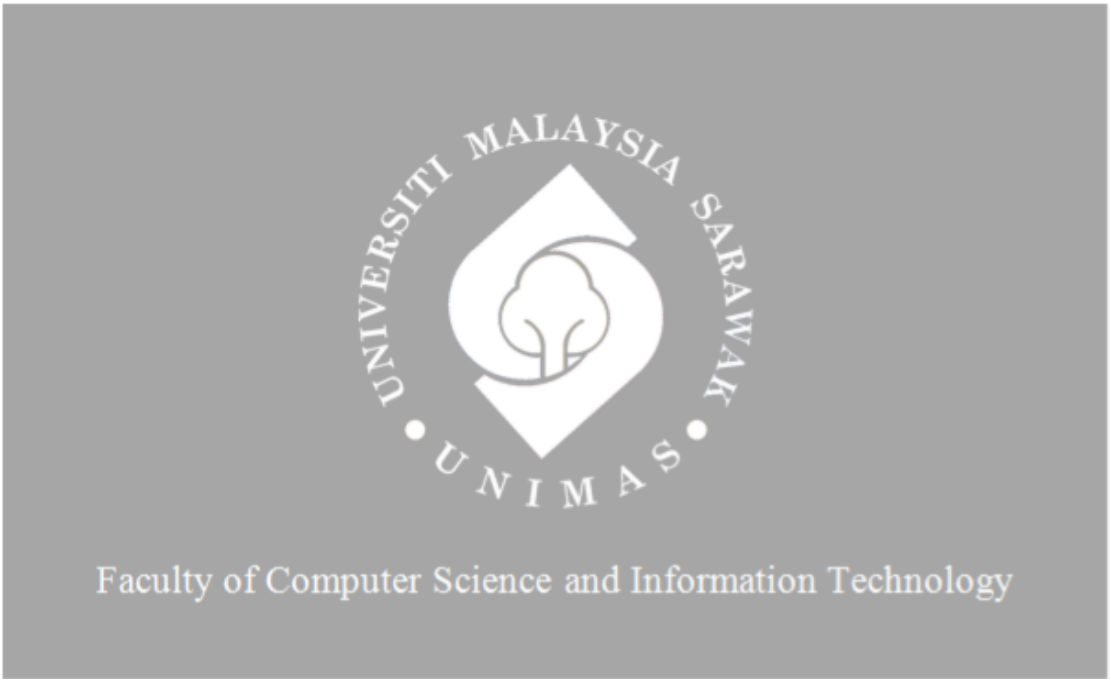

Predicting US Stock Prices Using Long Short-Term Memory (LSTM)

DAYANG AFIQAH LIYANA BINTI ABANG EHSAN

Bachelor of Computer Science with Honours

(Multimedia Computing)

2023

UNIVERSITI MALAYSIA SARAWAK

THESIS STATUS ENDORSEMENT FORM

TITLE Predicting US Stock Prices Using Long Short-Term Memory (LSTM)

ACADEMIC SESSION: 2022/2023

DAYANG AFIQAH LIYANA BINTI ABANG EHSAN

(CAPITAL LETTERS)

hereby agree that this Thesis* shall be kept at the Centre for Academic Information Services, Universiti Malaysia Sarawak, subject to the following terms and conditions:

1. The Thesis is solely owned by Universiti Malaysia Sarawak
2. The Centre for Academic Information Services is given full rights to produce copies for educational purposes only
3. The Centre for Academic Information Services is given full rights to do digitization in order to develop local content database
4. The Centre for Academic Information Services is given full rights to produce copies of this Thesis as part of its exchange item program between Higher Learning Institutions [or for the purpose of interlibrary loan between HLI]
5. ** Please tick (√)

☐	CONFIDENTIAL	(Contains classified information bounded by the OFFICIAL SECRETS ACT 1972)
☐	RESTRICTED	(Contains restricted information as dictated by the body or organization where the research was conducted)
☑	UNRESTRICTED	

Validated by

DR SHAHIEE ABD RAHMAN
Senior Lecturer (DS52)
Computational Science Programme
Faculty of Computer Science & Information Technology
Universiti Malaysia Sarawak

(AUTHOR'S SIGNATURE) (SUPERVISOR'S SIGNATURE)

Permanent Address

No.20, Lorong Depo 11, Taman
Sebah Heights, Jalan Depo,
Petra Jaya, 93050, Kuching, Sarawak

Date: 1/7/2023 Date:1/7/2023

Note * Thesis refers to PhD, Master, and Bachelor Degree
** For Confidential or Restricted materials, please attach relevant documents from relevant organizations / authorities

Declaration

I, Dayang Afiqah Liyana Binti Abang Ehsan, hereby declare that the thesis entitled "Predicting US Stock Prices Using Long Short-Term Memory (LSTM)" is entirely based on my own original work, except for appropriate quotations and citations. I also declare that it has not been submitted for any other degree at the Universiti Malaysia Sarawak (UNIMAS) previously or currently.

Signature,

……..

(DAYANG AFIQAH LIYANA BINTI ABANG EHSAN)

Faculty of Computer Science and Information Technology

Date: 25th January 2023

Universiti Malaysia Sarawak

Acknowledgement

First and foremost, all praise to Allah for all His blessings and guidance to me for successfully completing this project after spending about four months of hard works along with numerous challenges and problems that I have encountered.

I would like to express my heartfelt appreciation to my supervisor, Dr. Shapi-ee bin Abd Rahman, and my co-supervisor, Dr. Nurul Syuhada binti Ismail, for granting me the opportunity to conduct this research and for their invaluable guidance and support throughout the process. I am deeply grateful to both Dr. Shapi-ee bin Abd Rahman and Dr. Nurul Syuhada binti Ismail for being a constant source of inspiration and motivation, and for providing valuable references that have contributed to the completion of this research.

My deepest gratitude for my parents and my family, Izzati Nadhirah Binti Mohammad Saberi, Anis Hazirah Binti Mohammad Saberi, Yusha ‘Athirah Binti Rozman and Husna Amani Binti Rozman. Special thanks to my friends, Nur’azra Alia Nisa Binti Zulpakar, Syahazwani Nurain Binti Shariman Faizul, Nurfazlina Binti Yusuf, Nurin Alya Binti Haris, Siti Rubiah Binti Muslim, Dayang Nurazyyati Binti Awang Suffian and Elia Sari Binti Odita who also have helped me a lot in giving guidance and information regarding my research.

Lastly thank you to those who are not mentioned above and whoever was involved directly and indirectly in completing this final year project.

Thank You,

DAYANG AFIQAH LIYANA BINTI ABANG EHSAN

Abstract

A stock, commonly referred to as equity, represents an investment that signifies partial ownership in a company. Investors are concerned with two crucial aspects: the current price of their existing or potential investment and its projected selling price in the future. Predicting stock prices has always been of great interest to investors; however, it has proven to be a challenging task for researchers and analysts. The stock market is highly unpredictable, with numerous complex financial indicators. Consequently, financial analysts, researchers, and data scientists are continuously exploring analytical tools to uncover stock market patterns. In this study, historical stock price data is leveraged to predict the stock prices of selected US companies using a machine learning approach known as the Long Short-Term Memory (LSTM) Model, which is a specialized form of Recurrent Neural Network (RNN). The dataset comprises five years of AAPL and MSFT data obtained from Yahoo Finance, with consideration given to six relevant attributes. The LSTM model is employed to generate accurate and reliable predictions. The LSTM model holds several advantages in the realm of stock price prediction as it utilises historical stock price data to discern patterns and trends, enabling the forecasting of future price movements. This study focuses on employing the LSTM model to shed light on the potential for achieving precise stock price forecasts using machine learning techniques. Additionally, the Root Mean Square Error (RMSE) is employed as a supplementary performance measure alongside the LSTM model for stock price prediction.

Abstrak

Saham, biasanya dirujuk sebagai ekuiti, mewakili pelaburan yang menandakan pemilikan separa dalam syarikat. Pelabur bimbang dengan dua aspek penting: harga semasa pelaburan sedia ada atau berpotensi mereka dan harga jualan yang diunjurkan pada masa akan datang. Meramalkan harga saham sentiasa menarik minat pelabur; Walau bagaimanapun, ia telah terbukti menjadi tugas yang mencabar bagi penyelidik dan penganalisis. Pasaran saham sangat tidak dapat diramalkan, dengan banyak petunjuk kewangan yang kompleks. Akibatnya, penganalisis kewangan, penyelidik, dan saintis data terus meneroka alat analisis untuk mendedahkan corak pasaran saham. Dalam kajian ini, data harga saham sejarah dimanfaatkan untuk meramalkan harga saham syarikat AS terpilih menggunakan pendekatan pembelajaran mesin yang dikenali sebagai Model Memori Jangka Pendek Panjang (LSTM), yang merupakan bentuk khusus Rangkaian Neural Berulang (RNN). Set data terdiri daripada lima tahun data AAPL dan MSFT yang diperoleh daripada Yahoo Finance, dengan pertimbangan diberikan kepada enam atribut yang berkaitan. Model LSTM digunakan untuk menghasilkan ramalan yang tepat dan boleh dipercayai. Model LSTM memegang beberapa kelebihan dalam bidang ramalan harga saham kerana ia menggunakan data harga saham sejarah untuk membezakan corak dan trend, membolehkan ramalan pergerakan harga masa depan. Kajian ini memberi tumpuan kepada penggunaan model LSTM untuk memberi penerangan tentang potensi untuk mencapai ramalan harga saham yang tepat menggunakan teknik pembelajaran mesin. Selain itu, Ralat Root Mean Square (RMSE) digunakan sebagai langkah prestasi tambahan bersama model LSTM untuk ramalan harga saham.

Table of Contents

Table of Contents

List of Tables

List of Figures

List of Equations

Chapter 1: Introduction

1.1 Introduction

Every investor should consider two crucial aspects: the current price of their investment or the one they plan to buy, and its future selling price. However, investors consistently examine past pricing history and use it to influence their future investment decisions. Some investors avoid purchasing stocks or indices that have experienced rapid price increases because they expect a correction. Conversely, others avoid buying stocks that are in decline, assuming the downward trend will persist. The primary goal of stock price prediction is to generate substantial profits (Hagenau et al., 2013). Predicting the performance of the stock market poses a challenging task. Various factors, including physical and psychological influences, rational and irrational behaviour, and others, also affect these predictions. All these factors contribute to the dynamic and volatile nature of stock prices. Consequently, accurately predicting stock prices becomes extremely difficult (Yu & Yan, 2019). Due to the abundance of data available, there is ongoing exploration by financial analysts, researchers, and data scientists to uncover patterns in the stock market. The potential of data analysis to be a game changer in this field has been investigated. Numerous studies have focused on utilising machine learning in quantitative finance, which allows for predicting the prices of various assets, optimizing investment strategies, and other related processes (Moghar & Hamiche, 2020). Generally, machine learning refers to algorithmic techniques that employ computers to discover patterns solely based on data, without requiring explicit programming instructions. These techniques have the capability to unveil previously unnoticed patterns and insights, thereby enabling the generation of highly accurate forecasts (Ghosh et al., 2019). In the field of quantitative finance, there are numerous models available that can be combined with machine learning to predict the future value of assets. These models offer a way to integrate diverse sources of information, creating a unique and effective tool. Recent advancements in

machine learning techniques, such as neural networks, gradient boosted regression trees, support vector machines, and random forest, have emerged from the fusion of statistics and learning models. These algorithms are capable of capturing complex patterns characterized by non-linearity and uncovering relationships that are challenging to identify using linear algorithms. Moreover, these algorithms demonstrate superior effectiveness and handle multicollinearity better than linear regression techniques. The research paper "Trends and Applications of Machine Learning in Quantitative Finance" (Emerson et al., 2019) focuses primarily on the investment process, which is a fundamental component of finance. This encompasses predicting investment returns, modelling risks, and constructing portfolios. In this regard, a group of machine learning algorithms based on Recurrent Neural Network (RNN) proves to be valuable for forecasting and predicting financial market prices. Another study examines the accuracy of autoregressive integrated moving average (ARIMA) and long short-term memory (LSTM) as prediction algorithms for time series data. When applied to financial data, it was found that LSTM outperformed ARIMA by a significant margin (Siami Namin & Siami Namin, 2018). The objective of this project is to utilise a machine learning algorithm based on LSTM RNN to forecast stock prices in the US stock market and assess the performance of the Long Short-Term Memory (LSTM) Model.

1.2 Problem Statement

Investors have a keen interest in predicting stock prices as they seek profitable investment opportunities. However, many investors lack a reliable strategy to make informed decisions that yield significant profits. Consequently, there is a growing interest among investors to gain insights into the future state of the stock market, enabling them to make intelligent and lucrative investments. Nevertheless, predicting stock prices has proven to be a challenging task for researchers and analysts. The stock market is influenced by a multitude of

factors, including both psychological and physical elements, rational and irrational behaviour, investor sentiment, and more. The unpredictable nature of stock market fluctuations, coupled with the complexity of numerous financial indicators, further compounds the difficulty of accurate prediction. Consequently, there is a need to analyse vast amounts of data to identify patterns in the stock market. Financial analysts, researchers, and data scientists continue to explore analytics tools in their quest to uncover stock market patterns. The application of data analysis in this domain has the potential to bring about significant advancements. To predict future trends, researchers employ Machine Learning techniques using historical stock price data. Leveraging these Machine Learning approaches allows for the discovery of previously unnoticed patterns and insights, enabling the generation of highly accurate forecasts (Ghosh et al., 2019). For this project, the Long Short-Term Memory (LSTM) Model will be employed to predict stock prices. LSTM is a specialized variant of recurrent neural network (RNN) architecture designed to better capture temporal sequences and their long-term relationships compared to traditional RNNs (Van Houdt et al., 2020; Sak et al., 2014). RNNs, in general, are neural networks devised to overcome the limitations of conventional neural networks. They possess a "memory" that stores information about previous calculations. In essence, RNNs are a type of artificial neural network in which the connections between units form a directed graph along a sequence (Patel et al., 2018). These networks incorporate internal feedback loops that facilitate information retention. Theoretically, RNNs are capable of handling "long-term dependencies" effectively. Unfortunately, in practice, RNNs seem unable to effectively learn long-term dependencies, giving rise to what is known as the "Vanishing gradient problem." The neural network employs the gradient descent algorithm to update its weights. However, as the network progresses through lower layers, the gradients diminish significantly. Consequently, the network reaches a point where there is little room for improvement as the gradients remain almost constant. This adjustment adversely affects the network's output.

When the gradient difference becomes exceedingly small, the network fails to learn anything new, and the output remains unchanged. On the other hand, exploding gradients lead to unstable training due to fluctuating weights, while vanishing gradients result in excessively long training times and, in some cases, render training impossible (Tsun et al., 2019). Consequently, a network grappling with vanishing gradient problems fails to converge on an effective solution, rendering both scenarios undesirable in neural network training. Hence, it necessitates the development of novel training methods and architectures to tackle these challenges. In response, in 1997, Hochreiter and Schmidhuber introduced the Long Short-Term Memory (LSTM) Model, a specialized type of RNN that overcomes the limitations of traditional RNNs by effectively learning long-term dependencies. Over time, LSTMs have undergone refinements to address the challenge of handling long-term dependencies. The LSTM architecture tackles this issue by incorporating memory cells and gate units into neural networks (Tsun et al., 2019). LSTMs were specifically designed with a unique "memory cell" structure that enables the retention or discarding of information as it traverses through the network. Furthermore, LSTMs employ three distinct gates (input, forget, and output) to regulate the flow of data and determine what information to remember and what to forget. In a typical formulation, memory cells store information encountered in their cell state. When an input is received, the output is determined based on a combination of the cell state (representing previous information) and the updated cell state. As subsequent inputs are fed into the memory cell, the updated state and new input can be utilised to compute the updated output. With the ability to retain long-term dependencies, LSTMs are well-suited for tasks like time series prediction, such as forecasting a company's share price. Unlike traditional RNNs, LSTMs have gate mechanisms that govern the transfer of data, providing finer control over what information is stored and what is disregarded. This enhanced control allows LSTMs to effectively handle non-linear relationships present in financial data, thereby enabling more accurate predictions.

Moreover, LSTM models demonstrate proficiency in handling extensive datasets with numerous time-steps, leveraging historical data to make more precise forecasts regarding future trends. Consequently, LSTM emerges as a favourable choice for predicting the future performance and growth of a company's share price.

1.3 Scope

This project's scope is to model, tune and evaluate the performance of Long Short-Term Memory (LSTM) Model, which is a special type of Recurrent Neural Network, in predicting the stock prices of companies listed on the US stock markets based on data on the date, opening stock price, closing stock price, previous close price, lowest price and highest price. This project will use five years of data in the prediction phase. Some of the software and sources that will be used to conduct this project includes Yahoo Finance and RapidMiner.

1.4 Objectives

The objectives of this project are:

- To collect data of several public companies listed on the US stock exchange.
- To predict the stock prices of the selected companies using Long Short-Term Memory (LSTM).
- To evaluate the performance of the Long Short-Term Memory (LSTM) Model.

1.5 Brief Methodology

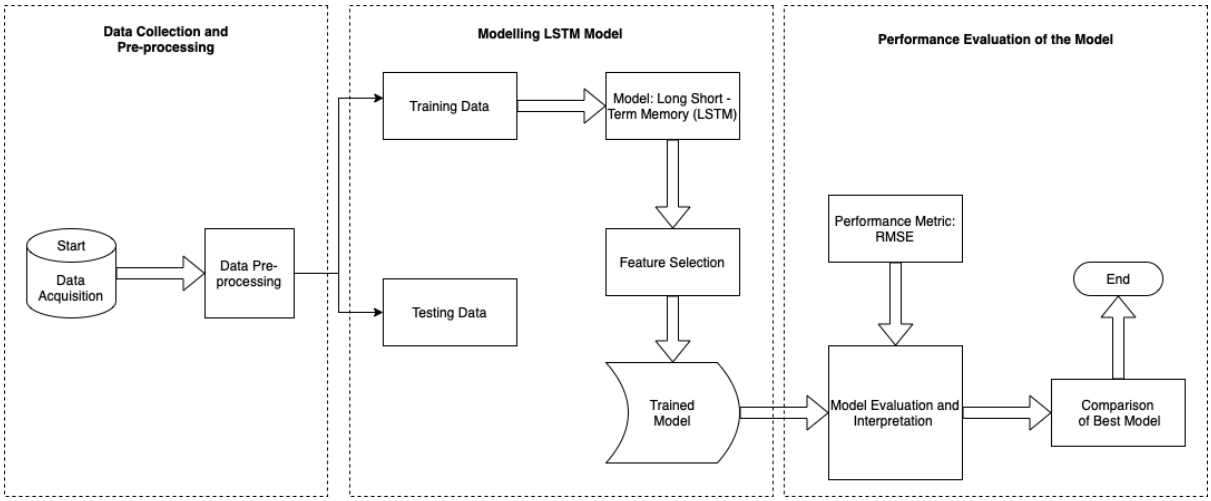

Figure 1.1: Framework of the study, starts from data collection to data pre-processing to splitting data and then modelling and training the LSTM model and lastly to model visualisation and performance evaluation

Phase 1: Data Collection and Pre-processing

For this project, the first phase is to collect 5 years' worth of data of past stock prices from selected companies using Yahoo Finance. Date, open share price, close share price, high share price, low share price and previous close prices are used.

The next part is data pre-processing. Before modelling the LSTM model, it is important to conduct data pre-processing. It is a crucial stage in machine learning research. Frequently, data collection procedures are poorly governed, resulting in out-of-range numbers, missing values, etc. Analysing data that has not been thoroughly examined for such issues can lead to incorrect conclusions. Before conducting an analysis, the representation and quality of the data are therefore of the utmost importance. Data pre-processing is the most crucial element of a machine learning project, particularly when dealing with computational data. Data pre-processing involves data cleaning, splitting data into training and testing set and featuring selection.

Phase 2: Modelling and Training

Moving to the subsequent phase involves the modelling and training of the LSTM model. The LSTM units are employed as foundational components for the layers of a recurrent neural network (RNN), commonly known as an LSTM network. These LSTM units enable RNNs to retain inputs over an extended period by leveraging a computer-like memory structure. Within an LSTM unit, information can be written, deleted, and read from this memory. Moreover, the LSTM memory cell encompasses three gates: the forget gate, input gate, and output gate. The forget gate determines when specific elements of the cell state should be replaced with more recent data. The input gate governs the conditions for storing or updating data in the cell state. Finally, the output gate determines the information transmitted to the subsequent network node (Nivethitha et al., 2019). In featuring selection process earlier, data attributes are chosen and are going to be fed to the LSTM model. Since this project utilise RapidMiner, there is no need to create the LSTM model from scratch since the platform offers pre-built LSTM model components that are specifically designed for machine learning applications like stock price prediction. Therefore, in this phase it would be more to configuring, optimizing, and training the existing LSTM model in RapidMiner. Modelling involves customizing the architecture and parameters of the LSTM model to suit the specific requirements of the predicting task and data. This includes making choices regarding the number of layers, neurons, activation functions, and other architectural elements. After configuring the model, the next step is tuning, which aims to optimize the model's performance. Tuning entails experimenting with various hyperparameter settings, such as learning rate, batch size, regularization techniques, and optimization algorithms. The goal is to identify the optimal combination of hyperparameters that generates the most accurate predictions. Once the modelling and tuning stages are complete, the LSTM model is ready for training. During training, the model is fed with the training data, enabling it to learn and adjust its internal

parameters to capture the underlying patterns and trends in the historical stock price data. This training process equips the model with the ability to make predictions based on the acquired patterns.

Phase 3: Performance Evaluation and Model Visualisation

In this phase, the trained LSTM model will be evaluated with the test set and the Root Mean Square Error (RMSE) metric is used to assess the model's performance. RMSE is a commonly used metric for evaluating the accuracy of a model's predictions. It measures the differences between actual and predicted values, known as residuals. Usually, it is beneficial to graphically depict the predicted prices. Stock price data will be represented as a 2D graph in RapidMiner where the current and predicted close price data will be compared. Predictions and existing data will be utilised to determine precision.

1.6 Significance of Project

This research on stock market prediction offers the potential advantage of significantly reducing losses. Many investors make the mistake of insufficient research before attempting to predict the market, leading them to use ineffective forecasting methods. They often rely on gut feelings or arbitrary estimates when investing in stocks, hoping for price increases and profits. However, by understanding and utilising appropriate prediction strategies, investors can minimize their losses. Making well-informed decisions using reliable stock market prediction tools has the potential to substantially increase profits.

Furthermore, stock market forecasting provides the benefit of consistent results. Given the high volatility of the stock market, there is no guarantee that investors will always be profitable, even if they employ various tactics and formulas to forecast the future. They may experience gains on some days and losses of equal magnitude on others. In such situations, consistency becomes crucial, regardless of the extent of gains and losses, in order to

consistently provide investors with favourable returns. Accurate predictions of the stock market help investors achieve the consistency necessary to generate higher returns in this volatile market.

1.7 Project Schedule

Figure 1.2 and 1.3 depicts the Gantt Chart for this project:

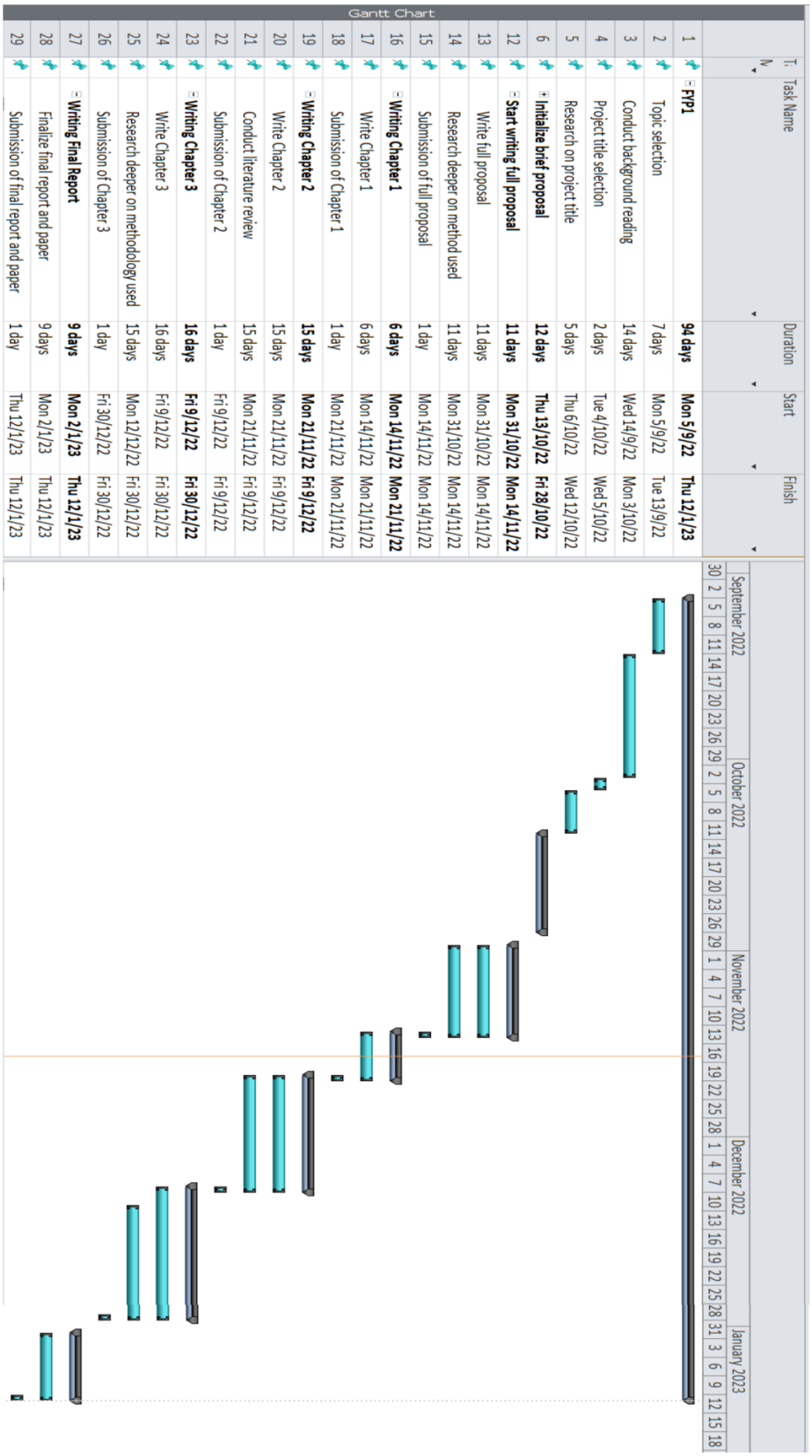

Figure 1.2: Gantt Chart for FYP 1

Task Mode	Task Name	Duration	Start	Finish
	Revised Structure of FYP1	15 days	Mon 3/20/23	Fri 4/7/23
	Writing Chapter 4	**28 days**	**Sat 4/8/23**	**Tue 5/16/23**
	Conduct Reading on Chapter 4 and Start Implementing the Topic	27 days	Sat 4/8/23	Mon 5/15/23
	Submission Of First Draft of Chapter 4	1 day	Tue 5/16/23	Tue 5/16/23
	Writing Chapter 5,6 and Abstract	**9 days**	**Wed 5/17/23**	**Mon 5/29/23**
	Amend Chapter 4	2 days	Wed 5/17/23	Thu 5/18/23
	Write Chapter 5,6 and Abstract	7 days	Fri 5/19/23	Sun 5/28/23
	Submission of First Draft of Chapter 5,6 and Abstract	0 days	Mon 5/29/23	Mon 5/29/23
	Writing Draft of Full Report	**10 days**	**Tue 5/30/23**	**Sat 6/10/23**
	Revise and Amend FYP Report	9 days	Tue 5/30/23	Fri 6/9/23
	Submission of First Draft of Full FYP Report	0 days	Sat 6/10/23	Sat 6/10/23
	Writing Final Report	**12 days**	**Sun 6/11/23**	**Sat 6/24/23**
	Amend and Finalize FYP Report	11 days	Sun 6/11/23	Fri 6/23/23
	Submission of Final Report	1 day	Sat 6/24/23	Sat 6/24/23
	FYP Symposium	**2 days**	**Thu 6/29/23**	**Fri 6/30/23**
	Presentation for FYP	2 days	Thu 6/29/23	Fri 6/30/23
	Amendment and Modification for FYP	**23 days**	**Thu 6/29/23**	**Sun 7/30/23**
	Amend FYP Report Based on Feedbacks	23 days	Thu 6/29/23	Sat 7/29/23
	Submission of Final Report	1 day	Sun 7/30/23	Sun 7/30/23

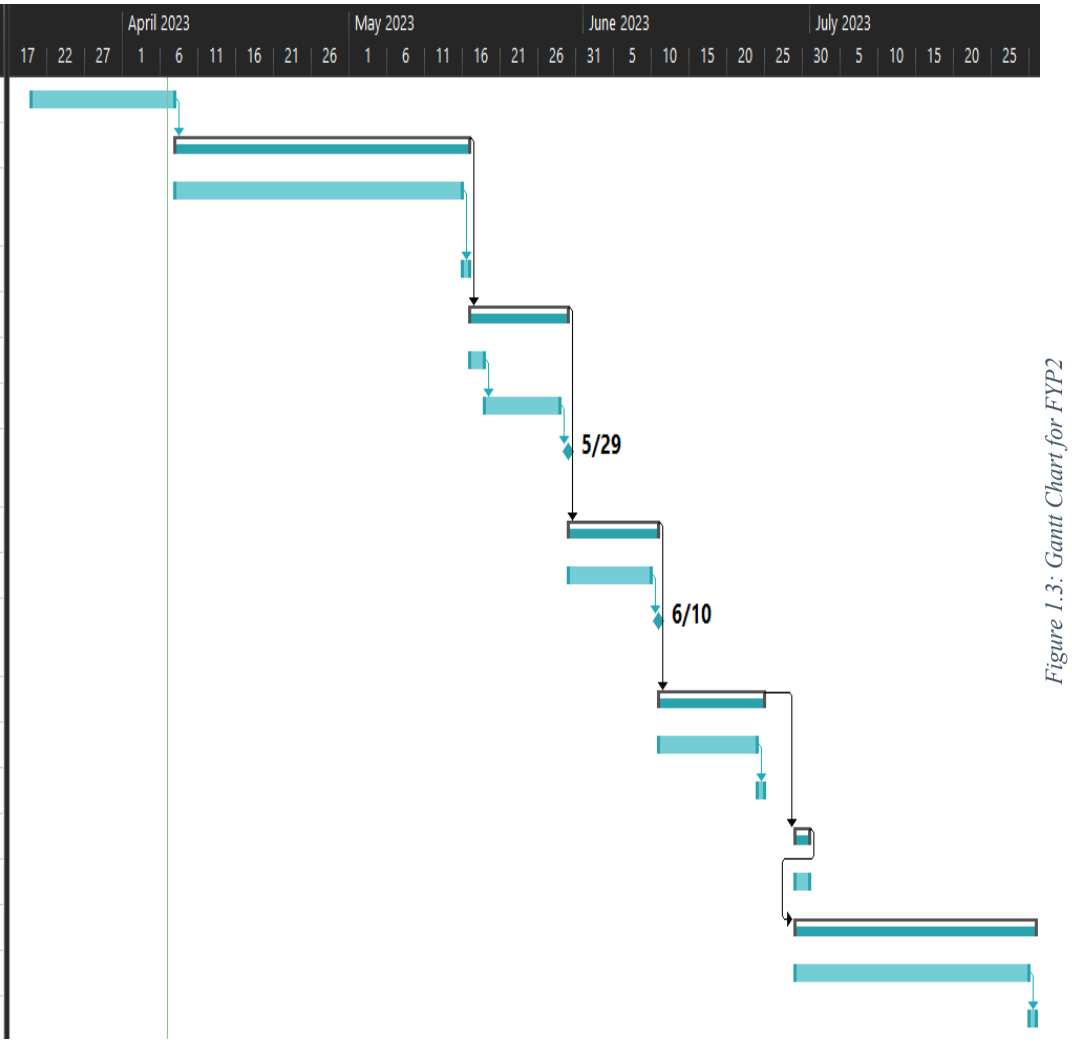

Figure 1.3: Gantt Chart for FYP2