Estimating Stock Price Based on Information From Financial Statement Using Machine Learning Approach

Mr. Thitikun Kunathananon

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science
Department of Computer Engineering
FACULTY OF ENGINEERING
Chulalongkorn University
Academic Year 2022

การประมาณมูลค่าของหุ้นด้วยข้อมูลจากงบการเงินโดยใช้การเรียนรู้ของเครื่อง

นายฐิติคุณ คุณะธนานนท์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2565

Thesis Title	Estimating Stock Price Based on Information From Financial Statement Using Machine Learning Approach
By	Mr. Thitikun Kunathananon
Field of Study	Computer Science
Thesis Advisor	Dr. PITTIPOL KANTAVAT

Accepted by the FACULTY OF ENGINEERING, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Science

..... Dean of the FACULTY OF ENGINEERING

(Professor SUPOT TEACHAVORASINSKUN)

THESIS COMMITTEE

..... Chairman

(Professor Dr. BOONSERM KIJSIRIKUL)

..... Thesis Advisor

(Dr. PITTIPOL KANTAVAT)

..... External Examiner

(Assistant Professor Dr. Kridsda Nimmanunta)

ฐิติคุณ คุณะธนานนท์ : การประมาณมูลค่าของหุ้นด้วยข้อมูลจากงบการเงินโดยใช้การเรียนรู้ของเครื่อง. (Estimating Stock Price Based on Information From Financial Statement Using Machine Learning Approach) อ.ที่ปรึกษาหลัก : ดร. พิตติพล คันธวัฒน์

การศึกษานี้นำเสนอเครื่องมือใหม่สำหรับนักลงทุนในตลาดหลักทรัพย์ที่สร้างขึ้นจากแบบจำลอง Long Short-Term Memory (LSTM) เพื่อทำนายราคาหุ้น โดยการวิเคราะห์รายงานทางการเงินและข้อมูลตลาดหลักทรัพย์อย่างมีประสิทธิภาพ LSTM ให้คำตอบที่รวดเร็ว ไม่มีผลอคติจากการทำนายและมีต้นทุนต่ำในการทำนายราคาหุ้น เพื่อเพิ่มกำไรสำหรับนักลงทุนและลดการขาดทุน ผลการศึกษาเชื่อมโยงว่า LSTM สามารถจับความสัมพันธ์ที่ซับซ้อนในข้อมูลและทำนายราคาหุ้นได้อย่างมีประสิทธิภาพ การวิจัยนี้เน้นให้ความสำคัญกับศักยภาพของ LSTM เป็นเครื่องมือที่มีประโยชน์และเป็นนวัตกรรมใหม่สำหรับนักลงทุนและสถาบันในตลาดหลักทรัพย์

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อนิสิต .....

ปีการศึกษา 2565 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6370076721 : MAJOR COMPUTER SCIENCE

KEYWORD: Long Short-Term Memory (LSTM), Machine Learning, Fundamental analysis, Financial statements, Stock price estimation

Thitikun Kunathananon : Estimating Stock Price Based on Information From Financial Statement Using Machine Learning Approach. Advisor: Dr. PITTIPOL KANTAVAT

This study introduces a new tool for stock market investors and institutions constructed from Long Short-Term Memory (LSTM) for predicting stock prices. By effectively analyzing financial statements and stock market data, LSTM provides a fast, unbiased, low-cost solution for stock price prediction, intending to increase profits for investors and reduce losses. The study results indicate that LSTM can maintain effectively captures complex relationships in the data and predicts stock prices. This research highlights the potential of LSTM as a valuable and innovative tool for investors and institutions in the stock market.

Field of Study:	Computer Science	Student's Signature .....
Academic Year:	2022	Advisor's Signature .....

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to my thesis advisor, Dr. Pittipol Kantavat, for his invaluable assistance and unwavering encouragement throughout the duration of this research. I am extremely appreciative of his instruction and guidance. Without the continuous support I have received from him, I would not have made it this far, and this thesis would not have been successfully completed. Furthermore, I extend my thanks to Prof. Dr. Boonserm Kijsirikul, Asst. Prof. Dr. Kridsda Nimmanunta, and others for their suggestions and assistance. Finally, I am deeply grateful to my parents and friends for their unwavering support throughout the entire research period.

Thitikun Kunathananon

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

วิทยานิพนธ์

(THESIS)

ชื่อเรื่อง (ภาษาไทย)	การประมาณมูลค่าของหุ้นด้วยข้อมูลจากงบการเงินโดยใช้การเรียนรู้ของเครื่อง
ชื่อเรื่อง (ภาษาอังกฤษ)	Estimating Stock Price Based on Information From Financial Statement Using Machine Learning Approach
เสนอโดย	นาย ฐิติคุณ คุณะธนานนท์
รหัสนิสิต	6370076721
หลักสูตร	วิทยาศาสตร์มหาบัณฑิต
ภาควิชา	วิศวกรรมคอมพิวเตอร์
คณะ	วิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
สถานที่ติดต่อ	ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย กรุงเทพ 10330
โทรศัพท์	(+66)2-218-6956, (+66)2-218-6957
อีเมล์	6370076721@student.chula.ac.th
อาจารย์ที่ปรึกษา	อ.ดร. พิตติพล คันธวัฒน์
คำสำคัญ (ภาษาไทย)	วิธีหน่วยความจำระยะสั้นระยะยาว, การเรียนรู้ของเครื่อง, การวิเคราะห์เชิงพื้นฐาน, งบการเงิน, การประมาณราคาหลักทรัพย์
คำสำคัญ (ภาษาอังกฤษ)	Long Short-Term Memory (LSTM), Machine learning, Fundamental analysis, Financial statements, Stock price estimation

Chapter1

Introduction

1.1 Introduction

The stock market is widely recognized as a venue for generating substantial returns through investment. These returns may manifest as dividends and capital gains resulting from price appreciation. However, it is important to note that stock prices are subject to volatility, contingent upon the underlying companies' financial performance.

There are two predominant approaches to stock investment analysis: technical and fundamental [1], [2]. Technical analysis involves the evaluation of stock prices and trading patterns to make predictions about future price movements. This approach considers factors such as previous prices, highs, and lows to gauge future price values. On the other hand, fundamental analysis utilizes financial statements, such as income statements and balance sheets, to analyze a company's financial performance and estimate its stock price. Financial ratios, such as the price-to-earnings (P/E) ratio, the price-to-book value (P/BV) ratio, the price-to-sales (P/S) ratio, and price-to-current cash flow (P/CF) ratio, are used to make these estimations. For example, Zdenko Prohaska and his colleagues [3] have developed a computational program to determine intrinsic value based on balance sheet information and a calculated equation.

In recent times, Traditional methods for stock price prediction can be subject to bias from various sources, including human bias, time bias, and cost bias. Analysts and traders may be influenced by their own biases or the biases of the companies they cover and may focus too much on recent trends or ignore longer-term patterns. Traditional methods can also be costly and time-consuming, creating a bias towards more established and well-funded firms. Machine Learning (ML) has gained widespread use in a diverse range of research areas, including the analysis of stock

prices through both technical and fundamental methods. For example of technical analysis, Rama Krishna and his team [4] used LSTM to forecast daily closing prices based on previous opening, closing, high, and low prices. Sandeep Patalay and his colleagues [5] utilized K-means clustering to predict stock prices and select the optimal portfolio fit. Similarly, in the realm of fundamental analysis, Shuo Han and Rung-Ching Chen [6] employed Support Vector Machine (SVM) to predict stock prices based on financial statements in comparison with the Outstanding Achievement Growth Rate. Junyoung Heo and Jin Yong Yang [7] also used SVM, but in conjunction with financial ratio data, to predict stock price trends and compare with expert estimations.

The utilization of Long Short-Term Memory (LSTM) for different applications is demonstrated. Zhen Sun and Shangmei Zhao [8] employed ML models, such as Multiple Linear Regression, Random Forest, and Long Short-Term Memory network (LSTM), to predict stock prices based on historical closing prices. Würtz and Göhner [9] proposed a model for analyzing driving styles based on LSTM RNNs, showing the potential of using this technique for analyzing and predicting human behavior in real-time applications. Benchaji et al. [10] proposed an LSTM RNN-based model for detecting credit card fraud, highlighting the potential of LSTM RNNs for fraud detection in financial transactions. Meenakshi and Mohamed Shanavas [11] proposed a shared input-based LSTM RNN model for predicting semantic similarity between sentences, demonstrating the potential of using LSTM RNNs for natural language processing tasks. Syarif et al. [12] proposed an LSTM RNN-based model for generating gamelan melodies, showcasing the potential of using LSTM RNNs for creative applications such as music generation and composition. Overall, these papers demonstrate the versatility of LSTM RNNs in various domains and highlight their potential for analyzing human behavior, processing natural language, and generating creative outputs.

LSTM was also applied to stock price prediction because the stock price is one of the sequential data. LSTM is more suitable for time series data than other Machine Learning models. In comparing stock price prediction performance between LSTM and Support Vector Machine (SVM), Zhixi Li and Vincent Tam [13] found that LSTM consistently demonstrated superior results. Adil Moghar and Mhanmed Hamiche [14] also applied LSTM to predict daily stock prices and obtained satisfactory results, allowing them to track the price evolution over time.

This research proposes a method for estimating stock value based on information from financial statements through the use of LSTM in a fundamental analysis approach. The financial statement data for SET50 and MAI in Thailand for the period 2006 to 2021 were obtained from the Stock Exchange of Thailand (SET), while the stock prices for SET50 and MAI from 2006 to 2021 were retrieved from Yahoo finance.

1.2 Research Objective

We propose a new method to estimate the intrinsic value of stock price based on information from financial statements using machine learning and expect it to assist investors and analysts in the future.

1.3 Scope of work

1. This research is based on information from a stock with data prior to 2005 from the Stock Exchange of Thailand (SET) using data from SET50 and MAI every quarter during 2005 to 2021.
2. We consider the stock in SET50 as the date of 4 Jan 2022.
3. We predict the percent of different average prices between the present quarter and next quarter of each stock in SET50.

1.4 Expected outcomes

1. To apply technology like machine to be capable in estimating the intrinsic value of stock price.
2. To become another option for analyzing stock prices in decision making for future investment.
3. To apply technology like machine with commercial and investment in Thailand.

1.5 Ordering in this proposal

This proposal separates into 7 parts as below.

1. Introduction
2. Literature Review
3. Related Theories
4. Methodology
5. Experiment Dataset
6. Experiment Result
7. Conclusion

Chapter 2

Literature Review

In this study, we conducted a review of related literature to gather insights on similar analysis and techniques applied in our research. Specifically, we analyzed two categories of literature: those that utilized fundamental analysis, including research combined with Machine Learning, and those that applied LSTM on time-series data for stock price prediction.

2.1 A computer program for fundamental analysis of stocks

In 2011, Zdenko Prohaska and his team [3] developed a computer program with the objective of computing stock prices based on data extracted from companies' balance sheets and income statements, as shown in Figure 1. The dataset encompassed various financial metrics, including price-earnings ratio, dividend yield, and debt-equity ratio, among others. Notably, this research primarily relied on equation-based estimations rather than employing machine learning techniques for predictive purposes.

Figure 1 Standard main menu (screenshot) from Literature Review [3]

2.2 A Model for Stock Price Predictions Using Deep Learning Techniques

Rama Krishna and his team [4] placed their primary focus on utilizing deep learning techniques to make predictions about stock prices. Their approach involved studying a wide range of relevant literature to evaluate existing methods critically. They developed a model that employed a combination of RNN-LSTM, a dynamic neural network, and an expansion of input variables. This approach provided a strong nonlinear framework for predicting various time series databases, including stock market indices, currency exchange rates, and electricity prices in deregulated energy markets. The team employed the tanh activation function in the output layer. Further, it enhanced the accuracy of stock market prediction by implementing a simple self-recurrent neural network using LSTM and keras ; the result is shown in Figure 2.

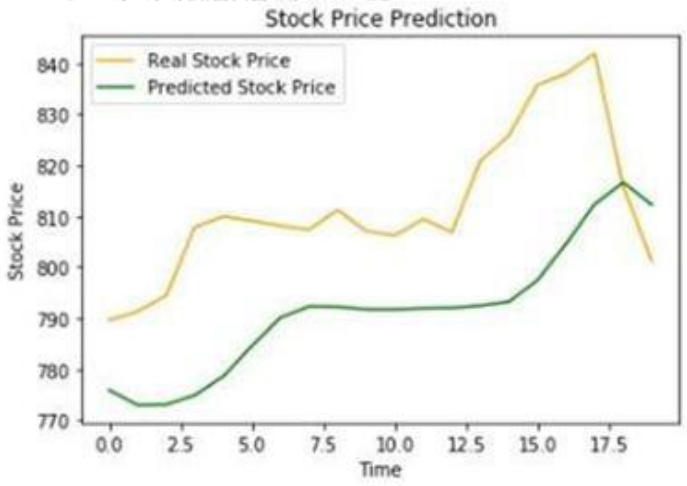


Figure 2 Comparison between actual values and predicted values from Literature Review [2]

2.3 Machine Learning in Stock Price Forecast

Zhen Sun and Shangmei Zhao [8] provided a comprehensive exploration of the application of machine learning techniques in predicting stock prices. The study highlights the limitations of traditional approaches and showcases the potential of advanced algorithms such as Multiple Linear Regression, random forests, and deep

learning models. The authors emphasize the significance of feature selection, preprocessing techniques, and evaluation metrics in developing accurate prediction models; examples of the results are shown in Table 1 and Table 2. Furthermore, the investigation of external factors demonstrates the value of incorporating additional sources of information to improve forecasting accuracy. Overall, this review contributes to the existing body of knowledge and offers valuable insights for researchers and practitioners in stock price prediction using machine learning methods.

Table 1 Forecasting result of the test within the sample based on closing price from Literature Review [8]

	RMSE	MAPE
Multiple Linear Regression	0.051	0.387
Random Forest	0.036	0.144
LSTM networks	1.420	2.585

Table 2 Forecasting result of the test outside the sample based on closing price from Literature Review [8]

	RMSE	MAPE
Multiple Linear Regression	0.236	0.423
Random Forest	0.283	0.561
LSTM networks	1.399	3.215

2.4 Using SVM with Financial Statement Analysis for Prediction of Stocks

In the study by Shuo Han and Rung-Ching Chen [6], a stock price estimation method using Support Vector Machine (SVM) was proposed. The authors used

financial statements data from the Shanghai and Shenzhen stock exchanges to predict stock prices. The data included Earnings Per Share (EPS), Book Value Per Share (BVPS), and Net Profit Growth Rate (NPGR). This paper focused on comparing the accuracy rates of eight experiments, as shown in Table 3, to aid in the decision-making process regarding selecting target stocks. Additionally, the effectiveness of the primary indices chosen for the prediction process can be evaluated by analyzing the accuracy rates. Suppose these indices do not contribute significantly to improving the accuracy rate. In that case, the paper suggests reevaluating or adjusting other financial indices from the financial statement until the optimal combination is identified for accurate stock prediction. This approach allows different stockholders to customize their predictions by selecting their preferred financial indices. The theoretical model depicted in Figure 3 visually represents the framework utilized in the study. The results of the study demonstrated that the model using financial statement data was more accurate in comparison to a base prediction model using only the Outstanding Achievement Growth Rate (OAGR) data provided by experts and professionals.

Table 3 Prediction modules from Literature Review [6]

Experiment	SVM	OAGR	EPS	BVPS	NPGR
1	√	√			
2	√	√	√		
3	√	√		√	
4	√	√			√
5	√	√	√	√	
6	√	√	√		√
7	√	√		√	√
8	√	√	√	√	√

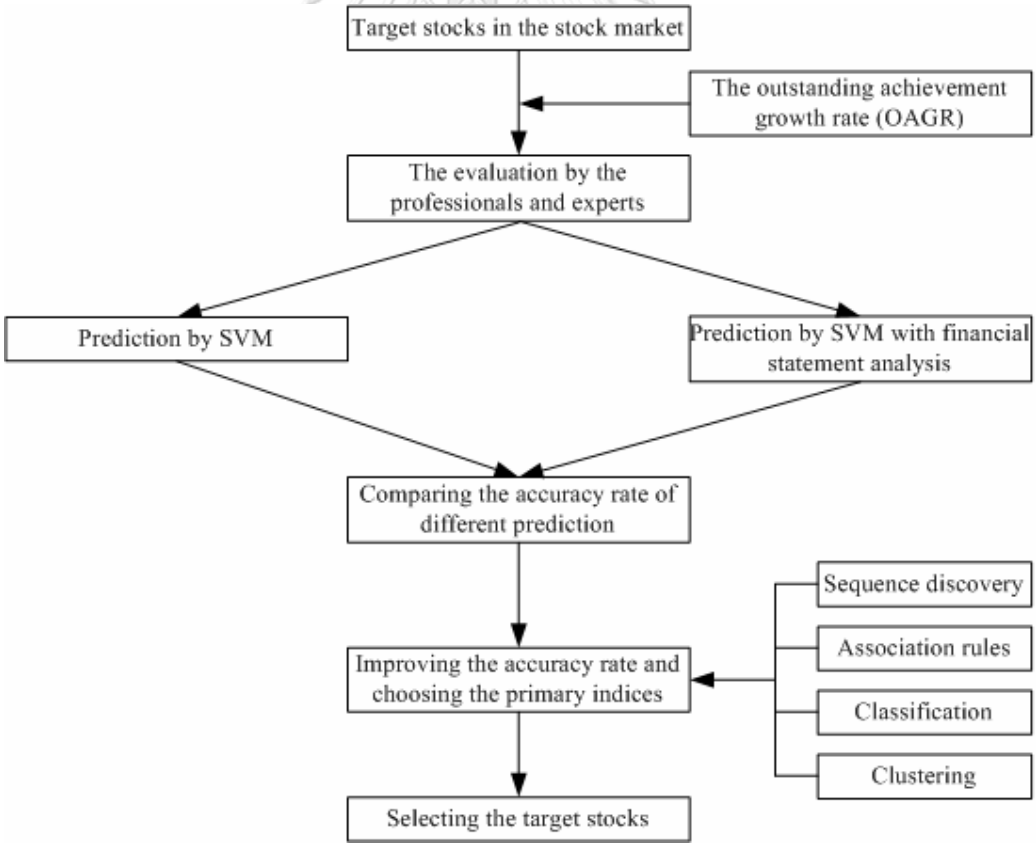


Figure 3 Theoretical model from Literature Review [6]

2.5 Stock Price Prediction Based on Financial Statements Using SVM

Junyoung Heo and Jin Yong Yang [7] focused on utilizing Support Vector Machines (SVM) for predicting stock prices based on financial statements in the KOSPI 200. The authors recognize the significance of financial statements as a crucial source of information for forecasting stock market trends. They address the limitations of traditional approaches and emphasize the potential of SVM in capturing complex patterns and relationships within financial data. Feature selection, preprocessing techniques, and model optimization are identified as essential factors in enhancing the accuracy of stock price prediction. Furthermore, the study investigates the influence of various financial ratios and indicators, including earnings per share (EPS), price-earnings ratio (P/E), and Net Profit Growth Rate (NPGR), extracted from financial statements on the predictive performance of SVM models, as a result, is shown in Figure 4. By providing insights into the effectiveness of SVM in leveraging financial statements for stock price prediction and highlighting relevant features and techniques, this research significantly contributes to the existing knowledge in the field.

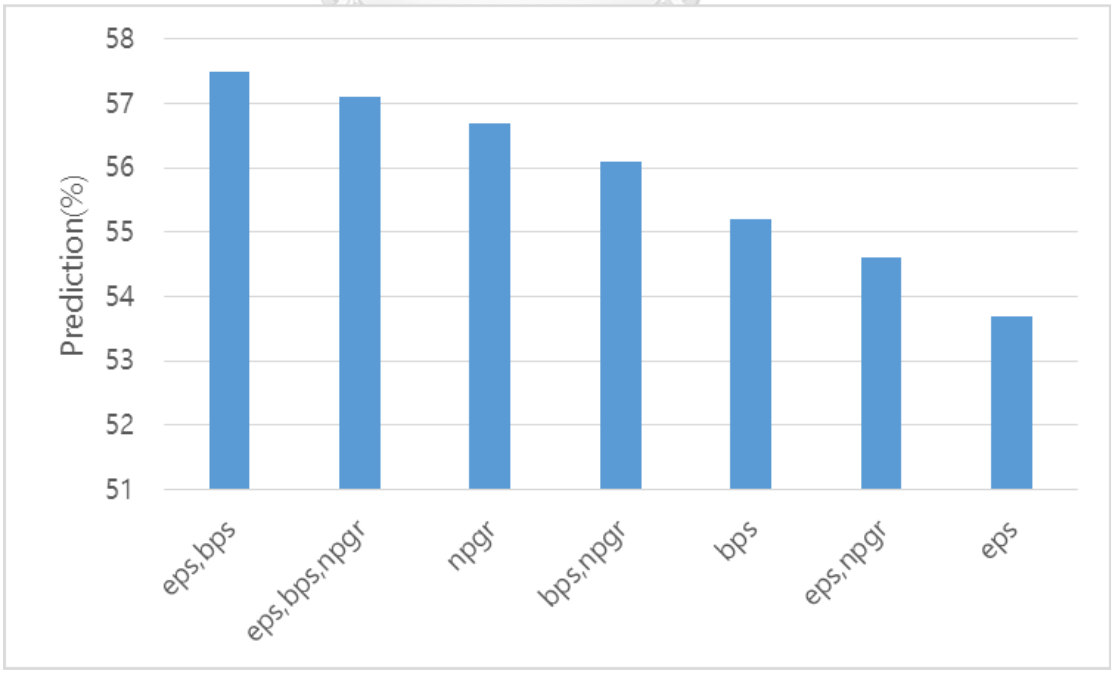


Figure 4 Prediction Result on SVM from Literature Review [7]

2.6 A comparative study of a recurrent neural network and support vector machine for predicting price movements of stocks of different volatilites

Zhixi Li and Vincent Tam [13] compared the performance of Recurrent Neural Networks (RNN) and Support Vector Machines (SVM) in predicting the price movements of stocks with varying levels of volatility. The authors acknowledge the challenges associated with predicting price movements in dynamic and volatile stock markets. They explore the potential of RNN and SVM as machine learning techniques capable of capturing the complexities and patterns within stock market data. The study highlights the importance of selecting appropriate input features and preprocessing techniques to improve the accuracy of predictions. Furthermore, the authors examine the impact of different volatilities on the performance of RNN and SVM models. By conducting a comparative analysis, this research provides valuable insights into the strengths and weaknesses of RNN and SVM in predicting price movements, catering to stocks with different levels of volatility. The results indicated that the PCA-LSTM and PLS-LSTM are the best performers among all the enhanced learning models for the low-volatility and high-volatility stocks respectively, as shown in Figure 5.

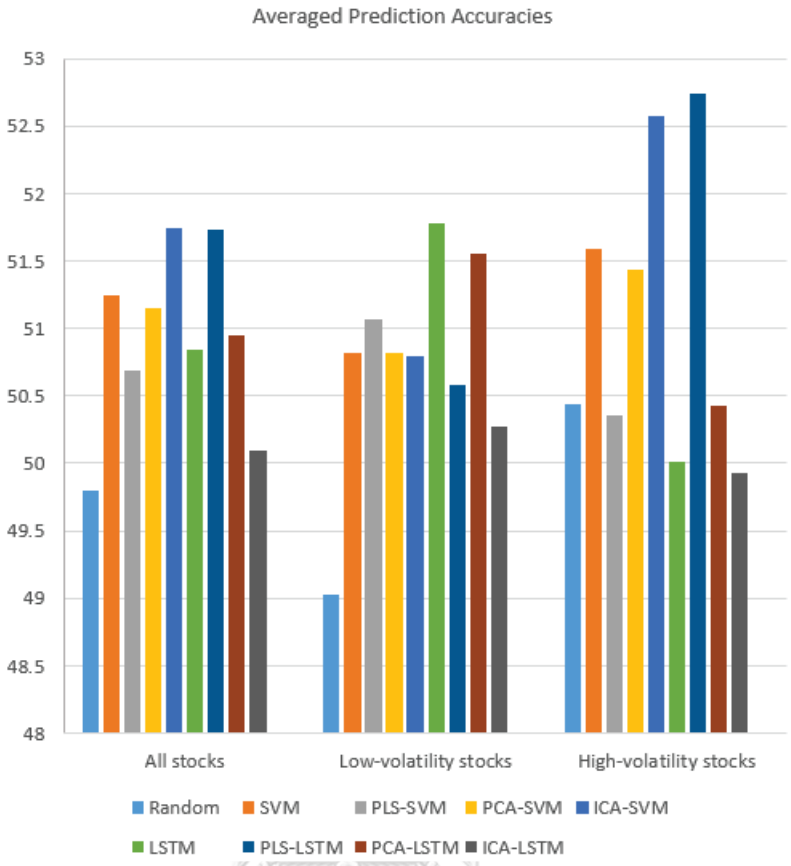


Figure 5 The Performance Comparison of Averaged Prediction Accuracies from Literature Review [13]

Chapter 3

Related Theories

In this chapter, we separate theories into four parts, 1. Linear Regression 2. Deep neural network(DNN) 3. Long Short-Term Memory (LSTM) 4. The error of prediction

3.1 Linear Regression [15]

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It is a widely used technique in many fields, including economics, finance, engineering, and social sciences. In this chapter, we will discuss the theory behind linear regression, its assumptions, and various techniques used to estimate the model. The basic linear regression model can be expressed below.

$$f(X) = \hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{P} X_j \hat{\beta}_j + \varepsilon \qquad (3.1)$$

where Y is the dependent variable, X_1, X_2, ..., X_j are the independent variables, β_0 is the intercept term, β_1, β_2, ..., β_j are the coefficients of the independent variables, and ε is the error term.

The objective of linear regression is to estimate the values of β_0, β_1, β_2, ..., β_j, such that the sum of squared errors (SSE) is minimized. SSE is defined as the sum of the squared differences between the actual and predicted values of the dependent variable.

There are two types of linear regression: simple linear regression and multiple linear regression. Simple linear regression involves only one independent variable, while multiple linear regression involves two or more independent variables.

Assumptions of linear regression include linearity, independence, homoscedasticity, and normality. Linearity assumes that there is a linear relationship

between the dependent variable and the independent variables. Independence assumes that the observations are independent of each other. Homoscedasticity assumes that the variance of the errors is constant across all levels of the independent variables. Normality assumes that the errors are normally distributed.

There are various techniques used to estimate the coefficients of the linear regression model, including the ordinary least squares (OLS) method, maximum likelihood estimation (MLE), and Bayesian methods. The OLS method is the most commonly used method for estimating the coefficients.

In conclusion, linear regression is a powerful statistical tool used to model the relationship between a dependent variable and one or more independent variables. It has many applications in various fields and is a fundamental technique in statistical analysis. The assumptions of linearity, independence, homoscedasticity, and normality are critical in ensuring the accuracy of the model, and various techniques are used to estimate the coefficients of the model.

3.2 Deep neural network (DNN) [16]

Deep neural networks (DNNs) are a type of artificial neural network that has gained popularity in recent years for their ability to learn complex patterns and features from large datasets.

DNNs are composed of multiple layers of artificial neurons, with each layer processing the output of the previous layer, as shown in Figure 6. The input layer receives the input data, and the output layer produces the output of the model. The intermediate layers are called hidden layers, and they extract features from the input data.

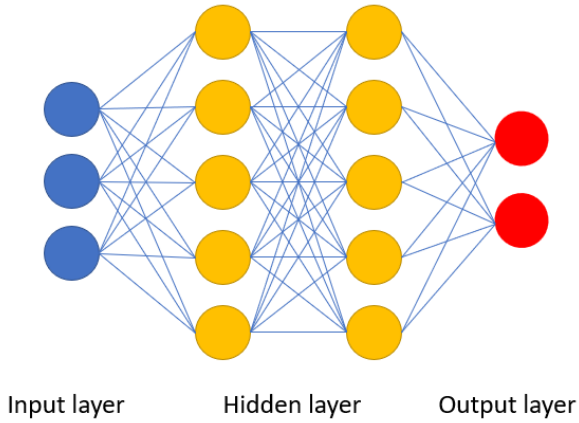


Figure 6 Artificial Neuron Networks

The neurons in a DNN are organized into interconnected layers, with each neuron receiving inputs from the previous layer and producing an output for the next layer. The inputs to a neuron are weighted, and a bias term is added to the weighted sum. The weighted sum is then passed through an activation function, which introduces nonlinearity into the model.

Let x be the input data to the DNN, y be the output, and e and b be the weights and biases, respectively. The output of a neuron in the DNN is calculated as follows:

$$\hat{y} = \phi(\overline{X} \cdot \overline{W}) = \phi(\sum_{j=1}^{d} w_j x_j) \quad (3.2)$$

Where ϕ is the activation function, and a is the output of the neuron.

The training process of a DNN involves updating the weights and biases of the neurons in the network to minimize the difference between the predicted outputs and the actual outputs. This is done using an optimization algorithm such as gradient descent, which iteratively adjusts the weights and biases to minimize the loss function.

There are several variations of DNNs, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and autoencoders. CNNs are commonly

used for image recognition and processing tasks, while RNNs are used for sequential data processing tasks such as natural language processing. Autoencoders are used for unsupervised learning tasks such as dimensionality reduction.

The success of DNNs is attributed to their ability to learn high-level abstractions from raw data, their ability to handle large datasets, and their ability to parallelize computations on GPUs. However, training DNNs can be computationally expensive and time-consuming, and overfitting can be a significant problem if the model is too complex or the training dataset is too small.

In conclusion, DNNs are a powerful tool for learning complex patterns and features from large datasets. Their architecture and training process enable them to learn hierarchical representations of data, and their success has led to their widespread adoption in various fields such as computer vision, natural language processing, and speech recognition. However, their training can be computationally expensive, and overfitting can be a significant problem.

3.3 Long Short-Term Memory (LSTM) [17]

Long Short-Term Memory (LSTM) is a type of recurrent neural network that has gained popularity in recent years due to its ability to handle long-term dependencies in sequential data. Unlike traditional recurrent neural networks, LSTMs have a memory cell that can store information for an extended period of time, allowing them to remember important information even when there are long gaps between relevant inputs. As a result, LSTMs have been used in a wide range of applications, including natural language processing, speech recognition, and time series analysis.

LSTM was introduced by Hochreiter and Schmidhuber in 1997 as a solution to the vanishing gradient problem in RNNs. The vanishing gradient problem occurs when the gradient of the loss function with respect to the weights of the RNN becomes

very small as it propagates through time, making it difficult to train the RNN to remember long-term dependencies.

LSTM solves this problem by introducing a memory cell and three gates that control the flow of information into and out of the memory cell. The three gates are the input gate, forget gate, and output gate, and they are controlled by sigmoid activation functions that output values between 0 and 1.

The input gate controls the flow of information from the input and previous hidden state into the memory cell. It is defined by the following equations:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3.3}$$

where i_t is the input gate activation, x_t is the input at time t, h_{t-1} is the hidden state at time t-1, W_i and U_i are the weight matrices, and b_i is the bias vector.

The forget gate controls the flow of information from the memory cell into the hidden state. It determines what information to keep or discard from the memory cell. It is defined by the following equations:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_r) \tag{3.4}$$

where f_t is the forget gate activation, W_f and U_f are the weight matrices, and b_f is the bias vector.

The output gate controls the flow of information from the memory cell to the hidden state and output. It is defined by the following equations:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{3.5}$$

where o_t is the output gate activation, W_o and U_o are the weight matrices, and b_o is the bias vector.

The memory cell is updated by the following equations:

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{3.6}$$

where C_t is the memory cell, W_c are the weight matrices, and b_c is the bias vector.

The output of the LSTM at time t is given by:

$$h_t = o_t * \tanh(C_t) \tag{3.7}$$

where h_t is the hidden state at time t.

During training, the LSTM is trained to minimize the difference between the predicted output and the actual output using an optimization algorithm such as gradient descent. Figure 7 illustrates the structure of the LSTM model.

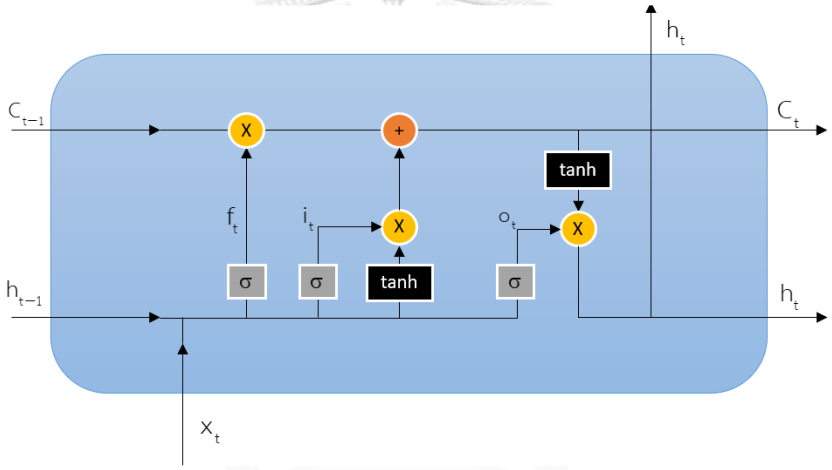


Figure 7 The architecture of LSTM [17]

In conclusion, LSTM is a powerful type of RNN that is designed to overcome the vanishing gradient problem by introducing a memory cell and three gates that control the flow of information into and out of the memory cell. The gates are controlled by sigmoid activation functions, and the memory cell is updated by a combination of the input and forget gates. The output of the LSTM is given by the output gate and the memory cell.

3.4 The error of prediction [18]

The error between prediction value and real value can be calculated from various indicators. Popular indicators such as Mean

Absolute Error, Mean Squared Error, and Root Mean Squared Error up to a type of prediction, they formulate shown as below.

$$MAE = \frac{\sum_{i=1}^{n} | y_i - \widehat{y}_i |}{n} \tag{3.8}$$

Let MAE is Mean Absolute Error

y_i is the real value position i

$\widehat{y}_i$ is the prediction value position i

n is a quantity of data

Chapter 4

Methodology and Dataset

4.1 Methodology

The proposed model in this study employs LSTM, as depicted in Fig. 8, to analyze financial statement data in quarter intervals as time series data. The output of the LSTM model is the predicted percentage change in stock price from the previous quarter. The model's accuracy is evaluated using the MAE cost function, which calculates the percentage difference between the actual and predicted values.

We used financial statement data and macroeconomic indicators as inputs for our LSTM model to predict stock price changes. Inputs include Total assets, Total equity, Total liability, Earning per share, Closing price, Price to Earning Per Share, Market Capitalization, SET index, Interest Rate, Inflection Rate, and Gross Domestic Product. The model's output is the prediction of percent change in stock price from the previous quarter, and we evaluated it using the MAE.

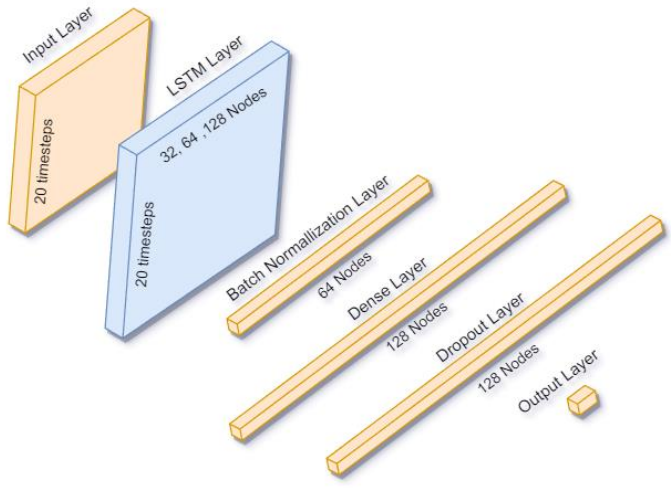


Figure 8 Our proposed LSTM architecture

4.2 Data Description

In this study, we have selected a stock with data prior to 2005 from The Securities Exchange of Thailand (SET) and the Market for Alternative Investment (MAI). The data from SET has been collected on a quarterly basis and encompasses individual and market information. The Bank of Thailand has provided yearly data that includes inflection and interest rates, while the National Economic and Social Development Board has provided yearly Gross Domestic Product (GDP) data. The model utilized in this study outputs the daily adjusted close price, which is adjusted for stock splits, dividends, and capital gain distributions, and obtained from Yahoo Finance. The input factors and descriptions are shown in Table 4.

Table 4 Input Factor and Description

Factor	Description
M_TOTAL_ASSET	Total assets of each stock
M_TOTAL_EQUITY	Total equity of each stock
M_TOTAL_LIABILITY	Total liability of each stock
Z_PAR	Par of each stock
M_NET_PROFIT	Net profit of each stock
R_EPS	Earning per share of each stock
R_PE	Price to Earning Per Share of each stock
R_PB	Price to Book Value Ratio of each stock
M_BOOK_VALUE	Book value of each stock
R_TURNOVER	Turn over rate of each stock
M_MKT_CAP	Market Capitalization of each stock
R_INDEX_CLOSE	Closing SET index
M_MKT_PE	Market price to earnings
M_MKT_PBV	Market price to book value
M_MKT_YIELD	Market yield
M_MKT_CAP_SET	Market Capitalization of SET

Factor	Description
Infection rate	Inflection Rate
Interest	Interest Rate
GDP	Gross Domestic Product

4.3 Data Splitting

The data collected between 2005 and 2021 from the AGRO, FINANCIAL, POPCORN, RESOURCE, SERVICE, and TECH industries was partitioned into three separate segments to train, validate, and test the model. The list of stocks for SET 50 includes ADVANC, AOT, BANPU, BBL, BDMS, BH, BTS, CPALL, CPF, CPN, DTAC, EGCO, HMPRO, INTUCH, IRPC, KBANK, KCE, KTB, KTC, LH, MINT, PTT, PTTEP, RATCH, SCC, TOP, TTB, TU, and TRUE. Similarly, the list of stocks for MAI comprises ACAP, AF, BOL, ROOK, CIG, CMO, CPR, DV8, IRCP, KASET, MBAX, NEWS, PICO, PPM, PROUD, SALEE, SLM, SSS, SWC, TAPAC, TMW, TNH, TRT, UBIS, UEC, UKEM, UMS, and YUASA.

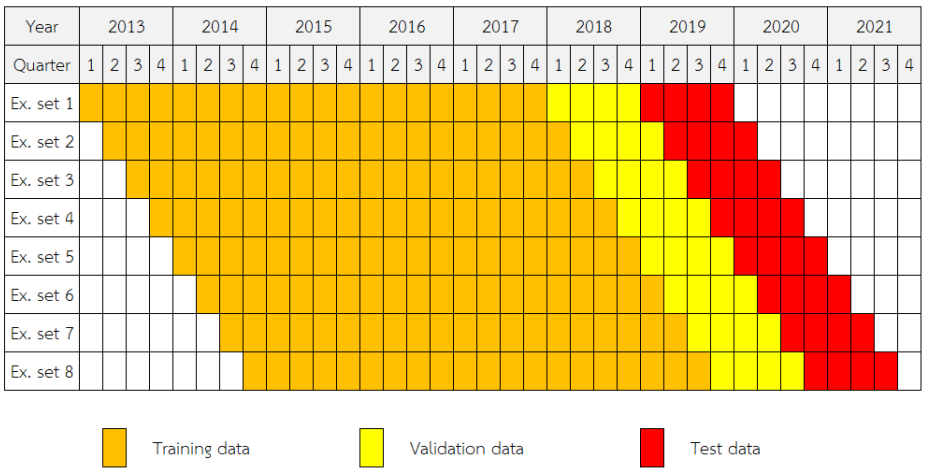


Figure 9 Separation of data for experiment

The data has been separated into twenty quarters (five years) of training data, four quarters (one year following the training period) of validation data, and four quarters (one year following the validation period) of test data, as indicated in Fig. 9.

4.4 Data Output

For the stock price, we use an adjusted close price which changed for splits, dividends, and capital gain distributions. The closed daily adjusted price is calculated to find the average price by quarter as (8.1).

$$Avg = \frac{\sum_{i=1}^{n} x}{n} \tag{8.1}$$

To provide meaningful and interpretable outputs, the price change is labeled as a percentage as (8.2). Finally, since the predictions are also represented as percentages, we evaluate the performance of the model using the MAE. This evaluation metric allows us to determine the percentage error of the model and provides valuable information on its accuracy and reliability.

$$Output = \frac{Stock\,Price_{n+1} - Stock\,Price_{n}}{Stock\,Price_{n}} \tag{8.2}$$

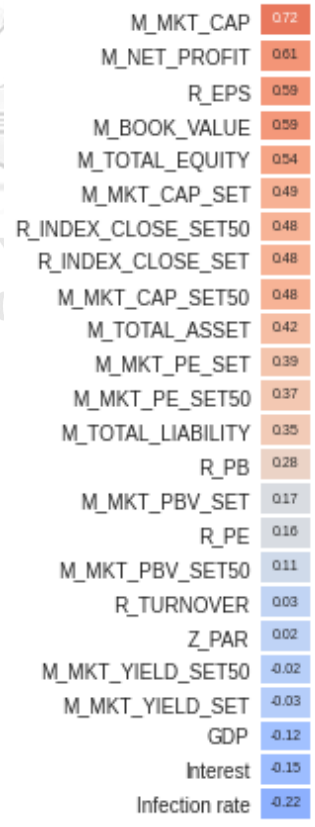
Chapter 5

Experiment Result

In this chapter, we show our results from the experiment. First is the correlation of factors, showing the relation between two variables. Second is the prediction performance compares between models. Final is the prediction performance showing separate errors in the considered industry.

5.1 The correlation between the various factors and stock price

In Fig. 10, the correlation between the various factors and the stock price in SET50 is depicted. The figure shows that the stock price is positively correlated with variables such as Equity, Net profit, Earning per Share (EPS), Market Capitalization, and SET Index. This indicates that an increase in the former variables and a decrease in



the latter variables would result in an increase in the stock price.

Figure 10 The correlation between the various factors and the stock price

5.2 The prediction performance comparison between models

We conducted our experiment using three models: linear regression, Deep neural network (DNN), and LSTM. The result of our experiments shows in Table 5, that the LSTM model performed the best for predicting stock prices in SET50, while the Deep neural network model, configured as Fig. 11, performs similarly well for predicting stock prices in MAI. Overall, the LSTM model is considered to be the most suitable for forecasting stock prices. In both cases, the models were able to effectively capture the relationships between the input features and the stock prices, leading to accurate predictions.

Table 5 MAE in Linear regression, DNN AND LSTM

Model	SET50	MAI
Linear regression	10.8037	16.2857
DNN	10.3686	15.9394
LSTM	9.6848	15.9683

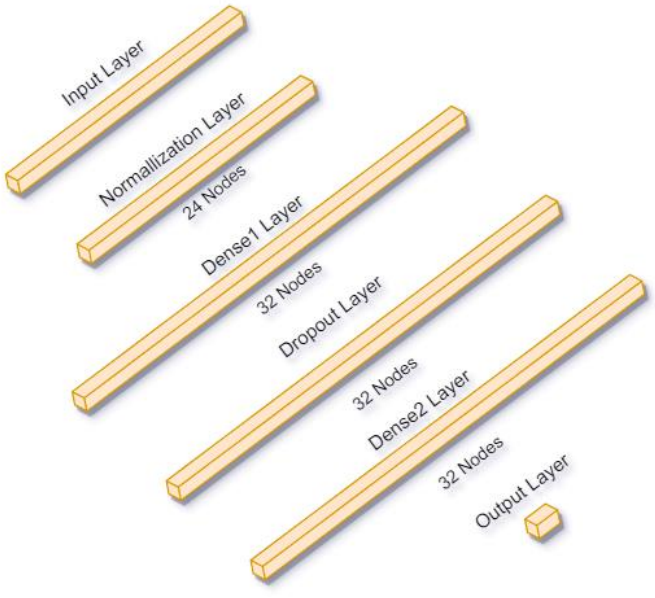


Figure 11 Our architecture of DNN

5.3 The prediction performance by industry

As previously mentioned, the LSTM model is considered to be the optimal choice for stock price prediction. We trained LSTM model using training data and configuration detailed in Table 6. The performance of the model was evaluated using the MAE as described in (6). LSTM layer was optimized by performing a grid search for the number of nodes, with the options of 32, 64, and 128, using validation data. The configuration with the lowest MAE was selected by evaluating its performance on the validation data. Finally, the overall MAE for stocks from the SET and MAI indices were calculated using test data.

The MAEs of the test data for each industry from SET50 and MAI are shown in Table 7 and Table 8. The industry in SET50 and MAI was officially separated from SET [19], [20].

Table 6 Configuration of LSTM

Parameters	**Configurations**
Batch size	128
Optimizer	Adam
Learning rate	1×10^{-3} with decay factor 0.95
Stopping algorithms	500 epochs

Table 7 The Prediction Performance of SET50

Industry	**MAE**
Overall SET50	9.6849
AGRO	8.4667
FINCIAL	11.4178

Industry	MAE
POPCORN	7.3147
RESOURC	9.0209
SERVICE	8.8628
TECH	12.0783

Table 8 The Prediction Performance of MAI

Industry	MAE
Overall MAI	15.9683
AGRO	11.0600
FINCIAL	16.5924
POPCORN	22.2104
RESOURC	12.1313
SERVICE	14.8421
TECH	27.2657

5.4 Discussion

This superiority of LSTM over linear regression and DNN can be attributed to its ability to capture the time-dependent relationships in the data. Unlike linear regression and DNN, LSTM models have the ability to maintain when processing time series data, which enables the neural network to effectively capture the complex relationships between financial statement data and stock prices. These relationships, which can be impacted by past events, current conditions, and future trends, are

critical in accurately predicting stock prices and making informed investment decisions. The study found that the performance of stock price predictions for SET50 was more accurate than for MAI. The LSTM model had an average MAE of 9.68% for SET50 and 15.96% for MAI in terms of the percent change in stock prices.

Various positive factors could affect the performance of prediction models for stock market indices, including the liquidity of the index, industry composition, and data quality. Negative factors for stock price prediction may vary based on macroeconomic factors and company-specific issues such as declining revenues, increasing debt levels, regulatory changes, weak corporate governance, and poor management decisions. The SET may have better prediction performance than MAI due to larger, more actively traded stocks, more liquidity, and its role as a barometer of the Thai economy.

The analysis of the LSTM model's performance in predicting stock prices across different industries reveals interesting findings. Specifically, when examining the AGRO, RESOURC, and SERVICE industry, the LSTM model demonstrates a superior performance compared to the average MAE in both the SET50 and MAI indices. This implies that the LSTM model exhibits a higher accuracy in forecasting stock prices within these sectors. On the other hand, the model's performance in predicting stock prices in the FINANCIAL and TECH industries shows a contrasting trend. In these sectors, the LSTM model exhibits MAE values higher than the average observed in both the SET50 and MAI indices. These results suggest that the LSTM model might face challenges or limitations when forecasting stock prices in the FINANCIAL and TECH sectors. Further analysis and investigation are necessary to understand the underlying reasons behind these divergent performances and explore potential strategies for improvement.

Chapter 6

Conclusion

In conclusion, this study has demonstrated the potential of Long Short-Term Memory (LSTM) as a valuable tool for predicting stock prices in the stock market. By effectively analyzing financial statements and stock market data, LSTM models offer several advantages, including the ability to capture long-term dependencies, handle noisy data, reduce human bias, and provide low-cost solutions for investors and institutions to increase profits and reduce losses. Our findings indicate that LSTM outperforms other models, such as linear regression and DNN, in predicting stock prices, particularly for stocks in SET50. Additionally, we found that the performance of LSTM may vary among different industries, with some industries exhibiting higher prediction performance than others. It is important to note that our study results are limited by the data and the period in which it was collected, and future research may focus on improving the accuracy of the model by considering additional factors and incorporating more recent data. Nonetheless, our study highlights the potential of advanced techniques such as LSTM in financial prediction and provides a foundation for further exploration in this field. Overall, this research underscores the importance of utilizing innovative tools such as LSTM to enhance financial decision-making in the stock market.

Appendix

Appendix A

Result of SET50 price prediction from the LSTM model.

Stock in AGRO	MAE
CPF	7.403228
MINT	9.997531
TU	7.999396

Stock in FINCIAL	MAE
BBL	6.877397
KBANK	10.07945
KTB	9.424863
KTC	20.32331
TTB	10.38391

Stock in POPCORN	MAE
CPN	7.974479
LH	7.475497
SCC	6.494213

Stock in RESOURC	MAE
BANPU	13.38512
EGCO	7.787029
IRPC	10.79715
PTT	7.064859
PTTEP	9.482544
RATCH	6.17376
TOP	8.4558

Stock in SERVICE	MAE
AOT	10.25638
BDMS	8.537809
BH	9.556235
BTS	7.152931
CPALL	8.464057
HMPRO	9.114387

Stock in TECH	MAE
ADVANC	8.266741
DTAC	11.31367
INTUCH	8.225502
KCE	17.97928
TRUE	14.49992
Grand Total	9.684858

Appendix B

Result of MAI price prediction from the LSTM model.

Stock in AGRO	MAE
KASET	11.06005

Stock in FINCIAL	MAE
ACAP	25.95603
AF	8.36966
BROOK	15.45153

Stock in POPCORN	MAE
PROUD	18.05375
PSG	35.00865
SSS	22.96119
TAPAC	25.61619

Stock in RESOURC	MAE
TRT	7.685114
UMS	16.45401

Stock in SERVICE	MAE
BOL	11.23122
CMO	13.3306
DV8	19.53321
NEWS	21.82109
PICO	12.77254
SLM	20.59235
TNH	9.049526

Stock in TECH	MAE
IRCP	27.2657

Appendix C

Result of SET50 and MAI price prediction from the linear regression and DNN model.

Model	SET50	MAI
Linear regression	10.8037	16.2857
DNN	10.3686	15.9394

REFERENCES

1. Nti, I.K., A.F. Adekoya, and B.A. Weyori, *A systematic review of fundamental and technical analysis of stock market predictions.* Artificial Intelligence Review, 2020. **53**: p. 3007-3057.
2. Eiamkanitchat, N., T. Moontuy, and S. Ramingwong, *Fundamental analysis and technical analysis integrated system for stock filtration.* Cluster Computing, 2017. **20**: p. 883-894.
3. Prohaska, Z., I. Uroda, and S. Suljić, *SP — A computer program for fundamental analysis of stocks*, in *2011 Proceedings of the 34th International Convention MIPRO*. 2011, IEEE.
4. Krishna, V.R., et al., *A Model for Stock Price Predictions Using Deep Learning Techniques.* International Journal of Advanced Trends in Computer Science and Engineering, 2020. **9**: p. 8266-8271.
5. Patalaya, S. and M.R. Bandlamudi, *Stock Price Prediction and Portfolio Selection Using Artificial Intelligence.* Asia Pacific Journal of Information Systems, 2020. **30**: p. 31-52.
6. Han, S. and R.-C. Chen, *Using SVM with Financial Statement Analysis for Prediction of Stocks.* Communications of the IIMA, 2007. **7**(4): p. 63-72.
7. Heo, J. and J.Y. Yang, *Stock Price Prediction Based on Financial Statements Using SVM.* International Journal of Hybrid Information Technology, 2016. **9**: p. 57-66.
8. Sun, Z. and S. Zhao, *Machine Learning in Stock Price Forecast*, in *E3S Web of Conferences 214*. 2020.
9. Würtz, S. and U. Göhner, *Driving Style Analysis Using Recurrent Neural Networks with LSTM Cells.* Journal of Advances in Information Technology, 2020. **11**.
10. Benchaji, I., S. Douzi, and B.E. Ouahidi, *Credit Card Fraud Detection Model Based on LSTM Recurrent Neural Networks.* Journal of Advances in Information Technology, 2021. **12**.
11. Meenakshi, D. and A.R.M. Shanavas, *Novel Shared Input Based LSTM for*

Semantic Similarity Prediction. Journal of Advances in Information Technology, 2022. **13**.

12. Syarif, A.M., et al., *Gamelan Melody Generation Using LSTM Networks Controlled by Composition Meter Rules and Special Notes.* Journal of Advances in Information Technology, 2023. **14**.
13. Li, Z. and V. Tam, *A comparative study of a recurrent neural network and support vector machine for predicting price movements of stocks of different volatilites.* IEEE Symposium Series on Computational Intelligence, 2017.
14. Moghar, A. and M. Hamiche, *Stock Market Prediction Using LSTM Recurrent Neural Network.* International Workshop on Statistical Methods and Artificial Intelligence, 2020.
15. Hastie, T., R. Tibshirani, and J. Friedman, *Data Mining and Inference and Prediction, Springer Series in Statistics*. 2008: Springer.
16. Aggarwal, C.C., *Neural Networks and Deep Learning*. 2008: Springer.
17. Hochreiter, S. and J. Schmidhuber, *Long Short-Term Memory*. 1997: Neural computation.
18. Sammut, C. and G.I. Webb, *Encyclopedia of Machine Learning*. 2010: Springer.
19. Thailand, T.S.E.o. *Mai Index*. 2021 [cited 2023 19 Jan 2023]; Available from: https://www.set.or.th/th/market/index/mai/profile.
20. Thailand, T.S.E.o. *SET Industry Group Index and Sector Index*. 2021 [cited 2023 16 Jan]; Available from: https://www.set.or.th/th/market/index/set/industry-sector-profile.

VITA

NAME	ฐิติคุณ คุณะธนานนท์
DATE OF BIRTH	24 สิงหาคม 2537
PLACE OF BIRTH	กรุงเทพ ประเทศไทย
INSTITUTIONS ATTENDED	วิศวกรรมศาสตรบัณฑิต จุฬาลงกรณ์มหาวิทยาลัย