificate authority (CA) certificates file for SSL verification are specified. Following the establishment of this connection, the code iterates through a collection of documents stored in the variable “overall\_info”. Each document is then indexed into Elasticsearch using the index method, where “index\_name” specifies the name of the target index and “doc” is the content of the document being inserted. This process is typically used to populate an Elasticsearch index with data that can later be queried and analyzed.

```

from elasticsearch import Elasticsearch

es = Elasticsearch(
    host,
    api_key=api_key,
    ca_certs=ca_certs
)

# insert document
for doc in overall_info:
    es.index(index=index_name, body=doc)

```

Figure 4.3.1 Python Script of Inserting Document to Elasticsearch

### 4.4 Generative AI

The example code snippet for tasking GPT models, as depicted in the following figure, demonstrates how to efficiently use the OpenAI API for sentiment analysis. The script initiates by importing the OpenAI library and initializing the client with an API key. It constructs a

conversation sequence structured with specifically labelled messages. A “system” message first sets the context, specifying the task of analysing news sentiment for a designated company. This is followed by a “user” message that presents the news article and directs the AI to evaluate and provide a sentiment score, strictly in numerical format. The model’s response, containing the sentiment score, can then be retrieved from the content of the response object.

```
from openai import OpenAI

ask = OpenAI(
    api_key = openai_key
)

response = ask.chat.completions.create(
    model=model,
    messages=[
        {"role": "system", "content": system_content.format(company)},
        {"role": "user", "content": user_content.format(article)}
    ]
)

result = response.choices[0].message.content
```

Figure 4.4.1 Python Script of Tasking Model

## 4.5 Predictive Model

This code segment [Figure 4.5.1] fits an ARDL model, estimates the necessary parameters, and forecasts future stock prices based on the sentiment data.

```
from statsmodels.tsa.api import ARDL

forecast_steps = int(len(data) * 0.2)
model = ARDL(data[:-forecast_steps]['staStock'], 1, data[:-forecast_steps][['staSent']], order={'staSent': lag})
result = model.fit()
forecast = result.forecast(steps=forecast_steps, exog=data[-forecast_steps:][['staSent']])
```

Figure 4.5.1 Python Code Snippet of ARDL Implementation

Figure 4.5.2 illustrates the code used to preprocess non-stationarity stock data and potential seasonal patterns repeating every two periods. The logarithmic transformation applied to the data helps reduce skewness and stabilize variance, making the series more symmetrical and manageable for modeling. Subsequently, seasonal differencing is employed to eliminate any underlying seasonal effects that recur every two periods. This preprocessing step is crucial for

ensuring the data meets the stationarity assumptions required by many statistical modeling techniques.

```
data['staStock'] = np.log10(data['stock']) - np.log10(data['stock']).shift(2)
```

Figure 4.5.2 Python Code Snippet of Seasonal Differencing

# Chapter 5

## Result and Discussion

By following the proposed research model, several experiments are conducted, and outcomes will be discussed in this section.

### 5.1 Stock Data

This study undertakes an empirical investigation of stock price movements, specifically focusing on Public Bank and CIMB, which are prominent entities in the Malaysian financial sector. The dataset utilized encompasses the daily stock prices of both entities for the years 2022 to 2023, sourced from the Yahoo Finance API. This source provides comprehensive and reliable financial data, critical for the robust analysis intended in this research. Figure 5.1.1 shows a five sample of collected stock data.

|            | Open     | High     | Low      | Close    | Volume     | Dividends | Stock Splits |
|------------|----------|----------|----------|----------|------------|-----------|--------------|
| date       |          |          |          |          |            |           |              |
| 2022-01-03 | 3.682777 | 3.709335 | 3.665071 | 3.673924 | 4077500.0  | 0.0       | 0.0          |
| 2022-01-04 | 3.673924 | 3.682776 | 3.638512 | 3.656218 | 13730800.0 | 0.0       | 0.0          |
| 2022-01-05 | 3.656218 | 3.691630 | 3.647365 | 3.691630 | 12865000.0 | 0.0       | 0.0          |
| 2022-01-06 | 3.682777 | 3.691630 | 3.638513 | 3.647365 | 13660500.0 | 0.0       | 0.0          |
| 2022-01-07 | 3.656218 | 3.682777 | 3.647365 | 3.682777 | 10787000.0 | 0.0       | 0.0          |

Figure 5.1.1 Sample Five Rows of Stock Data

Figure 5.1.2 displays the daily stock prices of Public Bank Berhad from 2022 to 2023, using candlestick formatting to offer a detailed view of price movements. A candlestick chart is a type of financial chart used to describe price movements of a security, derivative, or currency. Each “candle” in the chart provides four key pieces of information: the opening price, the closing price, the high price, and the low price over a certain period. If the closing price is higher than the opening price, the candlestick is often coloured green, representing a price increase. Conversely, if the closing price is lower, the candlestick is coloured red, indicating a

price decrease. In this chart, fluctuations are evident as the colours shift between green and red, signalling periods of volatility influenced by investor sentiments and external factors. The sequences of red candlesticks typically indicate a downward trend in stock prices, while sequences of green suggest upward movements. For instance, there is a notable increase in green candlesticks before April 2023, suggesting a sharp rise in stock prices during this period.

The chart shows several periods of volatility, as evidenced by the significant rises and falls in stock prices. For example, between April 2022 and July 2022, the stock price saw a sharp increase, reaching a peak around July before experiencing a decline. This pattern suggests a volatile period where external factors or market sentiments may have driven prices up, followed by a correction. After the initial peak in July 2022, there is a noticeable downward trend that continues until around January 2023, where the stock reaches its lowest point on the chart. This is followed by a recovery phase where the stock price begins to climb again, suggesting a rebound possibly due to positive market news or financial performance.

Throughout 2023, particularly from February to October, the stock price shows resistance at higher levels around RM4.20, where it struggles to break through, and support levels around RM3.80, where it seems to bounce back upon touching these lower price points. These levels could be critical for traders looking for entry and exit points. The frequent alternation between red and green candlesticks, especially visible from January to July 2023, indicates a highly reactive market sentiment during this period. The market's reaction to news or global economic conditions might have been mixed, leading to these rapid changes.

Public Bank Stock Prices 2022 to 2023



Figure 5.1.2 Candlestick Chart of Public Bank Stock Prices 2022 to 2023

Figure 5.1.3 illustrates the daily stock prices of CIMB from 2022 to 2023, presented through a candlestick chart. This chart highlights a pattern of volatility similar to that observed in the Public Bank chart. The stock prices of CIMB show significant fluctuations within this period, reflecting investor responses to varying market conditions and external economic factors. For instance, a noticeable rise in stock prices can be observed starting around mid-2022, peaking towards the end of July 2022. This period is marked by a predominance of green candlesticks, suggesting a bullish market sentiment driving the prices upward.

However, following this peak, there is a clear trend of decreasing prices, with a succession of red candlesticks dominating the chart until early 2023, indicating a bearish phase. This decline could be attributed to market corrections or potentially adverse news affecting investor sentiment. Subsequently, the chart shows a recovery phase beginning around February 2023, where prices gradually begin to increase again, showcasing resilience or positive market influences at play.

Throughout the year 2023, the stock price oscillates between support levels at approximately RM4.00 and resistance levels near RM4.35. These price points represent key thresholds that the stock struggles to break past on several occasions, reflecting the tug-of-war between buyers and sellers influenced by ongoing market dynamics and news events. The alternating pattern of green and red candlesticks, particularly visible from mid-2023 onwards, underscores the ongoing volatility and the market's sensitivity to news and global economic conditions during this period.

CIMB Stock Prices 2022 to 2023



Figure 5.1.3 Candlestick Chart of CIMB Stock Prices 2022 to 2023

Among various stock indicators, stock opening prices are selected as the focal point for this study's prediction analysis. The opening price of a stock is considered a significant indicator as it reflects the market sentiment and information available to investors immediately before the market opens. By focusing on the opening prices, this research aims to capture the initial reactions of the market to overnight news and events, which are pivotal in setting the tone for the trading day.

## **5.2 News Data**

News data is collected by web scraping techniques as outlined in Chapter 3. A variety of news sources are selected to extract news information. These include The Star, New Straits Times, Malay Mail, Free Malaysia Today, Bernama, The Edge Malaysia, Yahoo News, The Sun, Business Today, and Sinar Daily. The first step in the web scraping process involves collecting possible URLs using the Google Custom Search JSON API. The keywords used for searches in this API include three components that need to be defined.

Table 5.2.1 outlines the list of the URLs for various news sources utilized in the data collection process. The site location is the first component in the Google search keyword. Using the URL to locate news sources is crucial because it ensures the accuracy and reliability of the data collected. URLs represent the unique address of each news outlet on the internet, providing a direct path to the specific content needed without the interference of third-party aggregators or unrelated content. This method enhances the precision of the scraping process, ensuring that the collected information is directly from the intended source.



Table 5.2.1 News Sources Directory

| <b>News Source</b>  | <b>URL</b>                     |
|---------------------|--------------------------------|
| The Star            | <i>thestar.com.my</i>          |
| New Straits Times   | <i>nst.com.my</i>              |
| Yahoo News          | <i>malaysia.news.yahoo.com</i> |
| Malay Mail          | <i>malaymail.com</i>           |
| Free Malaysia Today | <i>freemalaysiatoday.com/</i>  |
| Bernama             | <i>bernama.com/en/</i>         |
| The Edge Malaysia   | <i>theedgemalaysia.com</i>     |
| The Sun             | <i>thesun.my</i>               |
| Business Today      | <i>businesstoday.com.my</i>    |
| Sinar Daily         | <i>www.sinardaily.my</i>       |

The second component is the keyword that defines the desired result of the Google search. This is strictly related to the stock or company to ensure the information's relevancy to the stock market. "Public Bank Berhad" and "Commerce International Merchant Bankers Berhad" are the keywords used to focus the search results on specific financial information, company news, and stock performance updates for these entities. By including such targeted keywords, the search is refined to yield content that directly impacts stock market analysis and investor decision-making.

The third keyword is the date range. As the search incorporates this temporal parameter, it becomes possible to strategically target the retrieval of news within specified periods. Utilizing a smaller time range, such as three months, typically yields more results than a single query over a longer period, such as six months. This is because shorter time frames help avoid overwhelming the search algorithm with too broad a query, which can miss more recent or specific articles due to query limitations. However, conducting searches over shorter intervals comes at a higher cost, both in terms of processing time and resource utilization. Therefore, for this research, a six-month frequency was chosen to balance the depth and breadth of data collection with cost-effectiveness.

With the search keywords defined in Chapter 3 and incorporating these three components, URLs are systematically extracted from the API. Following the extraction of URLs, the scraping process and data management are conducted. Figure 5.2.1 shows the sample five rows

of news data, showcasing the structure and type of information available in the dataset. Meanwhile, the number of articles produced by each news source is shown in Figure 5.4.1.

|   | title                                             | source            | url                                               | published_date | article                                           |
|---|---------------------------------------------------|-------------------|---------------------------------------------------|----------------|---------------------------------------------------|
| 0 | Public Bank launches PB Golden Fortune campaign   | The Sun           | https://thesun.my/home-news/public-bank-launch... | 2022-01-04     | PETALING JAYA: In conjunction with the forthco... |
| 1 | 10 stocks brokers say to buy in 2022              | The Edge Malaysia | https://theedgemalaysia.com/article/top-10-sto... | 2022-01-05     | KUALA LUMPUR (Jan 5): After a tumultuous 2020 ... |
| 2 | Avaa Vanja legally prevents husband from takin... | Yahoo News        | https://malaysia.news.yahoo.com/avaa-vanja-leg... | 2022-01-05     | 5 Jan – Singer-actress Avaa Vanja was revealed... |
| 3 | There and back again: What the Omicron variant... | Yahoo News        | https://malaysia.news.yahoo.com/back-again-omi... | 2022-01-06     | KUALA LUMPUR, Jan 6 – The Omicron variant of t... |
| 4 | Kota Kinabalu MP: Lifting timber export ban an... | Yahoo News        | https://malaysia.news.yahoo.com/kota-kinabalu-... | 2022-01-07     | KOTA KINABALU, Jan 7 – A Sabah DAP lawmaker ha... |

Figure 5.2.1 Sample Five Rows of News Data

### 5.3 Model Selection Test

News sentiment analysis will be conducted using a foundational model that needs to be evaluated by a model selection test. Below, the results of the model selection testing phase are shown.

#### 5.3.1 Step 1: Format Testing

Table 5.3.1.1 details the results of format accuracy tests conducted on various GPT models, assessing their capability to generate outputs in a predefined numerical format. Each model was tested for its ability to produce responses that adhere to the required format specifications. The results confirm that all models—GPT-4, GPT-4 Turbo, GPT-4o, GPT-3.5 Turbo, and GPT-4o Mini—successfully met the criteria for numeric response formatting. The validated models, which demonstrated consistent accuracy in format adherence, will advance to the next phase of testing.

Table 5.3.1.1 Model Testing Result: Format

|                      | Numeric response | Max Response | Min Response |
|----------------------|------------------|--------------|--------------|
| <i>GPT-4</i>         | ✓                | 1.00         | -0.80        |
| <i>GPT-4 Turbo</i>   | ✓                | 0.90         | -0.80        |
| <i>GPT-4o</i>        | ✓                | 1.00         | -0.70        |
| <i>GPT-3.5 Turbo</i> | ✓                | 1.00         | -1.00        |
| <i>GPT-4o Mini</i>   | ✓                | 1.00         | -1.00        |

### 5.3.2 Step 2: Sentiment Trend Testing

Table 5.3.2.1 presents the Spearman’s rank correlation results for each model compared against the benchmark model GPT-4. The correlations, as indicated by Spearman’s rho, show how closely the sentiment analysis trends of the models align with those of GPT-4. Specifically, GPT-4 Turbo demonstrates the highest correlation with a rho of 0.8485, indicating a very strong alignment in sentiment trends and suggesting it is the most reliable model in mirroring the benchmark’s performance. GPT-4o follows with a rho of 0.793435, showing substantial but slightly weaker alignment compared to GPT-4 Turbo. GPT-3.5 Turbo and GPT-4o Mini exhibit lower correlations of 0.593245 and 0.704623, respectively, indicating less consistency with the benchmark. Among these, only GPT-4 Turbo meets the effective range of 0.80 to 1.00 for rho, qualifying it as the only valid model to progress to the next step. The extremely low p-values across all models underscore the reliability of the correlation results, confirming that the observed correlations are not due to random chance.

Table 5.3.2.1 Model Testing Result: Trends Comparison

| <b>Model</b>         | <b>Spearman’s Rho</b> | <b>P-value</b>         |
|----------------------|-----------------------|------------------------|
| <i>GPT-4</i>         | 1                     | 0                      |
| <i>GPT-4 Turbo</i>   | 0.848500              | $7.75 \times 10^{-29}$ |
| <i>GPT-4o</i>        | 0.793435              | $7.44 \times 10^{-23}$ |
| <i>GPT-3.5 Turbo</i> | 0.593245              | $7.83 \times 10^{-11}$ |
| <i>GPT-4o Mini</i>   | 0.704623              | $2.83 \times 10^{-16}$ |

Below, the comparison of visualized sentiment trends is presented. As verified by the Spearman’s Rank Correlation, GPT-4 Turbo displays a highly similar sentiment trend to the benchmark GPT-4, showcasing its reliability in sentiment analysis. The GPT-4o model also illustrates a similar movement, though with slightly less alignment compared to GPT-4 Turbo. However, the remaining models, GPT-3.5 Turbo and GPT-4o Mini, exhibit more divergent patterns, highlighting inconsistencies and reduced correlation with the benchmark, which could impact their suitability for precise sentiment analysis tasks. These visual results align with the numerical correlation values previously discussed.

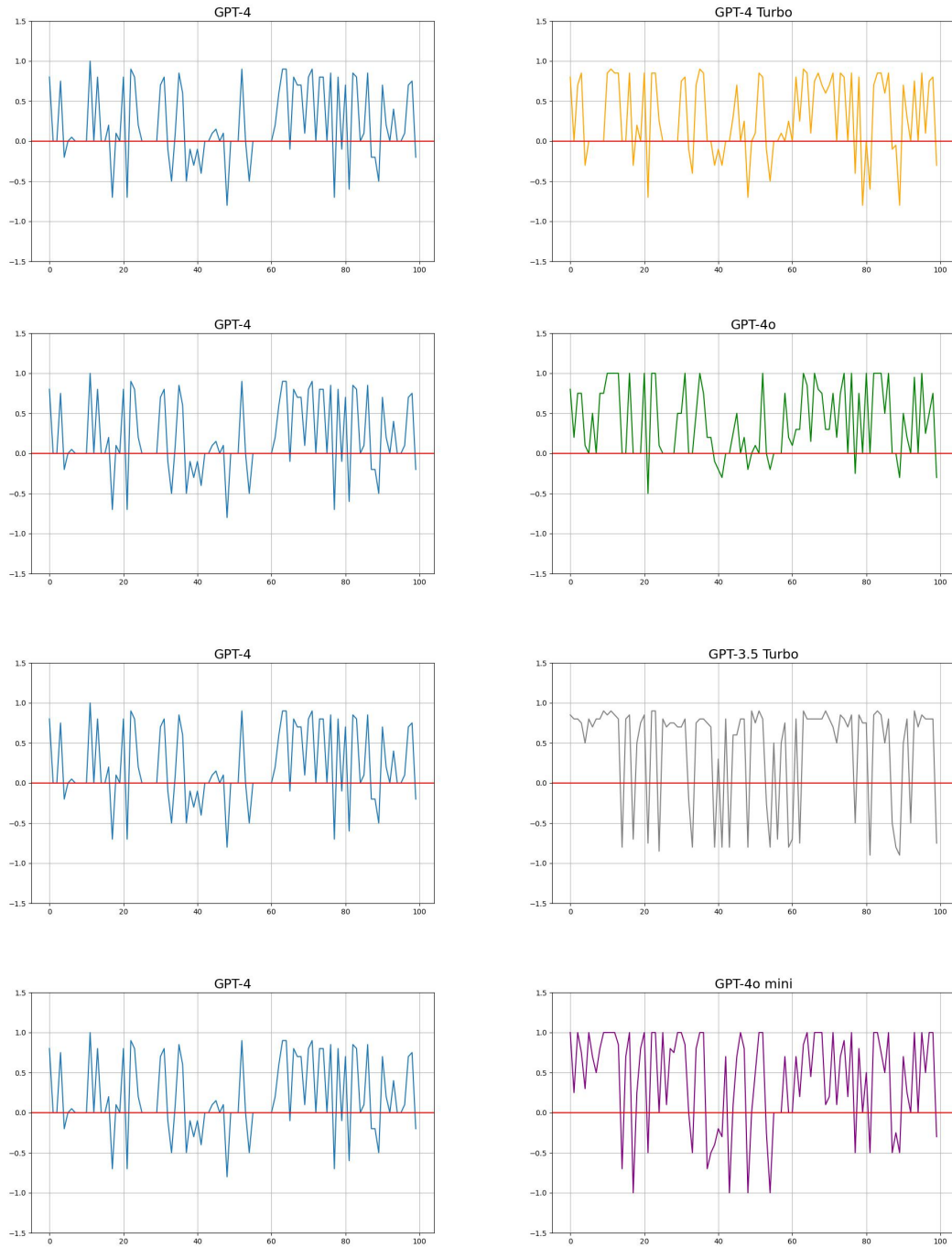


Figure 5.3.2.1 Comparison of Different Models Sentiment Trends

### 5.3.3 Step 3: Stability Testing

Table 5.3.1 illustrates the Spearman’s rank correlation results of the GPT-4 Turbo model across three different trials using the same prompt, with the initial test serving as the benchmark. The results from the subsequent tests yielded rhos of 0.917350 and 0.875720, respectively. Despite a slight decrease, these values remain within the effective range, demonstrating high stability and consistency in the model’s performance.

Table 5.3.1 Model Testing Result: Stability Test

| <b>Model</b>           | <b>Spearman’s Rho</b> | <b>P-value</b>         |
|------------------------|-----------------------|------------------------|
| <i>GPT-4 Turbo 1st</i> | 1                     | 0                      |
| <i>GPT-4 Turbo 2nd</i> | 0.917350              | $5.48 \times 10^{-41}$ |
| <i>GPT-4 Turbo 3rd</i> | 0.875720              | $9.39 \times 10^{-33}$ |

Consequently, GPT-4 Turbo has successfully met all testing criteria and is selected as the sentiment analysis model for this research. Meanwhile, the zero-prompting is shown to be effective on sentiment analytical capabilities for this study.

## 5.4 Sentiment Data

Sentiment scores are derived by the GPT-4 Turbo model, using the defined prompt and news articles collected previously.

Figure 5.4.1 presents an analysis of the number of articles and their mean sentiment from various news sources used in this study. The x-axis lists the news sources, ranging from Bernama to Yahoo News, while the y-axis on the left measures the count of articles from each source, represented by bars. The right y-axis represents the mean sentiment of articles from each source, shown by the red line.

Several observations can be noted from this figure. There is significant variability in the number of articles collected from each source. Yahoo News provides the highest number of articles, significantly more than others like Bernama and Business Today, which contribute the fewest. In addition, the mean sentiment score varies across sources. Notably, Yahoo News, despite having the highest number of articles, shows a relatively lower mean sentiment compared to sources like The Sun and The Edge Malaysia, which exhibit much higher

sentiment scores. This could indicate differences in editorial tone or subject matter focus among these sources.

There is no clear correlation between the volume of articles and their mean sentiment. For instance, The Sun has fewer articles compared to Yahoo News but a much higher mean sentiment. This suggests that the quantity of coverage does not necessarily relate to the positivity or negativity of the content. The disparity in sentiment and volume across sources indicates the need for careful consideration when integrating news data into predictive models. Higher sentiment scores from sources with fewer articles might disproportionately influence model outcomes if not properly normalized or weighted.

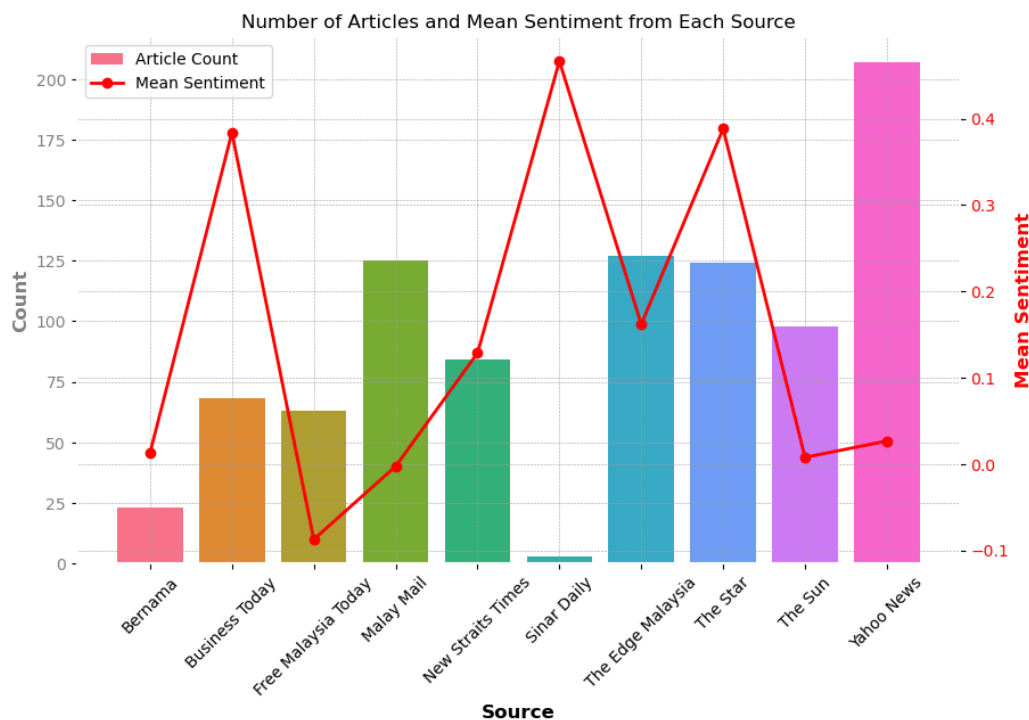


Figure 5.4.1 Number of Articles and Mean Sentiment from Each Source

## 5.5 Time series analysis

The stock price data and news sentiment scores are integrated by aligning the date of the stock prices with the publication dates of the news articles. The integration involves matching the daily stock prices with the mean sentiment scores from news articles published on the same day. This alignment helps in constructing a coherent time series dataset where each point consists of the stock price and its corresponding news sentiment score, providing a basis for further time series analysis.

|                   | <b>Stock</b> | <b>Sentiment</b> |
|-------------------|--------------|------------------|
| <b>2022-01-03</b> | 3.682776     | 0.850000         |
| <b>2022-01-04</b> | 3.673924     | 0.850000         |
| <b>2022-01-05</b> | 3.656218     | 0.425000         |
| <b>2022-01-06</b> | 3.682776     | 0.283333         |
| <b>2022-01-07</b> | 3.656218     | 0.237500         |

Figure 5.5.1 Sample Five Rows of Aligned Data

Before proceeding with further analysis, it is crucial to conduct several tests to ensure the integrity and stationarity of the integrated time series data. The data from Public Bank is selected to conduct test and fine tuning of the time series models. CIMB data will be used to evaluate if the analysis and prediction flow on Public Bank data can achieve the same conclusion.

### 5.5.1 ADF Test

The Augmented Dickey-Fuller test is conducted to assess the stationarity of the stock and sentiment data series. The Test Statistic reflects the outcome of the ADF test. A more negative test statistic indicates stronger evidence against the null hypothesis, which posits that the series has a unit root and is non-stationary. The P-value measures the probability of observing the computed statistic if the null hypothesis were true. A p-value lower than 0.05 suggests that the null hypothesis can be confidently rejected, indicating that the series is likely stationary. Critical Values represent thresholds that the test statistic must fall below to reject the null hypothesis at specified confidence levels (1%, 5%, and 10%), each corresponding to a potential risk of incorrectly rejecting the null hypothesis.

From the results [Table 5.5.1.1], the original stock data (I(0)) yielded a test statistic of -2.217 with a p-value of 0.200. This p-value exceeds typical significance levels, suggesting that the null hypothesis cannot be rejected for the original stock data, indicating it is non-stationary. Conversely, after differencing once (I(1)), the stock data shows a test statistic of -25.587 with a p-value of less than 0.001, significantly rejecting the null hypothesis at all conventional levels and confirming that the differenced stock data is stationary.

Similarly, the sentiment data, without differencing, presents a test statistic of -10.285 and a p-value of less than 0.001. This indicates that the sentiment data is stationary in its original form, strongly rejecting the null hypothesis.

Table 5.5.1.1 ADF Test Result

| Data                  | Test Statistic | P-value | Critical Values |        |        | Reject $H_0$ |
|-----------------------|----------------|---------|-----------------|--------|--------|--------------|
|                       |                |         | 1%              | 5%     | 10%    |              |
| <i>Stock I(0)</i>     | -2.217         | 0.200   | -3.439          | -2.866 | -2.569 | ✗            |
| <i>Stock I(1)</i>     | -25.587        | <0.001  | -3.439          | -2.866 | -2.569 | ✓            |
| <i>Sentiment I(0)</i> | -10.285        | <0.001  | -3.439          | -2.866 | -2.569 | ✓            |

## 5.5.2 Lags

By using ACF and PACF, this section delves into the practical application of these analytical tools to determine the optimal lag structure for the ARDL model employed in this study. In this step, the stock data is normalized by base 10 of logarithmic transformation before evaluation of autocorrelation.

Figure 5.5.2.1 displays the ACF for stock, extending over 100 lags. From the graph, it's evident that there is a significant autocorrelation at lag 0, which is always the case as a data point is perfectly correlated with itself. Notably, there's also a substantial spike at lag 1, suggesting a strong correlation between consecutive observations. This indicates that the value at one time point is closely related to its immediate predecessor, which is typical for time series data where successive measurements are often dependent.

After the first lag, the autocorrelations drop considerably and hover around zero, occasionally peaking slightly but remaining within the confidence bounds. This pattern suggests that there is little to no correlation between the observations as the lag increases beyond one time unit. The ACF quickly settling near zero implies that the influence of a given observation does not extend beyond its immediate successor, highlighting the potential memory-less nature of the process being analyzed beyond the first lag. This behavior is crucial for determining the order of autoregressive processes where the significant correlation at the first lag could suggest a potential AR(1) model fitting for this series.



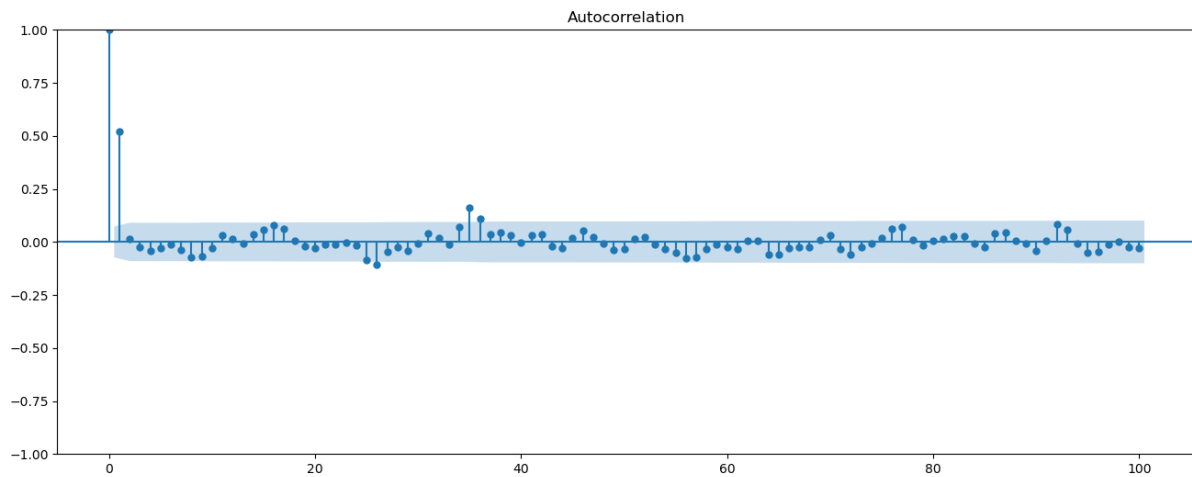


Figure 5.5.2.1 Autocorrelation Function for Stock Data

Figure 5.5.2.2 illustrates the PACF for stock data over 100 lags. The graph indicates a significant initial spike at lag 1, confirming a substantial direct correlation and suggesting the potential utility of an autoregressive component of order one. Notably, the PACF also shows a smaller but significant spike at lag 2 in the negative direction, indicating a possible oscillatory pattern which might suggest a correction or a slight reversal effect following the immediate influence captured at lag 1. Subsequent lags display smaller alternating positive and negative spikes, with diminishing magnitudes that gradually stabilize close to zero. These patterns suggest diminishing direct effects with increasing lag number, reinforcing the primary importance of the first two lags while indicating that higher-order terms may provide diminishing returns in capturing additional autocorrelation structure in the stock price data. Overall, an AR model incorporating the first two lags could potentially provide a more nuanced understanding of the stock price dynamics, capturing both the immediate effect and its immediate aftermath.

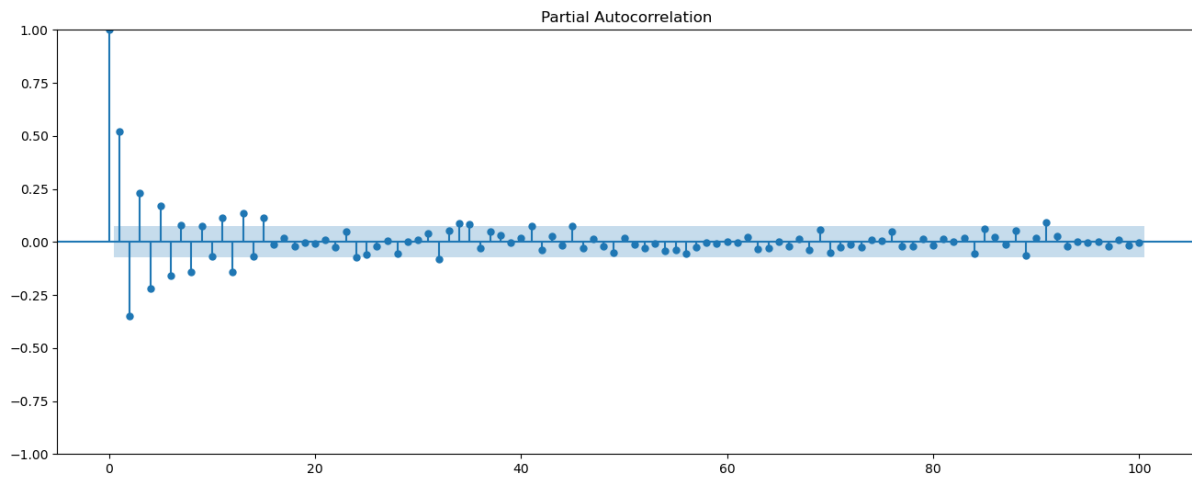


Figure 5.5.2.2 Partial Autocorrelation Function for Stock Data

Both ACF and PACF indicate significant autocorrelation at the first lag, confirming the importance of the immediate past value in predicting the current value. The PACF's additional insight into the second lag and the alternating pattern offers a deeper understanding of the data's behavior.

Figure 5.5.2.3 depicts the ACF for sentiment data analyzed over 100 lags. The ACF shows a unique, gradually decaying pattern from a strong initial positive autocorrelation at lag 0, which naturally measures a perfect correlation with itself. As the lags increase, the autocorrelation values demonstrate a consistent and smooth decay into negative territory, indicating a changing relationship in the sentiment data over time. Notably, the autocorrelations initially remain positive but decrease swiftly within the first few lags, and then start oscillating below zero. This sinusoidal-like pattern, where the autocorrelations drop below zero and gradually return to zero, suggests some periodic or cyclic behavior in the sentiment data, possibly echoing regular fluctuations in public sentiment or recurring themes in the news cycle.

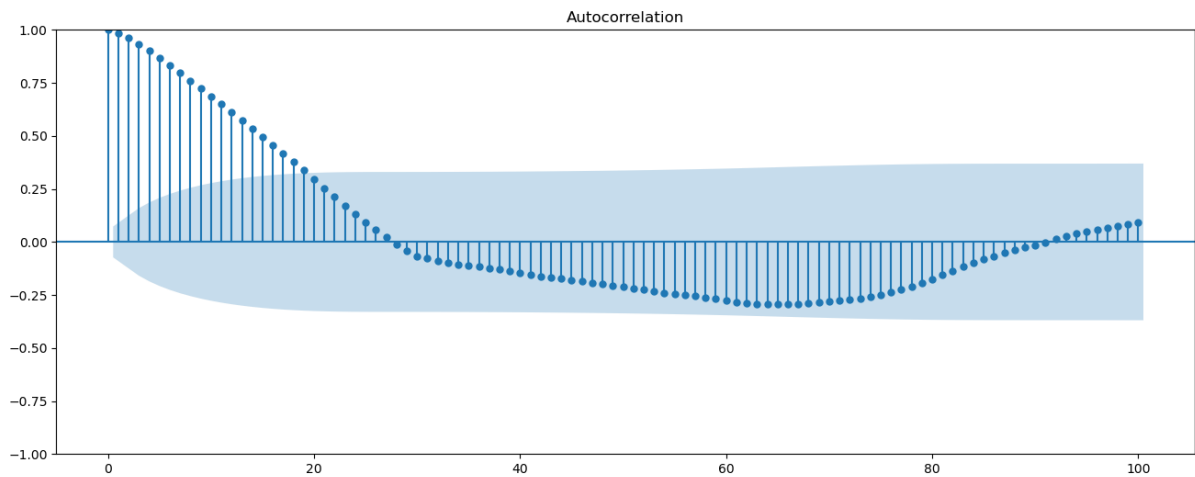


Figure 5.5.2.3 Autocorrelation Function for Sentiment Data

Figure 5.5.2.4 depicts the PACF for sentiment data analyzed over 100 lags. Initially, the PACF exhibits a substantial spike at lag 1, suggesting a significant direct correlation with the immediate previous value. This implies that yesterday’s sentiment has a substantial direct influence on today’s sentiment, supporting the inclusion of an AR(1) term in predictive modeling. Meanwhile, a significant negative spike at lag 2 stands out, indicating an inverse relationship immediately following the first lag. This could suggest a corrective effect where a particularly strong sentiment one day could lead to a rebound effect the next day.

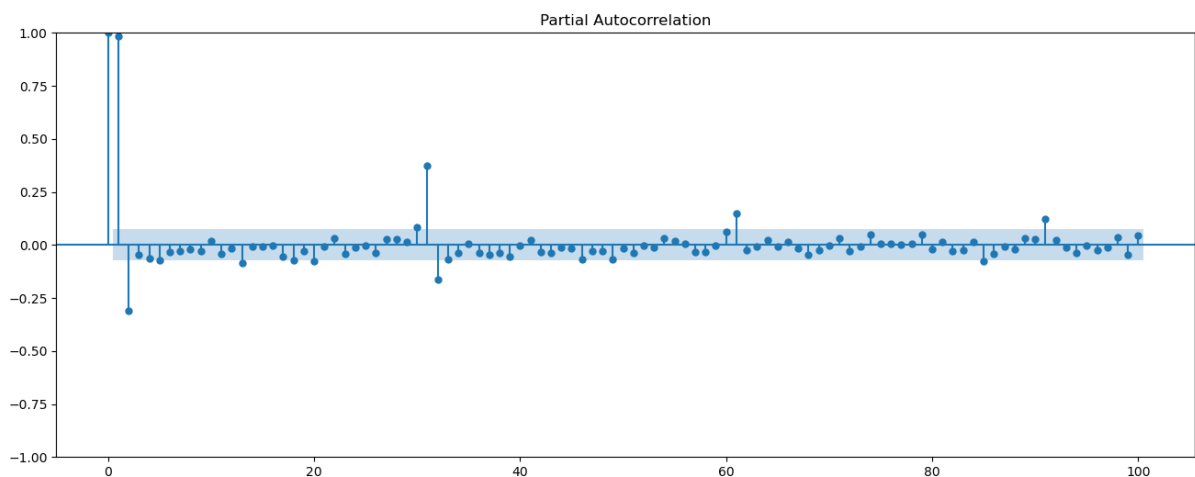


Figure 5.5.2.4 Partial Autocorrelation Function for Sentiment Data

These findings suggest that an AR(1) model could be particularly effective in capturing the primary dynamics of sentiment data, with considerations for adding corrective components to account for the rebound effect observed at lag 2.

However, when integrating both sentiment and stock data into an ARDL model, the selection of appropriate lags becomes a nuanced task that extends beyond the insights provided by the PACF and ACF. In an ARDL framework, the key is to determine the optimal lags that capture both the immediate and potentially delayed interactions between the variables effectively. While PACF and ACF are instrumental in identifying potential lags for individual time series, the specific lags still depend on the combined dynamic behavior as revealed through the ARDL modeling process.

## 5.6 Prediction

Based on the discussion in the previous section, a lag of 1 for stock data and lags of 1 to 2 for sentiment data have been selected to model the dynamic relationships within the ARDL framework. This configuration is informed by the PACF and ACF analyses, which indicated significant correlations at these specific lags.

To operationalize this, the ARDL equation for predicting future stock prices can be formalized as follows:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \gamma_1 X_{t-1} + \gamma_2 X_{t-2} + \epsilon_t \quad (7)$$

where:

- $Y_t$  is the log of stock prices at time  $t$ .
- $X_t$  represents the sentiment scores at time  $t$ .
- $\alpha, \beta, \gamma$  values are the coefficients to be estimated.
- $\epsilon_t$  is the error term.

The dataset is partitioned into training and testing subsets to rigorously evaluate the ARDL model's performance in forecasting stock prices based on sentiment data. Specifically, 80% of the data is allocated for training, involving the calibration of the ARDL model's parameters to optimally fit the historical data. The remaining 20% is reserved for testing, which provides an unbiased evaluation of the model's predictive accuracy on new, unseen data.

RMSE is employed to assess the accuracy of the ARDL model's predictions. Table 5.6.1 outlines the RMSE for each sentiment data lag when fit into the ARDL model. A lower RMSE indicates a model with higher predictive accuracy, suggesting that the corresponding lag configuration more effectively captures the dynamics influencing stock prices. The RMSE values as shown in Table 5.6.1 reveals that the lag 2 configuration results in a lower RMSE of

0.1080 compared to lag 1. This suggests that incorporating sentiment data from two days prior yields a more accurate prediction of stock prices within the ARDL model framework.

Table 5.6.1 Comparison of RMSE for Different Sentiment Data Lags

| Sentiment Data Lag | RMSE   |
|--------------------|--------|
| 1                  | 0.1261 |
| 2                  | 0.1080 |

By applying a lag of 2 for sentiment data, Figure 5.6.1 and Figure 5.6.2 showcases the ARDL model’s performance by comparing the initial stock prices with the forecasted values from August 2023 to January 2024. The graph deliberately starts in 2023 to emphasize the trend of the forecasted values. As can be observed, the forecasted trend closely follows the actual stock prices in several segments, capturing the general upward and downward movements. However, it also deviates at certain points, reflecting the inherent challenges of predicting financial markets with precision.

Besides, as the trend of forecasted values slightly lags behind the actual stock prices, it can serve as a useful reference for anticipating trend movements in advance. This slight lag in the forecast may provide investors and analysts with a predictive advantage, allowing them to foresee potential upward or downward shifts before they fully materialize in the market. By examining these forecasted trends, stakeholders can make more informed decisions, potentially gaining an edge by acting on these predictions prior to broader market recognition. This predictive feature of the ARDL model highlights its utility in financial forecasting, where even a slight lead time can significantly impact investment strategy and market positioning.

| ARDL Model Results |                  |                     |           |       |        |        |
|--------------------|------------------|---------------------|-----------|-------|--------|--------|
| =====              |                  |                     |           |       |        |        |
| Dep. Variable:     | staStock         | No. Observations:   | 581       |       |        |        |
| Model:             | ARDL(1, 2)       | Log Likelihood      | 2550.168  |       |        |        |
| Method:            | Conditional MLE  | S.D. of innovations | 0.003     |       |        |        |
| Date:              | Fri, 13 Sep 2024 | AIC                 | -5088.336 |       |        |        |
| Time:              | 15:32:31         | BIC                 | -5062.158 |       |        |        |
| Sample:            | 01-05-2022       | HQIC                | -5078.130 |       |        |        |
|                    | - 08-06-2023     |                     |           |       |        |        |
| =====              |                  |                     |           |       |        |        |
|                    | coef             | std err             | z         | P> z  | [0.025 | 0.975] |
| -----              |                  |                     |           |       |        |        |
| const              | -0.0001          | 0.000               | -0.580    | 0.562 | -0.001 | 0.000  |
| staStock.L1        | 0.0192           | 0.042               | 0.460     | 0.646 | -0.063 | 0.101  |
| staSent.L0         | 0.0052           | 0.006               | 0.866     | 0.387 | -0.007 | 0.017  |
| staSent.L1         | -0.0056          | 0.009               | -0.602    | 0.547 | -0.024 | 0.013  |
| staSent.L2         | 0.0018           | 0.006               | 0.330     | 0.742 | -0.009 | 0.013  |
| =====              |                  |                     |           |       |        |        |

Figure 5.6.1 ARDL Model Forecasting Results for a Sentiment Lag of 2

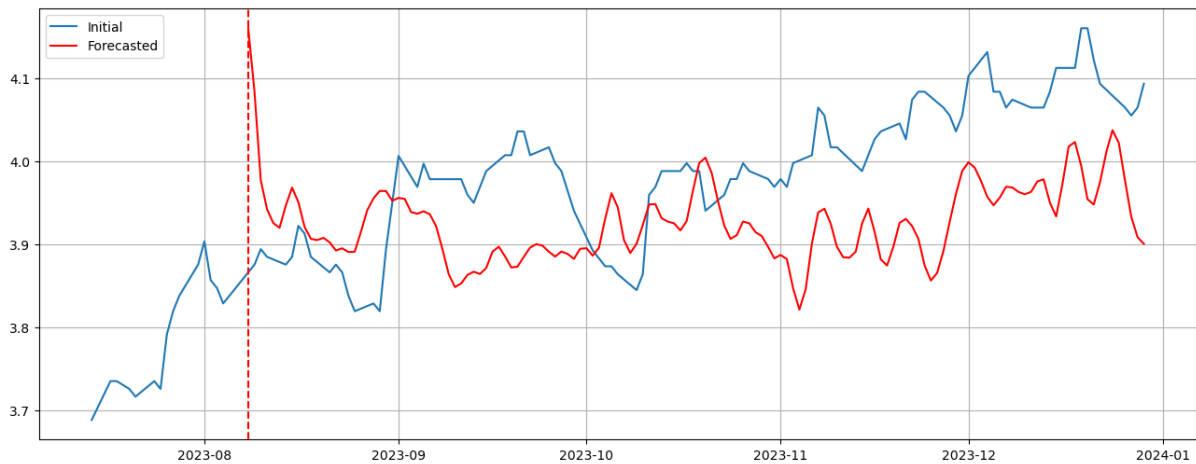


Figure 5.6.2 Comparison of Initial and Forecasted Stock Prices for Public Bank

Another dataset from CIMB is evaluated with the same settings and parameters as Public Bank. Figure 5.6.2 illustrates the forecasting performance of the ARDL model applied to CIMB’s stock data from August 2023 to January 2024. While the forecasted and actual stock prices display congruent trends at certain intervals, there are noticeable deviations in overall trajectory alignment. The forecasted values slightly lag behind the actual stock prices, similar to the trend observed with Public Bank. This lagging effect is a common characteristic seen in both datasets, indicating the model’s inherent response delay to market dynamics. The model captures several peaks and troughs, demonstrating its sensitivity to market dynamics, though not always aligning perfectly with the timing and magnitude of actual stock movements.

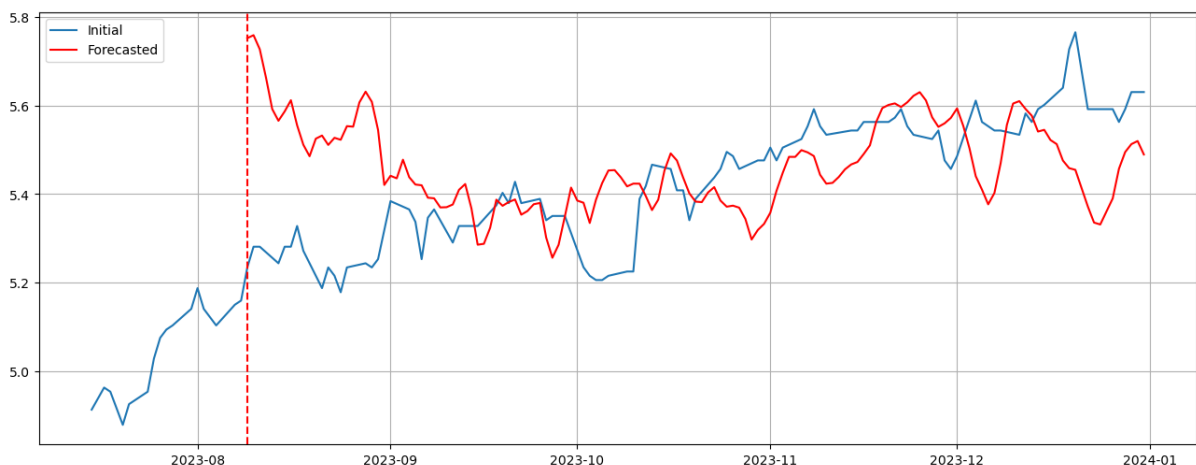


Figure 5.6.3 Comparison of Initial and Forecasted Stock Prices for CIMB

The performance of the ARDL model in forecasting stock prices for both Public Bank and CIMB reveals a nuanced efficacy. Although the model does not always match the actual stock prices with high precision, it successfully captures the general trends and oscillations in the market. The slight lag observed in the forecasted values relative to the actual prices, consistent across both datasets, suggests that while the model may not be optimal for precise timing predictions, it remains valuable for anticipating general market movements.

This predictive capability of the ARDL model directly contributes to the research's primary aim of evaluating the performance and potential of models in financial sentiment analysis and their applicability in predicting stock market movements. The successful capture of general market trends by the ARDL model, despite slight lags, demonstrates its usefulness in providing foresight into market dynamics, which aligns with the sub-objective of developing a simplified analysis process by correlating sentiment scores with stock trends.

## 5.7 Source Comparison

Table 5.7.1 presents a summary of the number of articles and the RMSE for each news source provided by Public Bank, providing insights into how the quantity of content and the predictive quality of sentiment data from different sources influence ARDL model stock market forecasting accuracy.

From the table, we can observe Malay Mail has a relatively low RMSE of 0.0702 despite having a moderate number of articles (125), suggesting that its sentiment data may be more predictive or more aligned with stock market movements. Yahoo News and The Star both have a high number of articles (127 and 124 respectively) with relatively low RMSE values (0.0984 and 0.0970 respectively), indicating effective sentiment analysis contributing to accurate stock price predictions. Free Malaysia Today and The Sun, with fewer articles (63 and 98 respectively), exhibit higher RMSE values (0.1521 and 0.1364 respectively), which could indicate less predictive accuracy or less relevant sentiment data impacting stock price movements. Notably, Bernama and Sinar Daily have dashes in the RMSE column due to the insufficient number of articles (23 and 3 respectively), making it difficult to compute a reliable RMSE.

These findings suggest that not only the quantity but also the quality of sentiment analysis from each news source plays a crucial role in forecasting stock prices. Sources like Malay Mail, despite not having the highest article count, demonstrate better forecasting capabilities possibly due to more relevant or accurate sentiment extraction from their articles. This analysis aligns

with the research objectives to evaluate how different news sources influence stock market movements and to streamline the use of sentiment analysis in stock prediction.

Table 5.7.1 Prediction Performance for Each News Sources

| <b>News Source</b>         | <b>Number of Article</b> | <b>RMSE</b> |
|----------------------------|--------------------------|-------------|
| <i>The Star</i>            | 124                      | 0.0970      |
| <i>New Straits Times</i>   | 84                       | 0.1206      |
| <i>Yahoo News</i>          | 127                      | 0.0984      |
| <i>Malay Mail</i>          | 125                      | 0.0702      |
| <i>Free Malaysia Today</i> | 63                       | 0.1521      |
| <i>Bernama</i>             | 23                       | -           |
| <i>The Edge Malaysia</i>   | 127                      | 0.1133      |
| <i>The Sun</i>             | 98                       | 0.1364      |
| <i>Business Today</i>      | 68                       | 0.0823      |
| <i>Sinar Daily</i>         | 3                        | -           |



# Chapter 6

## Conclusion

### 6.1 Conclusion

The realm of stock investment full of risk and uncertainty, and therefore, an effective stock prediction method is essential for investors and financial professionals. While many prediction techniques proposed, existing methods often overlook the influence of different news sources, specifically in Malaysia, and utilise outdated sentiment analysis techniques in the prediction model. This project aimed to integrate the state-of-the-art GPT model and propose a simplified prediction method in stock investment.

The research explored the efficacy of the ARDL model combined with GPT-4 Turbo for sentiment analysis to predict stock market movements, focusing on the impact of various news sources on stock price fluctuations. The sentiment analysis, powered by GPT-4 Turbo, provided a robust framework for evaluating the emotional tone of financial news. Although no model yields 100% accuracy in sentiment prediction, this tool served as a valuable benchmark in the study, aiding in the correlation of sentiment scores with market movements and offering insights into the predictive relevance of various news sources.

To achieve its objectives, the project designed and conducted comprehensive tests to evaluate the capabilities and performance of the GPT-4 Turbo model in analyzing the sentiment of financial news. This exploration proved instrumental in establishing a robust framework for sentiment analysis, although it also highlighted the inherent limitations of current AI technologies in achieving absolute accuracy. Besides, the research extensively investigated how different news sources influenced stock market movements. By utilizing the ARDL model integrated with sentiment scores derived from GPT-4 Turbo, the study quantified the impact of various news outlets. This analysis revealed significant variability in the predictive relevance of different sources, thereby providing valuable insights into which sources wield the most significant impact on stock prices.

A key aim was to develop a simplified process by comparing sentiment scores with stock trends, thereby streamlining the use of sentiment analysis in stock prediction. Through systematic data preprocessing and integration of sentiment analysis with the ARDL model, the

research demonstrated a practical approach to combining quantitative sentiment data with stock market analytics, though it also underscored the need for further refinement to enhance predictive accuracy.

The results and discussions provided a critical examination of the predictive capabilities of the ARDL model. Despite its sophisticated integration of sentiment analysis, the ARDL model demonstrated moderate success in accurately forecasting stock market trends. This was evidenced by the variability in RMSE values across different news sources, suggesting that while some sources provided valuable predictive insights, others might require more nuanced analysis techniques or could be less influential than hypothesized.

This research not only met its objectives but also opened avenues for future inquiries into the integration of more nuanced sentiment analysis techniques and hybrid predictive models that may offer greater accuracy and reliability in stock market forecasting.

## **6.2 Recommendations**

The integration of additional macroeconomic variables and alternative sentiment indicators to capture market dynamics more comprehensively. This could involve including factors like economic indicators, market indices, or even global events that significantly impact market movements. Moreover, there is a potential benefit in exploring hybrid models that combine the strengths of ARDL with other predictive algorithms, such as machine learning and deep learning approaches. These hybrid models could potentially improve both the accuracy and robustness of the predictions by leveraging the unique capabilities of each approach.

In addition, to keep pace with the rapidly changing financial markets, continuous updates and training on newer datasets are crucial for enhancing the GPT model's understanding of market linguistics and evolving sentiment. This involves retraining the models periodically with updated news articles, financial reports, and market data that reflect current market conditions and sentiment. Moreover, implementing ensemble techniques that combine the outputs of several sentiment analysis models could significantly refine the accuracy and reliability of sentiment scores. This ensemble approach can mitigate the weaknesses of individual models and provide a more balanced and nuanced analysis of market sentiment.

On the other hand, expanding the scope of data sources is vital for capturing a more holistic view of market sentiment. This could include broadening the data acquisition to encompass social media platforms, financial forums, and analyst reports, which can provide a wealth of real-time and diverse perspectives on market conditions. Moreover, utilizing advanced web

scraping techniques to harness real-time data would allow for dynamic model adjustments and predictions. These techniques ensure that the models have access to the most current data, enhancing the timeliness and relevance of the predictive outputs.

Future studies could explore the intersection of behavioural finance theories with computational models to understand the psychological factors driving market trends. Investigating the impact of geopolitical events or unexpected market shocks could also provide deeper insights into the resilience and adaptability of predictive models in finance.

## REFERENCES

- [1] N. Novemsky, & D. Kahneman, "The Boundaries of Loss Aversion." *Journal of Marketing Research*, 42(2), 119-128, 2005. Available: <https://doi.org/10.1509/jmkr.42.2.119.62292>
- [2] U. Schmidt, S. Traub, "An Experimental Test of Loss Aversion." *Journal of Risk and Uncertainty* 25, 233–249, 2002. Available: <https://doi.org/10.1023/A:1020923921649>
- [3] U. Schmidt, H. Zank, "What is Loss Aversion?." *J Risk Uncertainty* 30, 157–167, 2005. Available: <https://doi.org/10.1007/s11166-005-6564-6>
- [4] S. M. Tom, C. R. Fox, C. Trepel, & R. A. Poldrack,. "The Neural Basis of Loss Aversion in Decision-Making Under Risk." *Science*, 315(5811), 515-518, 2007. DOI:10.1126/science.1134239
- [5] R. Brealey, S. Myers, & F. Allen, "Principles of Corporate Finance (Finance, Insurance and Real Estate)." *McGraw-hill*, 2014.
- [6] F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," *Journal of Finance* 25, May. 1970, pp. 383–417.
- [7] A. W. Lo, "The adaptive markets hypothesis: Market efficiency from an evolutionary perspective." *Journal of Portfolio Management*, Forthcoming. 2004. Available: <http://stat.wharton.upenn.edu/~stele/Courses/434/434Context/EfficientMarket/AndyLoJPM2004.pdf>
- [8] A. W. Lo, "Reconciling efficient markets with behavioral finance: the adaptive markets hypothesis." *Journal of investment consulting*, 7(2), 2005, 21-44. Available: [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=1702447#paper-citations-widget](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1702447#paper-citations-widget)
- [9] R. Goonatilake, & S. Herath, "The volatility of the stock market and news." *International Research Journal of Finance and Economics*, 3(11), 2007, 53-65. Available: <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=567df765560dfb426c184bd34b47510eaaabb500>
- [10] S. R. Baker, N. Bloom, S. J. Davis, & K. J. Kost, "Policy news and stock market volatility (No. w25720)." *National Bureau of Economic Research*, March. 2019. doi: 10.3386/w25720
- [11] S. Merello, A. Picasso Ratto, Y. Ma, O. Luca and E. Cambria, "Investigating Timing and Impact of News on the Stock Market," *2018 IEEE International Conference on Data Mining Workshops (ICDMW), Singapore*, 2018, pp. 1348-1354, doi: 10.1109/ICDMW.2018.00191.
- [12] R. Ren, D. D. Wu and T. Liu, "Forecasting Stock Market Movement Direction Using Sentiment Analysis and Support Vector Machine," in *IEEE Systems Journal*, vol. 13, no. 1, pp. 760-770, March 2019, doi: 10.1109/JSYST.2018.2794462.
- [13] V. S. Pagolu, K. N. Reddy, G. Panda and B. Majhi, "Sentiment analysis of Twitter data for predicting stock market movements," *2016 International Conference on Signal Processing*,

*Communication, Power and Embedded System (SCOPEs)*, Paralakhemundi, India, 2016, pp. 1345-1350, doi: 10.1109/SCOPEs.2016.7955659.

[14] A. E. Khedr, S. E. Salama, and N. Yaseen, "Predicting stock market behavior using data mining technique and news sentiment analysis," *International Journal of Intelligent Systems and Applications*, vol. 9, no. 7, pp. 22–30, Jul. 2017, doi: 10.5815/ijisa.2017.07.03.

[15] T. H. Nguyen, & K. Shirai, "Topic modeling based sentiment analysis on social media for stock market prediction." In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, July. 2015, pp. 1354-1364. Available: <https://aclanthology.org/P15-1131.pdf>

[16] A. M. Abirami and V. Gayathri, "A survey on sentiment analysis methods and approach," *2016 Eighth International Conference on Advanced Computing (ICoAC)*, Chennai, India, 2017, pp. 72-76, doi: 10.1109/ICoAC.2017.7951748.

[17] Q. T. Ain, M. Ali, A. Riaz, A. Noureen, M. Kamran, B. Hayat, & A. Rehman, (2017). "Sentiment analysis using deep learning techniques: a review." *International Journal of Advanced Computer Science and Applications*, 2017, 8(6). Available: [https://www.academia.edu/download/74943484/Paper\\_57-Sentiment\\_Analysis\\_using\\_Deep\\_Learning.pdf](https://www.academia.edu/download/74943484/Paper_57-Sentiment_Analysis_using_Deep_Learning.pdf)

[18] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. T. Chitkushev and D. Trajanov, "Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers," in *IEEE Access*, vol. 8, pp. 131662-131682, 2020, doi: 10.1109/ACCESS.2020.3009626.

[19] A. Sadia, F. Khan, & F. Bashir, "An overview of lexicon-based approach for sentiment analysis." In *2018 3rd International Electrical Engineering Conference (IEEC 2018)*, Feb. 2018, pp. 1-6.

[20] R. Steinert and S. Altmann, "Linking microblogging sentiments to stock price movement: An application of GPT-4," Aug. 2023, [Online]. Available: <http://arxiv.org/abs/2308.16771>

[21] K. Kheiri and H. Karimi, "SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning," Jul. 2023, [Online]. Available: <http://arxiv.org/abs/2307.10234>

[22] S. Kothapalli and S. G. Totad, "A real-time weather forecasting and analysis," *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, Chennai, India, 2017, pp. 1567-1570, doi: 10.1109/ICPCSI.2017.8391974.

[23] R. Fildes and N. Kourentzes, "Validation and forecasting accuracy in models of climate change," *International Journal of Forecasting*, vol. 27, no. 4, pp. 968-995, 2011, ISSN 0169-2070.

[24] A. Chhabra, S. K. Singh, A. Sharma, S. Kumar, B. B. Gupta, V. Arya, and K. T. Chui, "Sustainable and intelligent time-series models for epidemic disease forecasting and analysis," *Sustainable Technology and Entrepreneurship*, vol. 3, no. 2, Art. no. 100064, 2024.

- [25] G. Nunnari and V. Nunnari, "Forecasting Monthly Sales Retail Time Series: A Case Study," *2017 IEEE 19th Conference on Business Informatics (CBI)*, Thessaloniki, Greece, 2017, pp. 1-6, doi: 10.1109/CBI.2017.57.
- [26] N. F. Ibrahim and X. Wang, "Decoding the sentiment dynamics of online retailing customers: Time series analysis of social media," *Computers in Human Behavior*, vol. 96, pp. 32-45, 2019.
- [27] Y. Peng and H. Jiang, "Leverage Financial News to Predict Stock Price Movements Using Word Embeddings and Deep Neural Networks," Jun. 2015, [Online]. Available: <http://arxiv.org/abs/1506.07220>
- [28] J. Kalyani, P. Bharathi, & P. Jyothi, "Stock trend prediction using news sentiment analysis." *arXiv preprint arXiv:1607.01958*, 2016.
- [29] L. Nemes and A. Kiss, "Prediction of stock values changes using sentiment analysis of stock news headlines," *Journal of Information and Telecommunication*, vol. 5, no. 3, pp. 375–394, 2021, doi: 10.1080/24751839.2021.1874252.
- [30] B. Fazlija and P. Harder, "Using Financial News Sentiment for Stock Price Direction Prediction," *Mathematics*, vol. 10, no. 13, Jul. 2022, doi: 10.3390/math10132156.
- [31] M. P. Cristescu, R. A. Nerisanu, D. A. Mara, and S. V. Oprea, "Using Market News Sentiment Analysis for Stock Market Prediction," *Mathematics*, vol. 10, no. 22, Nov. 2022, doi: 10.3390/math10224255.
- [32] S. Usmani and J. A. Shamsi, "LSTM based stock prediction using weighted and categorized financial news," *PLoS One*, vol. 18, no. 3 March, Mar. 2023, doi: 10.1371/journal.pone.0282234.
- [33] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners." [Online]. Available: <https://github.com/codelucas/newspaper>
- [34] A. Xie, C. Bruckmann, and T. Qian, "Sensitivity Analysis on Transferred Neural Architectures of BERT and GPT-2 for Financial Sentiment Analysis." [Online]. Available: [https://github.com/axie123/gpt\\_bert\\_transfer\\_arch](https://github.com/axie123/gpt_bert_transfer_arch).
- [35] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," May 2020, [Online]. Available: <http://arxiv.org/abs/2005.14165>
- [36] OpenAI, "GPT-4," Mar. 14, 2023. <https://openai.com/research/gpt-4> (accessed Sep. 09, 2023).
- [37] H. Taherdoost, H. (2021). "Data collection methods and tools for research; a step-by-step guide to choose data collection technique for academic and business research projects". *International Journal of Academic Research in Management (IJARM)*, 10(1), 10-38.

- [38] Google, “Refine web searches,” *Google Search Help*, 2023. [Online]. Available: <https://support.google.com/websearch/answer/2466433?hl=en>. [Accessed: Sept. 1, 2023].
- [39] Chengqing Zong, Rui Xia, and Jiajun Zhang, *Text Data Mining*, Singapore: Springer, 2021.
- [40] T. B. Brown. (2020). “Language models are few-shot learners.” *arXiv preprint arXiv:2005.14165*.
- [41] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, et al. (2022). “Chain-of-thought prompting elicits reasoning in large language models.” *Advances in neural information processing systems*, 35, 24824-24837.
- [42] OpenAI, “Models,” [Online]. Available: <https://platform.openai.com/docs/models>
- [43] A. Vaswani, (2017). “Attention is all you need.” *Advances in Neural Information Processing Systems*.
- [44] OpenAI, “Pricing,” [Online]. Available: <https://openai.com/api/pricing/>
- [45] S. E. Said & D. A. Dickey (1984). “Testing for unit roots in autoregressive-moving average models of unknown order.” *Biometrika*, 71(3), 599-607.

# FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

|                                                                                   |                          |
|-----------------------------------------------------------------------------------|--------------------------|
| <b>Trimester, Year:</b> June, 2024                                                | <b>Study week no.:</b> 6 |
| <b>Student Name &amp; ID:</b> Tan Lin 22ACB00188                                  |                          |
| <b>Supervisor:</b> Dr. Kh'ng Xin Yi                                               |                          |
| <b>Project Title:</b> Using Sentiment Analysis to Forecast Stock Short-term Trend |                          |

## 1. WORK DONE

Data is collected.

## 2. WORK TO BE DONE

Learn and apply Autoregressive Distributed Lag Model from the suggested websites.

## 3. PROBLEMS ENCOUNTERED

None

## 4. SELF EVALUATION OF THE PROGRESS

Moderate in overall progress



Supervisor's signature

Tan Lin

Student's signature



# FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

|                                                                                   |                          |
|-----------------------------------------------------------------------------------|--------------------------|
| <b>Trimester, Year:</b> June, 2024                                                | <b>Study week no.:</b> 8 |
| <b>Student Name &amp; ID:</b> Tan Lin 22ACB00188                                  |                          |
| <b>Supervisor:</b> Dr. Kh'ng Xin Yi                                               |                          |
| <b>Project Title:</b> Using Sentiment Analysis to Forecast Stock Short-term Trend |                          |

## 1. WORK DONE

ARDL prediction on stock movement.

## 2. WORK TO BE DONE

Fine tuning the model

## 3. PROBLEMS ENCOUNTERED

Model predictions are not effective

## 4. SELF EVALUATION OF THE PROGRESS

Moderate in overall progress



Supervisor's signature

Tan Lin

Student's signature

# FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

|                                                                                   |                           |
|-----------------------------------------------------------------------------------|---------------------------|
| <b>Trimester, Year:</b> June, 2024                                                | <b>Study week no.:</b> 10 |
| <b>Student Name &amp; ID:</b> Tan Lin 22ACB00188                                  |                           |
| <b>Supervisor:</b> Dr. Kh'ng Xin Yi                                               |                           |
| <b>Project Title:</b> Using Sentiment Analysis to Forecast Stock Short-term Trend |                           |

## 1. WORK DONE

Fine-tuned ARDL prediction result, compare news sources influences on stock movement.

## 2. WORK TO BE DONE

Fine tuning the model.

## 3. PROBLEMS ENCOUNTERED

Model predictions are not effective.

## 4. SELF EVALUATION OF THE PROGRESS

Moderate in overall progress



\_\_\_\_\_  
Supervisor's signature

Tan Lin

\_\_\_\_\_  
Student's signature

# POSTER

FICT

COMPUTER SCIENCE

# UTAR

# POSTER

## USING SENTIMENT ANALYSIS TO FORECAST STOCK SHORT-TERM TREND

### INTRODUCTION

In an era where financial markets are increasingly influenced by vast arrays of data, understanding the impact of news sentiment on stock movements is crucial. This study explores the efficacy of combining advanced AI models, specifically GPT-4 Turbo, with ARDL methodologies to enhance stock market prediction.

### REFERENCES

A selected list of academic papers and datasets that provided the foundation for this research.

### OBJECTIVES

- EVALUATE GPT MODELS FOR FINANCIAL SENTIMENT ANALYSIS.
- ASSESS THE INFLUENCE OF DIFFERENT NEWS SOURCES ON STOCK PRICES.
- DEVELOP A STREAMLINED PROCESS FOR INTEGRATING SENTIMENT ANALYSIS IN STOCK PREDICTIONS.

Public Bank Stock Prices 2022 to 2023



### CONTRIBUTION

This research advances the field by demonstrating the effectiveness of integrating advanced AI-driven sentiment analysis with econometric models to enhance the accuracy of stock market predictions.

### METHODOLOGY

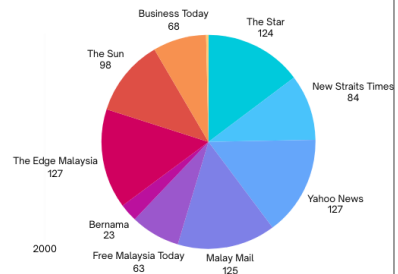
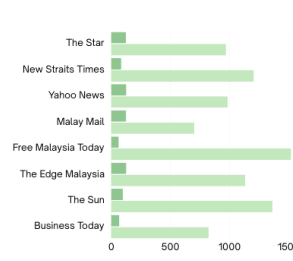
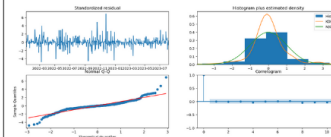
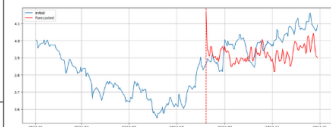
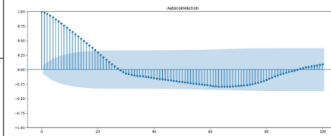
Employed the ARDL model to integrate sentiment scores derived from various news sources using GPT-4 Turbo.

- Scrap Data
- Sentiment Analysis
- Time Series Analysis

This approach allowed us to assess the dynamic interactions between news sentiment and stock market fluctuations over a specified period.

### RESULTS/FINDINGS

- Sentiment Analysis Accuracy:** "GPT-4 Turbo provided a robust framework for sentiment analysis, though it confirmed that no tool achieves 100% accuracy."
- News Source Impact:** "Analysis revealed significant variability in how different news sources influenced stock prices, with some sources showing more predictive power than others."
- Model Performance:** "The ARDL model showed moderate success in forecasting stock prices, suggesting further refinement is needed for enhanced accuracy."



### ANALYSIS

This study systematically applied the Autoregressive Distributed Lag (ARDL) model to assess the influence of news sentiments, extracted using the GPT-4 Turbo model, on stock market trends. By quantitatively examining sentiment data from various news sources and correlating it with stock price movements, we uncovered significant differences in the predictive power of each source. This methodological approach allowed to isolate the impact of sentiment on stock prices, providing a nuanced understanding of how news influences market behavior and highlighting the potential for using advanced AI in financial analytics to improve investment strategies.

### CONCLUSION

The study highlighted the potential of using advanced AI tools for financial market analysis, underscoring the need for continued enhancements in model accuracy and sentiment analysis techniques. Future work will focus on integrating additional data sources and refining AI algorithms to improve prediction models.

By Tan Lin  
Under supervision of Dr. Kh'ng Xin Yi

## PLAGIARISM CHECK RESULT

[check] FYP2\_Report.pdf

### ORIGINALITY REPORT

|                                |                               |                           |                             |
|--------------------------------|-------------------------------|---------------------------|-----------------------------|
| <b>10%</b><br>SIMILARITY INDEX | <b>4%</b><br>INTERNET SOURCES | <b>7%</b><br>PUBLICATIONS | <b>4%</b><br>STUDENT PAPERS |
|--------------------------------|-------------------------------|---------------------------|-----------------------------|

### PRIMARY SOURCES

|          |                                                                                                                                                                                                                                 |               |
|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------|
| <b>1</b> | <b>"Advances in Databases and Information Systems", Springer Science and Business Media LLC, 2022</b><br>Publication                                                                                                            | <b>2%</b>     |
| <b>2</b> | <b>Sachi Nandan Mohanty, Preethi Nanjundan, Tejaswini Kar. "Artificial Intelligence in Forecasting - Tools and Techniques", CRC Press, 2024</b><br>Publication                                                                  | <b>&lt;1%</b> |
| <b>3</b> | <b>fastercapital.com</b><br>Internet Source                                                                                                                                                                                     | <b>&lt;1%</b> |
| <b>4</b> | <b>Ayman E. Khedr, S.E.Salama, Nagwa Yaseen. "Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis", International Journal of Intelligent Systems and Applications, 2017</b><br>Publication | <b>&lt;1%</b> |
| <b>5</b> | <b>Chengqing Zong, Rui Xia, Jiajun Zhang. "Text Data Mining", Springer Science and Business Media LLC, 2021</b><br>Publication                                                                                                  | <b>&lt;1%</b> |

|                                                                                                                                                                  |            |                            |                  |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|----------------------------|------------------|
| <b>Universiti Tunku Abdul Rahman</b>                                                                                                                             |            |                            |                  |
| <b>Form Title : Supervisor's Comments on Originality Report Generated by Turnitin for Submission of Final Year Project Report (for Undergraduate Programmes)</b> |            |                            |                  |
| Form Number: FM-IAD-005                                                                                                                                          | Rev No.: 0 | Effective Date: 01/10/2013 | Page No.: 1 of 1 |




**FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY**

|                                     |                                                             |
|-------------------------------------|-------------------------------------------------------------|
| <b>Full Name(s) of Candidate(s)</b> | Tan Lin                                                     |
| <b>ID Number(s)</b>                 | 22ACB00188                                                  |
| <b>Programme / Course</b>           | CS                                                          |
| <b>Title of Final Year Project</b>  | Using Sentiment Analysis to Forecast Stock Short-term Trend |

| <b>Similarity</b>                                                                                                                                                                                                                                                                                                                                                                                                   | <b>Supervisor's Comments<br/>(Compulsory if parameters of originality exceeds the limits approved by UTAR)</b> |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------|
| <b>Overall similarity index: <u>10</u> %</b><br><br><b>Similarity by source</b><br>Internet Sources: <u>4</u> %<br>Publications: <u>7</u> %<br>Student Papers: <u>4</u> %                                                                                                                                                                                                                                           | Overall similarity index < 20%                                                                                 |
| <b>Number of individual sources listed of more than 3% similarity: <u>-</u></b>                                                                                                                                                                                                                                                                                                                                     |                                                                                                                |
| <b>Parameters of originality required and limits approved by UTAR are as Follows:</b><br>(i) Overall similarity index is 20% and below, and<br>(ii) Matching of individual sources listed must be less than 3% each, and<br>(iii) Matching texts in continuous block must not exceed 8 words<br><i>Note: Parameters (i) – (ii) shall exclude quotes, bibliography and text matches which are less than 8 words.</i> |                                                                                                                |

Note Supervisor/Candidate(s) is/are required to provide softcopy of full set of the originality report to Faculty/Institute

***Based on the above results, I hereby declare that I am satisfied with the originality of the Final Year Project Report submitted by my student(s) as named above.***



\_\_\_\_\_  
Signature of Supervisor

Name: Kh'ng Xin Yi

Date: 13/9/2024

\_\_\_\_\_  
Signature of Co-Supervisor

Name: \_\_\_\_\_

Date: \_\_\_\_\_



**UNIVERSITI TUNKU ABDUL RAHMAN**

**FACULTY OF INFORMATION & COMMUNICATION TECHNOLOGY  
(KAMPAR CAMPUS)**

**CHECKLIST FOR FYP2 THESIS SUBMISSION**

|                 |                  |
|-----------------|------------------|
| Student Id      | 22ACB00188       |
| Student Name    | Tan Lin          |
| Supervisor Name | Dr. Kh'ng Xin Yi |

| <b>TICK (√)</b> | <b>DOCUMENT ITEMS</b>                                                                                                                                             |
|-----------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                 | Your report must include all the items below. Put a tick on the left column after you have checked your report with respect to the corresponding item.            |
| √               | Title Page                                                                                                                                                        |
| √               | Signed Report Status Declaration Form                                                                                                                             |
| √               | Signed FYP Thesis Submission Form                                                                                                                                 |
| √               | Signed form of the Declaration of Originality                                                                                                                     |
| √               | Acknowledgement                                                                                                                                                   |
| √               | Abstract                                                                                                                                                          |
| √               | Table of Contents                                                                                                                                                 |
| √               | List of Figures (if applicable)                                                                                                                                   |
| √               | List of Tables (if applicable)                                                                                                                                    |
| √               | List of Symbols (if applicable)                                                                                                                                   |
| √               | List of Abbreviations (if applicable)                                                                                                                             |
| √               | Chapters / Content                                                                                                                                                |
| √               | Bibliography (or References)                                                                                                                                      |
| √               | All references in bibliography are cited in the thesis, especially in the chapter of literature review                                                            |
|                 | Appendices (if applicable)                                                                                                                                        |
| √               | Weekly Log                                                                                                                                                        |
| √               | Poster                                                                                                                                                            |
| √               | Signed Turnitin Report (Plagiarism Check Result - Form Number: FM-IAD-005)                                                                                        |
| √               | I agree 5 marks will be deducted due to incorrect format, declare wrongly the ticked of these items, and/or any dispute happening for these items in this report. |

\*Include this form (checklist) in the thesis (Bind together as the last page)

I, the author, have checked and confirmed all the items listed in the table are included in my report.

\_\_\_\_\_  
Tan Lin  
(Signature of Student)  
Date: 13/9/2024