Explainable Stock Price Prediction Using Technical Indicators With Short Thai Textual Information

Mr. Kittisak Prachyachuwong

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering in Computer Engineering
Department of Computer Engineering
FACULTY OF ENGINEERING
Chulalongkorn University
Academic Year 2020

การทำนายราคาหุ้นที่สามารถอธิบายได้โดยใช้ตัวชี้วัดทางเทคนิคกับข้อมูลภาษาไทยที่สั้น

นายกิตติศักดิ์ ปรัชญาชูวงศ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2563

Thesis Title	Explainable Stock Price Prediction Using Technical Indicators With Short Thai Textual Information
By	Mr. Kittisak Prachyachuwong
Field of Study	Computer Engineering
Thesis Advisor	Assistant Professor PEERAPON VATEEKUL, Ph.D.

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

..... Dean of the FACULTY OF ENGINEERING
(Professor SUPOT TEACHAVORASINSKUN, D.Eng.)

THESIS COMMITTEE

..... Chairman
(Professor BOONSERM KIJSIRIKUL, D.Eng.)

..... Thesis Advisor
(Assistant Professor PEERAPON VATEEKUL, Ph.D.)

..... Examiner
(DUANGDAO WICHADAKUL, Ph.D.)

..... External Examiner
(Thanapat Kangkachit, Ph.D.)

กิตติศักดิ์ ปรัชญาชูวงศ์ : การทำนายราคาหุ้นที่สามารถอธิบายได้โดยใช้ตัวชี้วัดทางเทคนิคกับข้อมูลภาษาไทยที่สั้น. (Explainable Stock Price Prediction Using Technical Indicators With Short Thai Textual Information) อ.ที่ปรึกษาหลัก : ผศ. ดร.พีรพล เวทีกูลPh.D.

การทำนายแนวโน้มของตลาดหุ้นได้รับความสนใจมาตั้งแต่อดีตจนถึงปัจจุบัน โดยเฉพาะอย่างยิ่งที่ในปัจจุบันสามารถเข้าถึงข้อมูลที่มีจำนวนมหาศาลได้อย่างง่ายดาย งานวิจัยก่อนหน้านี้ได้มีความพยายามคาดการณ์แนวโน้มของตลาดหุ้นโดยพิจารณาข้อมูลเชิงตัวอักษรเพียงอย่างเดียว อย่างไรก็ตาม มันสามารถพัฒนาให้มีประสิทธิ์ภาพดีขึ้นได้เมื่อพวกเขานำเทคนิคการฝังคำเข้ามาปรับใช้ ในบทความนี้ เราขอเสนอรูปแบบการเรียนรู้เชิงลึกเพื่อทำนายตลาดซื้อขายล่วงหน้าของประเทศไทย (TFEX) ที่มีความสามารถในการวิเคราะห์ข้อมูลทั้งตัวเลขและข้อความ พวกเราใช้หัวข้อข่าวเศรษฐกิจภาษาไทยจากแหล่งข้อมูลออนไลน์ต่างๆ เพื่อช่วยให้หัวข้อข่าวสะท้อนความสัมพันธ์ที่แท้จริงของตลาดได้ดียิ่งขึ้น พวกเราได้แบ่งหัวข้อข่าวออกเป็นดัชนีเฉพาะอุตสาหกรรม (เรียกอีกอย่างว่า“sector”) เพื่อสะท้อนการเคลื่อนไหวของหลักทรัพย์ที่มีพื้นฐานเดียวกัน วิธีการที่นำเสนอประกอบไปด้วย Long Short-Term Memory Network (LSTM) และสถาปัตยกรรม Bidirectional Encoder Representations (BERT) จาก Transformers เพื่อทำนายกิจกรรมตลาดหุ้นรายวันพวกเราได้ทำการประเมินประสิทธิภาพของแบบจำลองโดยพิจารณาจากความแม่นยำในการคาดการณ์และผลตอบแทนที่ได้รับจากการจำลองการซื้อและขาย ผลการทดลองแสดงให้เห็นว่าการปรับปรุงทั้งข้อมูลตัวเลขและข้อความของแต่ละภาคส่วนสามารถปรับปรุงประสิทธิภาพการทำนายและทำงานได้ดีกว่าโมเดลก่อนหน้าที่พวกเรานำมาใช้เปรียบเทียบในงานวิจัยนี้ทั้งหมด

สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....

ปีการศึกษา 2563 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6270018221 : MAJOR COMPUTER ENGINEERING

KEYWORD: Deep learning, Time Series Forecast, Natural language processing

Kittisak Prachyachuwong : Explainable Stock Price Prediction Using Technical Indicators With Short Thai Textual Information. Advisor: Asst. Prof. PEERAPON VATEEKUL, Ph.D.

A stock trend prediction has been in the spotlight from the past to the present. Fortunately, there is an enormous amount of information available nowadays. There were prior attempts that have tried to forecast the trend using textual information; however, it can be further improved since they relied on fixed word embedding, and it depends on the sentiment of the whole market. In this paper, we propose a deep learning model to predict the Thailand Futures Exchange (TFEX) with the ability to analyze both numerical and textual information. We have used Thai economic news headlines from various online sources. To obtain better news sentiment, we have divided the headlines into industry-specific indexes (also called "sectors") to reflect the movement of securities of the same fundamental. The proposed method consists of Long Short-Term Memory Network (LSTM) and Bidirectional Encoder Representations from Transformers (BERT) architectures to predict daily stock market activity. We have evaluated model performance by considering predictive accuracy and the returns obtained from the simulation of buying and selling. The experimental results demonstrate that enhancing both numerical and textual information of each sector can improve prediction performance and outperform all baselines.

Field of Study: Computer Engineering Student's Signature .....

Academic Year: 2020 Advisor's Signature .....

ACKNOWLEDGEMENTS

วิทยานิพนธ์และปริญญาโทนี้เป็นประสบการณ์ที่มีค่าที่สุดในชีวิตของผม ผมตัดสินใจเรียนคอมพิวเตอร์ในระดับปริญญาโท และผมก็ไม่ผิดหวัง ผมชอบที่จะได้ทำงานและเรียนรู้จากคนที่น่าทึ่งมากมายในด้านนี้ ผมรู้สึกขอบคุณสำหรับทุกการสนับสนุนที่ได้รับจากครอบครัว ที่ปรึกษา เพื่อน และผู้คนมากมายในช่วงเวลานี้เพื่อทำวิทยานิพนธ์ให้สำเร็จ ผมรู้สึกโชคดีที่ได้เป็นส่วนหนึ่งของกลุ่มเรียนรู้การทดลองของที่ปรึกษาของผม ผศ.ดร. พีรพล เวทีกูล ซึ่งเป็นอาจารย์ที่ทุ่มเทที่สุดเท่าที่ผมเคยเจอมา การฝึกสอน การแบ่งปันความรู้ และการแก้ปัญหาที่เขาสนับสนุนผมตลอดหลักสูตรนั้นน่าทึ่งมาก เพื่อนและเพื่อนร่วมงานในห้องปฏิบัติการ Datamind มีความเป็นมิตรและช่วยเหลือดี ผมไม่สามารถบรรลุสิ่งนี้ได้หากไม่ได้รับการสนับสนุนจากพวกเขาเช่นกัน ขอขอบคุณคณะกรรมการทุกท่าน ประกอบด้วย ศ.ดร.บุญเสริม กิจศิริกุล ดร.ดวงดาว วิชาดากุล ผศ.ดร.ธนากร ลิขิตาภิวัฒน์ และ ดร.ธนภัทร ฆังคะจิตร ที่ให้คำแนะนำอันมีค่าสำหรับวิทยานิพนธ์ฉบับนี้ ผมขอขอบคุณทุกฝ่ายที่สนับสนุนและมีส่วนร่วมในการวิจัยครั้งนี้ ขอบคุณมากครับ

Kittisak Prachyachuwong

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

1.1 Motivation

Since the efficient market hypothesis (EMH) has been proposed by Fama [1], it has become the mainstream among financial economists. It states that if the capital market is efficient, information will be quickly and equally disseminated in the market, and rational investors will be able to interpret the information correctly. Thus, when new information arises, it will be captured and reflected in the stock price immediately. The EMH is associated with the random walk theory, arguing that the future stock price is random and unpredictable. Neither the technical analysis nor fundamental analysis can be used to predict the future price and generate excess return. More specifically, the EMH states that if using the past price information (also known as the technical analysis), investors are unable to outperform the market. It is said to have weak-form efficiency. If using all available public information (also known as the fundamental analysis), including the historical price and volume, investors are unable to outperform the market. It is said to have semi-strong form efficiency. After all, if using all public and private information available, investors are still unable to outperform the market, then it is said to have strong form efficiency.

A number of studies attempt to test the hypothesis, providing empirical results on stock price predictability. Numerous empirical studies have shown that future returns can be obtained by following specific trading strategies using past price information or other public information. Some examples of early studies are as follows. First, the small-size stocks will generate a higher return than the large-size stocks [2]. Second, the low price-to-earnings-ratio stocks generate higher returns than the high price-to-earnings-ratio stocks [3]. Third, the past winning stocks tend to overreact and continue to generate positive returns over several months [4], and this trend will reverse over the long run. Fourth, low liquidity tends to be associated with higher

returns [5]. This empirical evidence shows that the excess returns are unable to justify by risk taken by investors. In other words, investors are able to generate excess returns by following this past price and public information. Although the above findings are often referred to as market anomalies, the implication is a persuasive strategy for investors to follow.

Since the year 2000, the rise of machines has become a fundamental shift in the financial markets. The computer now dominates the trading activity that used to be done by a human. More specifically, the traders use algorithms to acquire information and make trading decisions at the speed of light. There are various algorithms used by traders. Some use a market maker strategy to offer liquidity to other traders by posting the bid and offer in the market. Some predict price change by learning the price change in the market. An attempt to generating excess return has changed from forming a portfolio using a certain strategy to predicting the future price path. Nowadays, machine learning (ML) techniques play an essential role as a core algorithm to predict the future stock price.

Recently, deep learning has been a subfield in ML and is considered one of the emerging areas that are showing promising results. Many prior researches focused on modeling either numerical data or solely using textual data. [6, 7] applied neural network technical indicators, assuming that numerical properties can reflect all factors of the stock market. However, these methods have limitations as stock market behavior continues to respond to other external factors that can be captured in the news. [8-11] on applying text analysis, mimicking fundamental analysis such as news articles and financial reports to find relevant relationships and to predict stock market behavior. In real-world scenarios, most investors usually consider both numerical and textual data. [12-15] used news headlines to relate historical prices and technical indicators, and most of them agreed that using only the news headlines should be sufficient for stock prediction representing the whole article. [14] used an event embedding as a representative of events to help increase prediction efficiency. An

event is extracted using a tool called “Open Information Extraction (OpenIE)”, which converts the news into three entities: actor, action, and object. Unfortunately, OpenIE supports only documents in English. For the Thai stock market, there are only a few deep learning pieces of research. Moreover, it is surprising that there is only one research based on deep learning [16] utilizing both numerical and Thai textual information. It is quite challenging to process a Thai corpus since, such as the tokenization problem, arising from the fact that the Thai characters do not have the spaces used to divide words as in the English alphabet. Another significant challenge is the Thai news headlines used in this research with short sentences. When using the previous research techniques, they cannot convey the essence of the news. Therefore, it is crucial to use a pre-trained model in order to overcome the scarcity of information in the text data.

In this research, we aim to propose a novel deep learning model to forecast a stock market trend (called “stock indexes”). It utilizes both numerical data (historical data and technical indicators) and textual information (news headlines). The model architecture is a combination between LSTM [17] and a pre-trained BERT [18, 19] responsible for numerical and textual data, respectively. We focus on forecasting a stock index (the whole market) rather than an individual stock since it has a lower risk. Since a stock index is a combination of top individual stocks from various sectors, we propose to use news headlines to capture the market's movement at (“the sector level”) rather than the overall market. Thus, we build a separated textual model for each sector and then combine results of all sectors to reflect the whole market's movement. In the experiment, we focus on a stock market in Thailand using SET50, a stock index combining the top fifty individual stocks. Since the textual data are news headlines in Thai, all challenges of natural language processing (NLP) in Thai must be addressed. Furthermore, Multilingual BERT is chosen since it supports 104 languages, including Thai. The experiment was conducted on the stock index data from 2014 to 2020 and economic news headlines collected from various online sources. The results

showed that our sector-based framework could forecast the stock market trend more accurately than other baseline models. Additionally, a simple investment was simulated using the output from our forecasting model, and the results showed that a straightforward trading strategy could yield better annualized returns than the stock market and other models. This comparison can be used as a test for EMH.

1.2 Objective

We aim to propose a deep learning model to forecast stock price prediction based on both numerical and textual data, which is collected from news headlines (short text) in Thailand. Also, the model explainability is our focus of this thesis.

1.3 Scope of works

1. Textual data used in this research was taken from an online source. Focusing on economic news and related headlines.
2. The numerical data used in this research are historical prices (open / close / high / low), technical indicators. Based on research [14] which was generated by historical stock price data using data from Stock Exchange of Thailand (SET) 2016 – 2019 focusing on SET index.
3. Measuring model performance by considering predictive accuracy and the returning obtain from the simulation of buying and selling.
4. Verifying model performance by comparison the results of each data type only versus the combination of the two data types.

1.4 Expected results

1. Improve the deep learning model performance by using both numerical and textual type input.
2. Be able to use appropriate methods to deal with the challenges of applying Thai language.
3. Able to correlate news headlines and changing behavior of stock trends.

1.5 Research Plan

1. Studying the related works and literature review
2. Studying basic data sets and models used to predict stock market trends.
3. Experimental design
4. Summarize preliminary result
5. Thesis proposal topic examination
6. Further experiments as in the proposal
7. Evaluate experimental result and tuning model as needed
8. Academic paper publication
9. Conclude results and write up the thesis
10. Thesis examination

1.6 Publications

“Stock Trend Prediction Using Deep Learning Approach on Technical Indicator and Industrial Specific Information” Prachyachuwong K., Vateekul P. (2021). In: Information (Switzerland), Publisher: MDPI Multidisciplinary Digital Publishing Institute.

Table 1 Research Plan Gantt Chart

Research Plan	2020												2021							
	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8
1. Study the related works and literature review																				
2. Experimental design																				
3. Develop, implement and experiment models																				
4. Summarize preliminary result																				
5. Thesis proposal topic examination																				
6. Further experiments as in the proposal																				
7. Evaluate experimental result and tuning model as needed																				
8. Academic paper publication																				
9. Conclude results and write up the thesis																				
10. Thesis examination																				

CHAPTER 2

BACKGROUND KNOWLEDGE

2.1 Artificial Neural Network (ANN)

A neural network is a computer system inspired by the biological neural network depend on the system of organizational human brain. Brain cells are referred to as "neurons." Artificial neural networks can be divided into three main groups: input layer, output layer and hidden layer, where the hidden layer is in the middle between input layer and output layer. Firstly, each input is calculated to optimize weight then the weighted is passed to the activation that limits the amplitude of the output so, the hidden layer allows the model to capture the nonlinearity and complexity of the problems. Lastly, the output layer provides prediction, which can be a classification or regression value. At the same time, deep neural network (DNN) is an artificial neural network (ANN) is many hidden layers to be able to solve even more complex problems. Figure 1 illustrates the three main components of Artificial neural network (ANN)

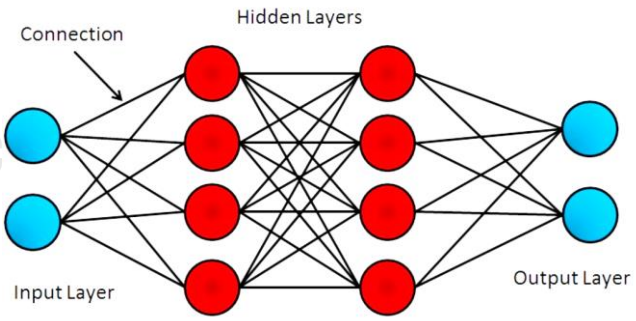

Figure 1 The main components of artificial neural networks [20].

2.2 Recurrent Neural Network (RNN)

Recurrent neural network (RNN) is designed to take sequential data with no determined limit on size. A neuron from the sequence is related to others and has an influence on its neighbors. Hence, this is what captures this relationship through input meaningfully. They remember what they had learned from the previous input while producing the output, making its decisions had influenced by what they had learned

from the past, causing a self-loop. Figure 2 illustrating a Recurrent Neural Network, with a hidden state that is meant to carry pertinent information from one input item in the sequences to others. RNN can be useful for financial data which are usually presented in forms of time series such as stock price movement. On the other hand, they are successful in solving many tasks of Natural Language Processing (NLP) including economic news headlines. There are various types of RNN. For example, Long-Short Term Memory (LSTM) that is another one often used because of better preserves long-term dependencies. More detail on LSTM will be covered in the next section.

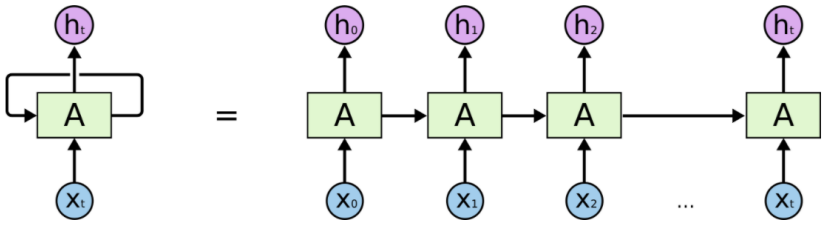

Figure 2 Shows the work of Recurrent Neural Network [20].

2.2.1 Long-Short Term Memory (LSTM)

Long-Short Term Memory (LSTM) is an architecture that extends the memory of recurrent neural networks and design to avoid the long period dependency problem such as the vanishing gradient problem which is where a neural network stop learning because the weights within the neural network become smaller and smaller. They introduce long-term memory into RNN by using four interacting gates. Each gate is like a switch that controls the read/write that helps them on how to forget or memorize the input data. The forgetting gate is used to decide whether a value should be forgotten then the next two gates are used to control which values should be updated or decided whether information should be discarded, respectively. The last gate is used to filtering out lavish values. Figure 3 illustrates the architecture of the Long-short term memory neural network. LSTM refers to the fact that it is a model for short-term memory which can last for a long period of time. LSTM is also ideal for predicting time

series where the period of time between important events cannot be determined exactly.

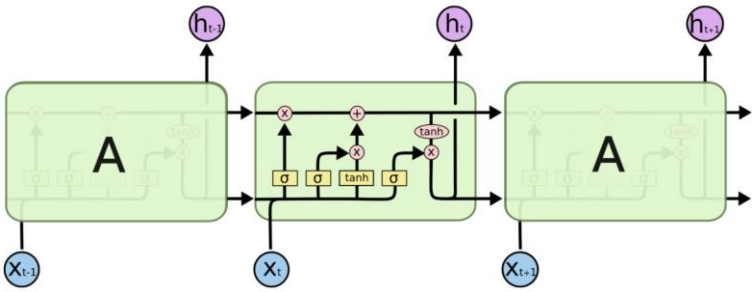

Figure 3 Demonstrates the functionality of the LSTM differential gate [21].

2.3 Textual Representation

Text representation is one of the most important processes for analyzing textual data. This is because computers cannot process text information directly. Therefore, they must be converted into numbers so that the computer can understand them. There are many ways to represent texts, as follows.

2.3.1 Word Embedding

Word embedding was one of the most popular document word representations of old times. It can capture the context of a word in a document where words that have the same meaning have a similar representation. Word embedding methods learn a vector representation for a vocabulary from a corpus of text. The process is usually joint with the neural network architecture on some tasks, such as text classification. There are various techniques to learn from text data. For example, Word2Vec [22] is high- quality word embedded. It can learn more efficiently in less space and save more time on learning that makes it possible to learn large embedded from the text that is much larger.

CHAPTER 3

LITERATURE REVIEW

3.1 Literature review

In this section, we are going to discuss various stock forecasting techniques. Since they can be related to EMH, this section will start from EMH's details that categorize markets into three levels. Then, recent techniques related to a stock market prediction will be discussed in two groups; first, a stock market prediction using only textual information mainly inspired by [18, 23]; second, a stock market prediction using both numerical and textual information largely inspired by [15, 16].

3.1.1 Efficient Market Hypothesis (EMH)

EMH is a hypothesis that stock market prices are already reflected in all relevant information. In other words, stock prices are reflected by investors' beliefs about future expectations. We can categorize the market's efficiencies into three levels [24]: (1) “a weak form”, assuming that if investors can beat the market using historical stock prices, also known as technical indicators, the stock market is not yet considered to be efficient; (2) “a semi-strong form”, assuming that if investors can beat the market using the first level data (historical data) with public data such as news, earnings, and other stock fundamentals, this is not yet considered to be efficient; (3) “a strong form”, which is the most efficient one. If investors have won against the market using the data from the first and second levels with private information, it can be concluded that the stock market is not yet considered to be highly efficient. In addition, to apply EMH concepts to our research, it can be concluded that the first two levels of EMH are widely used by considering both historical data and public data to forecast the market trend. Nevertheless, the strong-form level considers private information, which is illegal in trading; thus, it is not focused on in our research.

3.1.2 Stock Market Prediction Using Only Textual Information

Analyzing the stock market using relevant text is complex but exciting [12, 25-33]. For example, a model named AEnalyst was introduced by Lavrenko et al. [34]. Their goal was to predict intraday stock price trends by analyzing news articles published on the YAHOO finance homepage.

Mittermayer and Knolmayer [35] utilized multiple models to predict short-term market reactions to the news using text mining techniques. Their model predicts the one-day trend of the five major company indices. Wu et al. [25] predicted stock trends by choosing a representative set of ambiguous attributes (keywords) that affect each stock.

However, these simple methods are not suitable for conveying the headline's meaning and have several limitations, including the disclosure of rules that may govern market dynamics, making forecasting models unable to capture the impact of recent stock market trends.

One of the most definitive studies is the research of Ding et al. [12], which assumes that news can influence the stock market's behavior and the news of the previous day will influence the daily changes in the stock price. They try to create a representative event by using a process named open information extraction, which converts daily news into three representative data: actor, action, and object. The result shows that their research works well and outperforms early research by increasing the efficiency of prediction. They proved by trials that the news headlines should be sufficient for textual features as opposed to the whole article.

Shi et al. [36] implemented a hierarchical neural network for stock prediction on textual news input. Their structure embedded input into three layers; word representation, bigram phrases, and news headline layer, respectively. The results were brought to the feed-forward regression dense layer. Figure 4 illustrates a hierarchical neural network on textual news input data. The blue boxes represent the numerical

information. The light blue boxes represent the textual information, and the dark green boxes represent the classification tasks}.

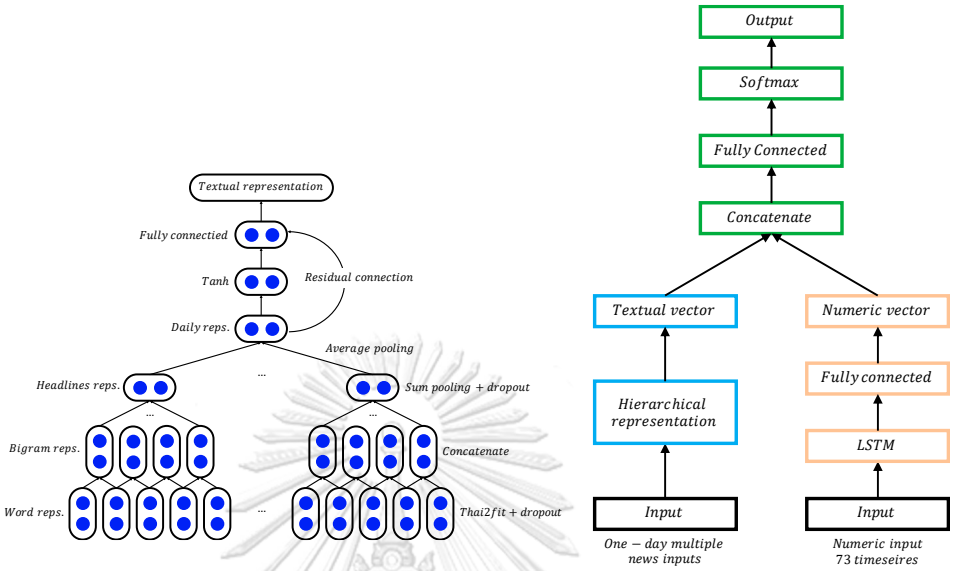

Figure 4 The hierarchical neural network (from Figure 2 in [36]). Detailed structure of a hierarchical representation (left). The overview architecture for the whole model (right).

Pisut et al. [15] have implemented the deep learning architecture for stock prediction by leveraging pre-trained word embedding to increase their model performance. Therefore, we customized the architecture from them and replaced the GloVe embedding [37] with FastText embedding [38], which is considered one of the best word embedding for Thai news headlines. Therefore, we took FastText embedding as one of our baseline models.

However, we noticed that the subword contextual text representation was significantly better than word global text representation fix embedding. The critical factor is that each type of pre-train can be observed from the BERT and FastText pre-train, ranging in size from large to small, respectively.

We will introduce more about model improvement and embedding generation. Ling et al. [39] have proposed a character-to-word model in which they combined character-level embedding with word-level embedding and then measured them

against five different language datasets, and their performance was satisfactory. Character-level embedding with a convolutional neural network (CNN) model is proposed by Kim et al. [40]. They proved that a highway network could accumulate performance from several language prediction tasks. Wehrmann et al. [41] applied character embedding with CNN to analyze the sentiment of Twitter.

However, using only character-level features is not satisfactory in terms of performance on Thai short-text classification because they rely on word segmentation, which sometimes leads to incorrect classification. Bidirectional Encoder Representations from Transformers (BERT) [18] provide the pre-trained vector representation of the words, which can be used further with the various AI models. BERT is trained using masked language modeling, which randomly masks some tokens in a text sequence, and then independently recovers the masked tokens by conditioning them on vectors. The BERT architecture is a framework that provides representation by a standard conditional probability both from the left and right contexts for all computed layers. Vector BERT was used in experiments to employ the transfer learning model to enhance the current prediction model's capabilities. Google also has a pre-trained model (a large-scale text corpus called Multilingual Cased (BERT-Base)), which undergoes initial training on the top 104 languages and is very suitable for the Thai language. Figure 5 illustrates the architecture of BERT for classification tasks.

Moreover, Bidirectional Encoder Representations from Transformers, one of the best language models available in the NLP research, outperforms many tasks. Hiew et al. [23] showed a significant enhancement of BERT in sentiment analysis compared to prior existing models. Next, Othan et al. [42] showed that utilizing deep learning models and a new-generation word embedding model called BERT could improve classification performance.

We utilized BERT because it provides a better text representation than the previous embedding. BERT is a deep, multi-layered neural network and generates two different vectors for the word “sentence” as they appear in two very different contexts (i.e., contextual vector). Moreover, BERT could solve our problems because they have BERT-Base Multilingual-Based available from Google, representing Thai words that convert a group of text characters into a numeric representation value. However, the disadvantage is that it took much compute-intensive inference, meaning that it would be costly if investors wanted to use it in scale-by-size production. However, the predictive performance was noticeably better.

Finally, to the best of our knowledge, this is the first study aimed at analyzing the direction of stocks by using news headlines to capture market movements at the sector level rather than the overall market, as the stock index is a combination of the top fifty stocks from various sectors. Capturing sector information can actually reflect the market trends.

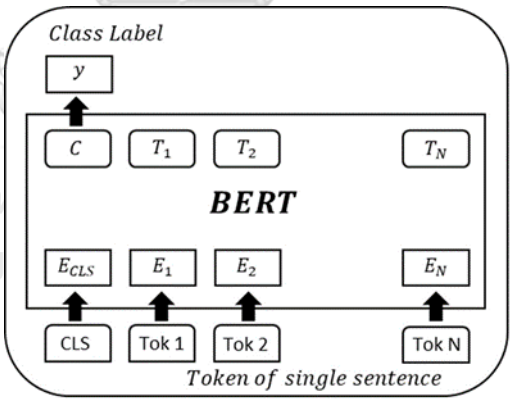

Figure 5 Illustrations of fine-tuning BERT for classification tasks (from Figure 4(b) in [18]) where each news (input) contains N tokens (words), and the output is a predicted class label y

3.1.3 Stock Market Prediction Using Numerical and Textual Information

Most prior research has focused on either text or numeric data as input but not both types of information. Nowadays, many researchers are interested in improving a model's performance by utilizing both types of data.

For the traditional machine learning techniques, Tantisantiwong et al. [43] recently proposed a framework to forecast SET50, a stock index in Thailand, using both numerical and textual information. For the textual information, they have gathered social media data with the advice of experts to manually define keywords for sentiments (positive and negative). Then, each document is labeled as positive or negative based on those predefined keywords. Next, sentiment keywords are extracted from those labeled documents. After that, each document is assigned a score called "the market composite sentiment index". Finally, a multiple linear regression is generated based on both numeric (historical data of SET50) and textual (the sentiment index) information along with other control variables. Although this work is interesting, it requires experts effort to define sentiment keywords. Furthermore, these keywords are not publicly available, and they can be changed at different periods of time, so they must be periodically updated. Therefore, this work is not included in our study due to the manual processes required of experts.

For the deep learning approach, Vargas et al. [14] are the first to present the concept of gathering two different types of information. They take the news headlines into word vectors by using the Word2Vec algorithm and then pass them to the convolutional neural network to extract the outstanding words. The results were brought to long short-term memory to find the relationships between all of the news. On the other hand, technical indicators were created from historical stock price information, which learns directly from long short-term memory. The results are obtained from the two types of data and are then combined and used to predict stock market behavior.

The work of Tanawat et al. [16] is another example that is very interesting. To the best of our knowledge, they are the first to extend the studying for the Thai stock market combining the Natural Language Processing (NLP) domain with technical indicators. There are various ideas and uses in Thai language headlines, more so than in other languages. For example, the tokenization problem. Thus, they have proposed

a modified hierarchical structure for textual representation, which is the highlight of the research by Shi et al. [36], but they customized the model by using the “Newmn” tokenization from pyThaiNLP and replaced the word2vec embedding with thai2fit embedding so that models can learn Thai news headlines. Figure 4 illustrates the architecture of the hierarchical textual representation. Their model was designed to optimize the explanation with vector representation from the word to bigram phrases, title, and daily news representation level, respectively. The results have shown that adding textual representation increased the profitability more so than previous research.

Another relevant research is presented by Pisut et al. [15]. They offer a deep learning model that can acquire textual and numerical data to use in predicting stock market trends. They took the event data generated from the mean of each day's event embedded vector as textual input and split the data into three parts: events from the past thirty days, events from the past seven days, and events from the past day, respectively. Event vectors dating back seven days and thirty days are fed into a convolutional neural network (CNN) to create a feature map that replaces essential events in the past and then combines them with event vectors from the previous day. On the other hand, historical price representation vectors and technical indicators are entered into long short-term memory (LSTM) and featured in the time series data analysis. Finally, the results obtained from each type of information are interconnected to predict the stock market trend.

In this research, we will consider both textual and numerical data, as many previous studies have shown that the efficacy obtained from both types of information is better than the performance focus on only textual or numerical data.

CHAPTER 4

PROPOSED FRAMEWORK

This section provides a deep learning method for predicting industrial stock trends by considering numerical and textual data together. The models focus on improving prediction accuracy and performance in the field of returns when using prediction results to simulate trading. We have divided this into two main topics: data preprocessing Section 4.1 and the proposed prediction model Section 4.2, respectively. For the model, there are three main components, as shown in Figure 6.

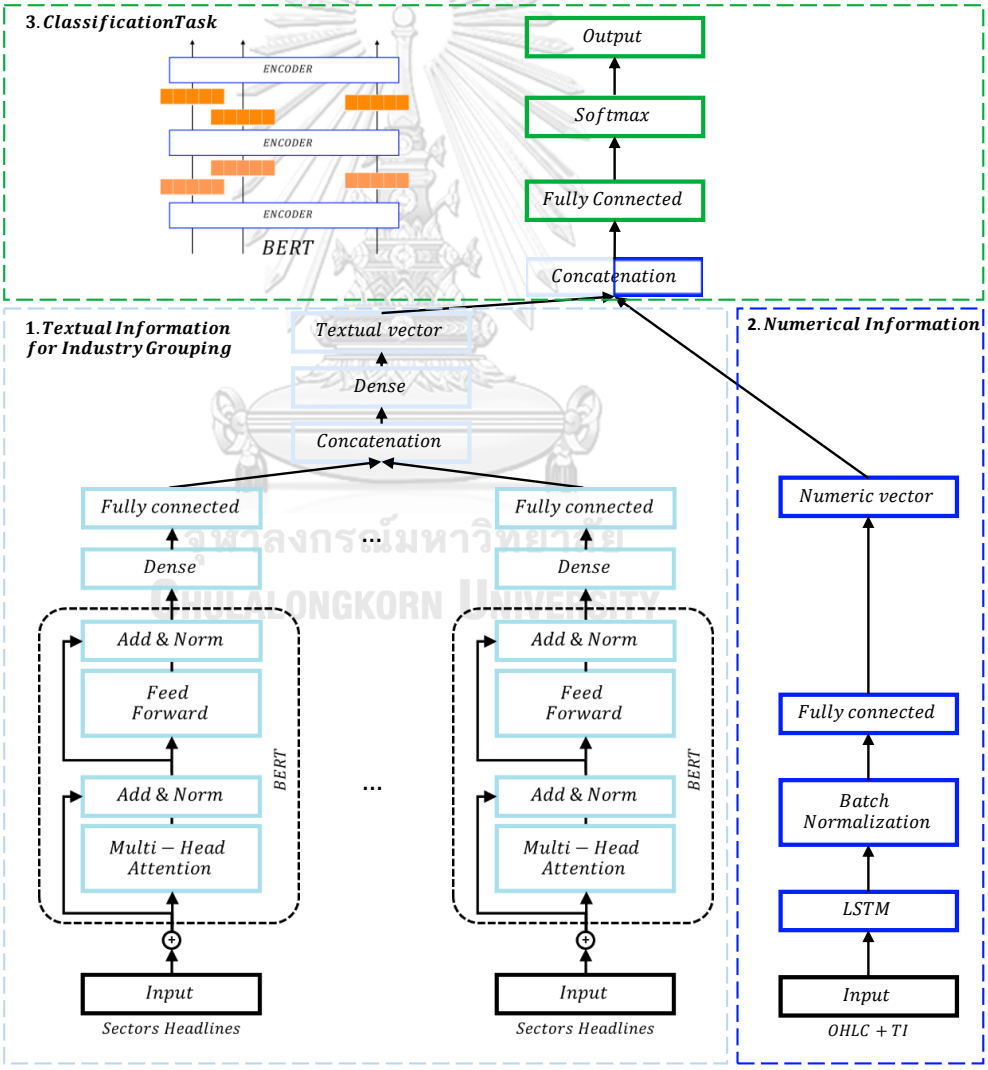

Figure 6 Our proposed model using numerical inputs and sector-based textual information

4.1 Data Preprocessing

This section describes the data preparation to be used as the input to a model for predicting industrial stocks' trends.

4.1.1 Textual Data Preprocessing

In this research, we have eliminated the Thai headlines between 2014 and 2020 from online sources and filtered only economic topics. Since the collected raw dataset is very dirty, different preprocessing techniques are used to clean up the dataset, e.g., lower casing, text without punctuation, stop word removal, removing duplicates, and trading symbol extraction methods.

As we focus on supervised deep learning strategies, it is necessary to label the collected Thai texts. One way to label documents is to read the news titles and labels them manually. This would provide results of a high level and ensure confidence. However, these qualities require a human to analyze the entire dataset. We have followed the authors of [44], who said they could automatically label documents using some market feedback, which is not perfect but quick and straightforward. We take a similar approach and groups documents that were released on the same day with the corresponding next-day market yield change. The value is calculated with Equation 1.

On the other hand, the labeling of textual data is used in a similar way. That is to say, the labeling obtained from the historical stock price is correlated with the textual data using the date as the correlation link. Like many previous kinds of research, we can divide the stock return ratio into three classes: upward, downward, and sideways, representing the significant dropping, rising, and steady stock trend on the next date, respectively. We set up a particular threshold to bin the return ratio percentage, based on the researching mentioned above, i.e., Downward (Rise_Percent (t) < −0.41%), Upward (Rise_Percent (t) > 0.87%), and sideways (−0.41% <= Rise_Percent (t) <= 0.87%).

$$Rise_{Percent} = \frac{Open_{Price(t+1)} - Open_{Price(t)}}{Open_{Price(t)}} \quad (1)$$

4.1.2 Numerical Data Preprocessing

We used open, high, low, and close as the historical prices in Table 1 and applied technical indicators, a feature that is based on historical stock prices, by mathematical equations presented in Reference [45]. As a result, there are a total of 15 technical indicators to make adjustments to the parameters as appropriate. Table 2 shows a list of technical indicators used in our experiment.

Table 2 Fundamental parameter summary.

Feature	Explanation
Open Price	The first share price at the start of daily trading
Close Price	The final share price at the end of daily trading
High Price	The highest share price during daily trading
Low Price	The lowest share price during daily trading

Table 3 List of 15 technical indicators.

RSI	CMO	WMA	PRO	William's %R
EMA	ROC	SMA	HMA	TripleEMA
DMI	PSI	CCI	CMFI	MACD

4.1.3 Data Normalization

The numerical inputs are in very different ranges; therefore, it is essential to standardize the dataset in a close range to enable the model's faster teaching. We use a z-score to convert data, which is the zero mean and standard deviation of the data. Where μ is the mean of the input x, and σ is the standard deviation of the input x.

$$z = \frac{(x - \mu)}{\sigma} \quad (2)$$

4.2 Proposed Model

4.2.1 Model Architecture

The model we present aims to predict stock market trends, focusing on finding the context of daily headlines and correlating different types of data using numerical and textual data. Initially, we start with a data stream from textual and numerical information and finally feed the data to a forecasting model with correlation inference. An overview of the framework is shown in Figure 6.

Firstly, we used a concept presented in [13], where it is reported that stocks belonging to the same industry tend to behave similarly. Grouping these stocks can improve the performance of the model. Therefore, we have divided the news headlines for each sector of the SET50 industry grouping, for which we create a custom stock index to suit our personal purposes. We need to emphasize the specific meaning of each stock within each industry segment, in accordance with Table 5, as we discussed in Section 4.1.2, by using the best existing aspect of an index to make it more usable for this research.

We use a method known as capitalization-weighted indexing because large stocks with market capitalization influence the dynamics of high indexes, such as the S&P500 (US), FTSE100 (UK), and SET50 (Thailand). Top-weighted best reflects the actual market, as the largest and most stable companies have the most influence on the index, while small or growing companies dictate the index's movement and have a smaller amount of value. The very definition of a weighted index has the highest weight and serves as a good gauge for the market's overall health. Equation 3 illustrates the capitalization-weighted indexing calculation by applying the SET50 calculation of the Stock Exchange of Thailand.

$$Index = \frac{\sum_{t=1}^{n}(Prices_{it} \times ListedShare_{it}) \times AdjustmentFactor_{it}}{AdjustedBMV} \times BaseValue \qquad (3)$$

Where $Prices_{it}$ is the price of each security constituting the index as of the calculation date. $ListedShare_{it}$ is the number of registered shares of each security

constituting the index as of the calculation date. $AdjustmentFactor_{it}$ is the weight limit rate of each asset in the index as of the calculation date, and $AdjustedBMV$ is the market capitalization of all securities that make up the index. This is weighted by the weight limit rate of each security in the index as of the base day.

1. We input each headline into the BERT model that we have employed (Multilingual Cased pre-training weights (BERT-Base)) to shape the language with our headline corpus. We have selected the first token from the BERT-Base output, which is often used for classification tasks to display the headlines. Next, we trained the model prediction until the loss function reaches convergence through the BERT architecture workflow. Lastly, we keep the best final embedding values and join each sector to represent the day's best headlines.
2. For numerical information, we feed the input into the LSTM network, which is widely used for time series data entry. The output from the LSTM is fed into a hidden layer that can indicate forecasts based on technical indicators and historical price data.
3. Finally, we concatenate all output from the textual representation vector, and the numeric vector then feeds them into the final hidden layers to predict the following day's stock trend prediction. The output of the model is a multiclass classification, where class "UP" represents an upward trend of the stock market, "DOWN" represents a downward trend of the stock market, and "STABLE" indicates that the trend of the stock market is in no definite direction or moving in a narrow range (sideways).

4.2.2 Training Process

To create a model, we need to prepare training, validation, and testing datasets for model construction, model tuning, and model assessment, respectively. In addition, we have tried to avoid overfitting results by following a setup of [46], which divides time-series data into multiple parts over time using a sliding window, as shown

in Figure 7. The performance of each model can be evaluated by averaging the results of all testing datasets rather than relying on just one testing dataset.

Our study has three datasets (#1, #2, and #3); this is called "3-fold cross-validation". Using a sliding window of 1 year, the periods of training, validation, and testing datasets are three years, around one year, and one year, respectively. Although Sutheebanjard and Premchaiswadi [47] recommend that it is sufficient to use only one year of data to train a model, this suggestion cannot be applied to our study. Their model is a multiple linear regression, while ours is a deep learning model, which is much more complex and commonly requires more training data.

Figure 7 Training, validating, and testing approach

CHAPTER 5

EXPERIMENTAL SETUP

5.1 Datasets

The experiment was conducted on the stock market in Thailand via a stock index called "SET50" along with textual data of news headlines in Thai mainly due to data availability. Nevertheless, the stock market in Thailand is one of the emerging markets of the world. The size of the market capitalization in Thailand is 2nd in Asia, as of 2019[1] and 20th in the world as of 2018[2]. Therefore, it should share common characteristics with other emerging markets. Moreover, it is common to utilize both numerical and textual information for stock market forecasting as stated in the semi-strong form level in EMH (in Section 3.1) Therefore, the proposed model can be applied to other major indexes, and the results in this study can somewhat represent other emerging markets.

5.1.1 Numerical Statistic

The EOD numbers (an end-of-day order is a buy or sell order for securities requested by an investor that is only open until the end of the day) used in our research come from the Stock Exchange of Thailand (SET50), corresponding to the period from 2 January 2014 to 14 February 2020, Table 3. The numerical data include price information, such as open, high, low, close, volume, as well as calculating the base value, such as the book value and the P/E ratio.

[1] http://www.set.or.th/en/news/econ_mkt_dev/overview_2019.html, accessed on 5 June 2021

[2] http://www.indexmundi.com/facts/indicators/CM.MKT.LCAP.CD/rankings, accessed on 5 June 2021

Table 4 Numerical data summary showing the total number of records (days).

No.	Data Period	Numerical Information		
		Training	Validating	Testing
#1	Jan.-2014 to Apr.-2018	1,117	109	351
#2	May-2014 to Mar.-2019	1,093	351	314
#3	Jun.-2015 to Feb.-2020	1,115	261	344

5.1.2 Textual Statistic

We only collected economic headlines, the source of which clearly groups economic news on a single topic, and only news headlines from various online sources corresponding to the EOD timeline. Due to our research scope, the entire Thai stock market index is studied, enabling all relevant economic news headlines. Therefore, we have to clean up approximately two-hundred-thousand headlines using the method presented in Section 3 to ensure tidiness and suitableness for further use. Finally, there are approximately one-hundred-thousand headlines being introduced every trading day, and we use these in this research. Detailed data statistics (Table 4) are included for each experiment, as not all tests use the same data boundary.

Table 5 Textual data summary showing the number of total news headlines; there are many news headlines per day.

No.	Data Period	Textual Information		
		Training	Validating	Testing
#1	Jan.-2014 to Apr.-2018	101,916	14,727	13,662
#2	May-2014 to Mar.-2019	101,875	14,802	14,398
#3	Jun.-2015 to Feb.-2020	101,860	13,722	14,228

In addition, we have classified the stocks in SET50. These are appropriate according to the seven industrial groups, as shown in Table 5. Next, we split the headlines for each sector for training, validation, and testing with three sets of sliding windows, as summarized in Table6.

Table 6 Industry grouping for the SET50 index.

Industry Symbols	Industry Group	Stock symbols
FINCIAL	Banking, finance, and securities	BBL, KBANK, KTB, SCB, MTC, TMB, TISCO, KTC, SAWAD, MTLS, TCAP, BLA
SERVICE	Transportation & Logistics, Health Care Services, Media & Publishing, Tourism & Leisure, Commerce	AOT, BEM, BTS, CPALL, HMPRO, CRC, BDMS, BH, BJC, GLOBAL, VGI
RESOURC	Energy & Utilities	BPP, EA, EGCO, GULF, GPSC, IRPC, PTT, PTTEP, RATCH, TOP, BGRIM, TTW
AGRO	Food and Beverage	CBG, CPF, MINT, TU, OSP
PROPCON	Construction services, Property Development, Construction materials	AWC, CPN, LH, SCC, TOA, WHA
TECH	Information & Communication Technology, Electronic components	ADVANC, DTAC, INTUCH, TRUE
INDUS	Petrochemicals & Chemicals, Packaging	IVL, PTTGC, SCGP

Table 7 Data statistic for industry grouping experiments.

No.	*Data period*	*Textual Information*			
		Sectors	*Training*	*Validating*	*Testing*
#1	Jul-2014 to Dec-2019	FINCIAL	27,662	4,067	4,030
		SERVICE	24,689	3,553	2,977
		RESOURC	16,176	2,215	2,236
		AGRO	11,480	1,723	1,598
		PROPCON	14,858	2,420	2,165
		TECH	4,808	569	440
		INDUS	2,070	349	458
#2	Aug-2014 to Jan-2020	FINCIAL	27,666	4,030	4,298
		SERVICE	24,595	3,625	3,132
		RESOURC	16,226	2,195	1,378
		AGRO	11,477	1,756	1,655
		PROPCON	14,884	2,408	2,341
		TECH	4,784	564	399
		INDUS	2,091	348	470
#3	Sep-2014 to Feb-2020	FINCIAL	27,760	3,731	4,245
		SERVICE	24,490	3,413	3,097
		RESOURC	16,195	2,034	2,299
		AGRO	11,453	1,670	1609

		PROPCON	14,919	2,222	2,298
		TECH	4,768	515	469
		INDUS	2,089	348	466

5.2 Baseline Model

This section discusses details of baseline models [15, 16] for performance comparison.

Numerical Input Only (LSTM (NUM)): This model uses only numerical information, as shown in the numerical module (the right box) in Figure 6. The inputs are time series of OHLC (4-time series) and technical indicators.

Textual Input Only (FastText): This model uses only textual information, which is different from the previous baseline.

Both Input (FastText + NUM): This model uses both numeric and textual information, as shown in Figure 6}.

For our model, BERT is chosen as a textual module with three variations. First, "BERT" refers to our model with only textual information. Second, "BERT + NUM" refers to our model with both numerical and textual information. Finally, "BERT_SEC + NUM" refers to our model with both sources of information in addition to a sector strategy for the textual data.

5.3 Evaluation Metrics

5.3.1 Performance Evaluation

As the current experiment being conducted is a supervised problem, a matrix is evaluated to compare results from different deep learning models that were implemented based on precision, accuracy, recall, and F1 [48, 49]. The metrics used to enhance performance are shown in Table 7, where "TP" represents the true positive

from the model, “TN” represents the true negative from the model, “FN” represents the false negative from the model, and “FP” represents the false positive from the model.

Table 8 Metrics for classification evaluations.

Metrics	Formula	Explanation
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Specifying the percentage of correct forecasts in all samples.
Recall	$\frac{TP}{TP + FN}$	Specifying the proportions of positive samples is classified as a positive sample.
Precision	$\frac{TP}{TP + FP}$	Identifying the proportion of real positive samples in the class that was classified as positive.
F1	$2 \times \frac{Precision \times Recall}{Precision + Recall}$	F1 is the precision and recall weighted harmonic mean.

5.3.2 Trading Profit

This work uses the stock buying simulation concept from Reference [46] to compare the predicted returns of the models we have presented. We only buy one contract at a time and set a stop loss at five percent of the total cost to prevent any forced margin and prevent the loss of all investments if that model predicts the wrong way. The conditions of the trading simulation are described below. (i) Make a buy when the model predicts an uptrend for the next day. (ii) Discard the stock when the model predicts a downtrend for the next day. (iii) If the conditions mentioned in the first two items are not met, hold shares. The formula shown in Equation 4 is only a formula that is suitable for simulating a bull market. However, the advantage of TFEX is that investors can trade both rising and falling stocks. Therefore, by simply swapping the execution of the equation, the stock will be traded in a downtrend accordingly.

$$Execution(t)\begin{cases} Buy\ n\ contracts;\ Prediction_{t-1} = UP\ and\ contracts\ = 0 \\ Sell\ n\ contracts;\ Prediction_{t-1} = DOWN\ and\ contracts\ > 0 \\ Hold;\ Otherwise \end{cases} \quad (4)$$

The profit or loss is calculated when the sell condition is reached by using the difference of price multiplied by the number of shares available. Additionally, if there are stocks on the last day of simulated trading, they are all sold at the closing price of that day. The equation for calculating profit or loss is shown in Equation 5.

$$Margin(t)\begin{cases} Execution(t) = Sell; Index\ mul. \times n \times (Exit\ Price - Entry\ Price) \\ t = T;\ Index\ mul. \times n \times (Close\ Price - Entry\ Price) \\ 0;\ Otherwise \end{cases} \quad (5)$$

The amount available at time t can be calculated based on the amount from the previous time, profit, or loss, as shown in Equation 6.

$$tMoney(t) = tMoney(t-1) + Margin(t) \quad (6)$$

The return is calculated every time the shares are sold. Furthermore, it is adjusted to the average annual return for easy comparison using Equation 7.

$$Annualized\ Return = \left\{(1 + Final\ tMoney)^{\frac{1}{\#Sim\ Day}}\right\} - 1 \quad (7)$$

where n = 1, index multiplier = 200, T is the end of time period, $tMoney$ denotes total money, and $\#Sim\ Day$ is the total of the simulation period.

5.4 Training and Hyperparameters

The computer resources used for this research have the following characteristics: Due to the large architectures we have designed, we can only scale up to 32 batch sizes. However, we use one of the most popular and most widely used optimization algorithms, Adaptive Moment Estimation (Adam), [50] with an initial learning rate of 0.001. We also apply a state-of-the-art approach, batch normalization [51], to our deep neural network to accelerate deep network training by reducing internal variable changes. Lastly, we train each model for a total of fifteen epochs and always choose the best result of our validation datasets; however, it takes us at least thirty-six hours to train each model.

CHAPTER 6

EXPERIMENTAL RESULTS

We conducted the study using the numerical and textual information described in the previous section. The performance is evaluated based on a "3-fold cross validation", which is the average result of three testing datasets (#1, #2, and #3). All models compared in the experiment are described and abbreviated in Section 5.2.

Tables 8 and 9 show a model comparison in terms of accuracy and F1, consecutively. Figure 8 illustrates a confusion matrix to provide details of our sector-based model (BERT_SEC + NUM), where rows and columns refer to actual and predicted classes, respectively. Note that a definition of each class (DOWN, STABLE, UP) is explained in Section 4.1.1 From the results, our sector-based model (BERT_SEC + NUM) is the winner with an average accuracy of 61.28% and F1 of 59.58%. In more detail, it achieves the highest accuracy in datasets #1 and #3 with 63.67% and 58.67%, respectively. It also shows approximately the same trends in terms of F1 achieving 62.30% and 56.34% in datasets #1 and #3, consecutively. Although the accuracy of the winner (61.28%) may not be promising, it is still higher than that of prior work (51.27%) [16]. This can be expected since a stock price prediction is considered a complicated problem due to its high volatility. Moreover, if we considered a confusion matrix in Figure 8, there are two serious error cases: lower-left (actual "UP" but predict "DOWN") and upper-right (actual "DOWN" but predict "UP"). Our model has failed these cases in minimal amounts of only 4.33%, 4%, and 4.17%, in datasets #1, #2, and #3, respectively.

Furthermore, we will discuss each module's effect on our model: transfer learning, numerical and textual information, and sector-wise strategy. Furthermore, a training strategy using our prediction results will be simulated to show annualized returns of our algorithm.

Table 9 Model comparison in terms of "accuracy" on testing data based on a 3-fold cross validation (#1, #2, #3 refers to the result of each fold), and the boldface represents the winner.

Group	Model	Accuracy (%)			
		#1	#2	#3	Avg.
Baseline	LSTM (NUM)	54.28	45.71	42.86	47.62
	FastText	54.50	58.33	52.83	55.22
	FastText + NUM	60.17	58.67	54.17	57.67
Ours	BERT	62.67	61.17	58.33	60.72
	BERT + NUM	62.17	**64.50**	55.67	60.78
	BERT_SEC + NUM	**63.67**	61.50	**58.67**	**61.28**

Table 10 Model comparison in terms of "F1-score" on testing data based on a 3-fold cross validation (#1, #2, #3 refers to the result of each fold), and the boldface is the winner.

Group	Model	F1-Score (%)			
		#1	#2	#3	Avg.
Baseline	LSTM (NUM)	54.63	43.75	40.44	46.27
	FastText	42.54	54.51	47.50	48.15
	FastText + NUM	58.69	56.34	54.91	56.65
Ours	BERT	56.09	58.47	**56.36**	56.97
	BERT + NUM	56.64	**60.13**	56.05	57.61
	BERT_SEC + NUM	**62.30**	60.10	56.34	**59.58**

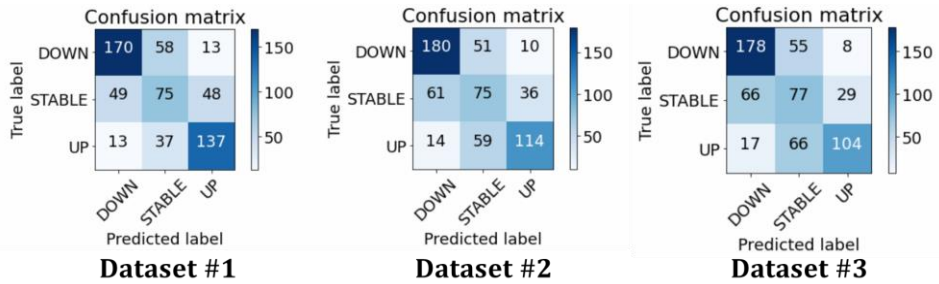

Figure 8 Confusion matrix of our sector-based model (BERT_SEC + NUM).

6.1 Effects of Transfer Learning

In this section, we aim to compare different language models: FastText vs. BERT. The results show that BERT outperforms FastText in both accuracy (60.72% vs. 55.22%) and F1 (56.97% vs. 48.15%). This demonstrates that BERT is the right choice for the textual module in our model. Unlike FastText, BERT can embed contextual information and handle unknown words by using subwords instead of a whole word. Furthermore, Multilingual BERT is pretrained with a huge training corpus.

6.2 Effects of Numerical and Textual Features

In this section, we aim to compare between an only textual model and a model with both information. The results confirmed that the model could be improved by using both information. For "BERT + NUM" vs. "BERT", the accuracy and F1 are 60.78% vs. 60.72% and 57.61% vs. 56.97%, respectively. For "FastText + NUM" vs. "FastText", the accuracy and F1 are 57.67% vs. 55.22% and 56.65% vs. 48.15%, consecutively.

6.3 Effects of Industry-Specific News Headlines (Sector)

In this section, we aim to analyze different embedding strategies for textual information. It should be more accurate if news headlines are embedded separately for each sector rather than the whole market. The results show that "BERT_SEC + NUM" outperforms "BERT + NUM" in both accuracy (61.28% vs. 60.78%) and F1 (59.58% vs. 57.61%). In conclusion, our sector-based model is the winner as it employs a contextualized language model (BERT), utilizes both sources of information, and handles textual information properly (sector-wise).

6.4 Annualized Return Based on Trading Simulation

As shown in Table 10, the results are favorable for dataset #2, while the other results are negative. However, the results should not be compared between the datasets because they are based on different market conditions. Therefore, we only compare the results for models that use the same dataset, even if we get negative returns. Nevertheless, our tests show that adding the numerical data increases the

model's annual return. The negative annual return may come from the trading strategies at this event. This strategy can lead to high transaction costs and negative returns when price action gains are less than trading costs. Moreover, the final model is selected based on the highest prediction accuracy on the validation data, leading to negative profits if the model makes accurate predictions during slight price movements. Moreover, a wrong forecast is made when the price changes significantly.

Based on EMH in Section 3.1, Table 10 shows that the average annualized return of the SET50 is -2.13%, which is the lowest return. Although it shows the highest return in dataset #1 (18.15%), it obtains the worst returns in datasets #2 and #3 (-7.17% and -17.76%, respectively). This demonstrates the volatility and unpredictability of the stock market; thus, it is crucial to have a trading strategy (e.g., a forecasting model) to help investors outperform the market. Furthermore, our sector-based model (BERT_SEC + TI) is a winner with the highest average annualized return (8.47%). The model is successful because it utilizes numerical data and textual data specifically for each industry (sector). Furthermore, the model using only historical data (LSTM (NUM)) gains the average annualized return at 2.47%, which outperforms SET50. Therefore, it can be concluded that Thailand's stock market efficiency is not considered a weak-form level.

Furthermore, when looking at investment behavior, models using only one type of information showed significantly more frequent trading behavior compared to our model. Thus, it can be concluded that our model has good investment behavior because it can effectively find clear signals to trade, thereby reducing the risks associated with investments and the fees associated with transactions. The result is an increase in the average annual return. Figure 9 showed investment behavior during trading simulation.

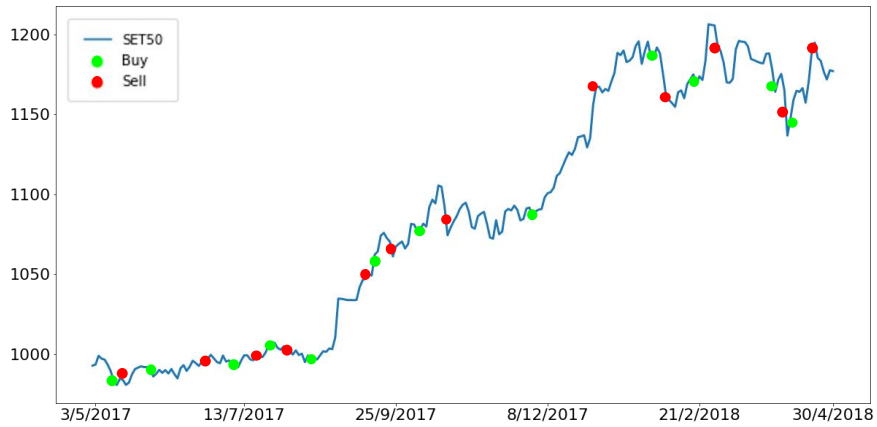

Figure 9 FINCIAL signals using our prediction model.

6.5 Interpretation and Textual Information Explainability

Initially, we chose the model presented as our priority implementation due to its visual explainability. We have applied self-attention to describe the model, in which weights Q, K, V are capable of linguistic recognition and can also be learned through model training. The result obtained after passing through Self-Attention is called Attention Weight. However, in architecture, many perspectives are considered to focus on the different priorities to cover. Therefore, we have adopted Multi-Head Self-Attention [52] to take advantage of the model's insights into the text input. We summarized the significance of all words (thousand levels) and listed the top 10 positive and negative words (contributed to predicting %yield) during the testing period as shown in Table 12.

Interestingly, we were able to extract some keywords, symbols, or tokens that the model typically observed participating in predictions. To summarize, the ability to interpret a model has two main benefits. First, it allows humans to validate the underlying insight. Secondly, it is also a debugging tool to aid in deep learning model development. For example, if the model emphasizes symbols like "@" or "&", that doesn't make much sense. We can go back to the cleaning process, removing it to reduce noise in the data. However, we still lack qualitative measurements of this model's interpretability, such as having domain experts to help validate those textual insights.

Table 11 Model comparison in terms of "Annualized Return" on testing data based on a 3-fold cross validation (#1, #2, #3 refers to the result of each fold), and boldface is the winner.

Group	Model	Annualized Return (%)			
		#1	#2	#3	Avg.
Baseline	SET50	**18.55**	-7.17	-17.76	-2.13
	LSTM (NUM)	8.40	11.10	-12.10	2.47
	FastText	6.90	8.60	-12.10	1.13
	FastText + NUM	15.30	-4.50	5.20	5.33
Ours	BERT	7.70	**12.00**	-11.90	2.60
	BERT + NUM	15.80	-3.90	6.30	6.06
	BERT_SEC + NUM	17.50	-3.30	**11.20**	**8.47**

Table 12 Top positive and negative words on test data in FINCIAL

Rank	(+)	English def.	(-)	English def.
1	ล้าน	[n] million	ไม่	[adj] no
2	แบงก์	[n] bank	กู้	[v] loan
3	รับ	[v] receive	หนี้	[n] debt
4	เพิ่ม	[v] increase	ลด	[v] decrease
5	ใหม่	[adj] new	ไทย	[n] Thailand
6	ปี	[n] year	ดอกเบี้ย	[n] interest
7	โต	[v] grow	ปล่อย	[v] release
8	สินเชื่อ	[n] credit	ยัง	[adv] yet
9	หนุน	[n] trig	เงิน	[n] money
10	ขึ้น	[v] rise	รัฐ	[n] government

CHAPTER 7

CONCLUSIONS

This research proposes a deep learning model to forecast the stock market trend (also called “a stock index”) based on both numerical (historical and technical indicators) and textual (news headlines) information. Since a stock index is a combination of many individual stocks from various sectors, we also propose to embed news into many industry-segment vectors. The experiments were conducted on SET50, a stock index in Thailand, along with news headlines in Thai. The results show that our sector-based model outperforms all baselines with an accuracy of 61.28% and F1 of 59.58%. Intensive experiments were provided to show that each proposed module can really improve the performance. Moreover, a trading strategy utilizing our prediction results was simulated. It achieves the highest annualized return of 8.47%.

For future studies, this research can be extended in two aspects. First, the model can be extended to other indexes, e.g., S&P500 in the United States and Nikkei 225 in Japan. However, the sector-based textual model must be tailored specifically for each market. Second, other external information can be integrated into our model to further improve the model’s performance. Especially during the epidemic (COVID-19) period, an announcement from the government and the number of infected patients can be included in our model.

REFERENCES

1. Fama, E.F., *Efficient market hypothesis.* Diss. PhD Thesis, Ph. D. dissertation, 1960.
2. Banz, R.W., *The relationship between return and market value of common stocks.* Journal of financial economics, 1981. **9**(1): p. 3-18.
3. Basu, S., *Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis.* The journal of Finance, 1977. **32**(3): p. 663-682.
4. Jegadeesh, N. and S. Titman, *Returns to buying winners and selling losers: Implications for stock market efficiency.* The Journal of finance, 1993. **48**(1): p. 65-91.
5. Amihud, Y. and H. Mendelson, *Liquidity and stock returns.* Financial Analysts Journal, 1986. **42**(3): p. 43-48.
6. Leigh, W., R. Purvis, and J.M. Ragusa, *Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural network, and genetic algorithm: a case study in romantic decision support.* Decision support systems, 2002. **32**(4): p. 361-377.
7. Mizuno, H., et al., *Application of neural network to technical analysis of stock market prediction.* Studies in Informatic and control, 1998. **7**(3): p. 111-120.
8. Gidofalvi, G. and C. Elkan, *Using news articles to predict stock price movements.* Department of Computer Science and Engineering, University of California, San Diego, 2001.
9. Gunduz, H. and Z. Cataltepe, *Borsa Istanbul (BIST) daily prediction using financial news and balanced feature selection.* Expert Systems with Applications, 2015. **42**(22): p. 9001-9011.
10. Schumaker, R.P. and H. Chen, *Textual analysis of stock market prediction using breaking financial news: The AZFin text system.* ACM Transactions on Information Systems (TOIS), 2009. **27**(2): p. 1-19.

11. Wang, B., H. Huang, and X. Wang, *A novel text mining approach to financial time series forecasting.* Neurocomputing, 2012. **83**: p. 136-145.
12. Ding, X., et al. *Deep learning for event-driven stock prediction*. in *Twenty-fourth international joint conference on artificial intelligence*. 2015.
13. Akita, R., et al. *Deep learning for stock prediction using numerical and textual information*. in *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*. 2016. IEEE.
14. Vargas, M.R., B.S. De Lima, and A.G. Evsukoff. *Deep learning for stock market prediction from financial news articles*. in *2017 IEEE international conference on computational intelligence and virtual environments for measurement systems and applications (CIVEMSA)*. 2017. IEEE.
15. Oncharoen, P. and P. Vateekul. *Deep learning for stock market prediction using event embedding and technical indicators*. in *2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA)*. 2018. IEEE.
16. Chiewhawan, T. and P. Vateekul. *Explainable deep learning for thai stock market prediction using textual representation and technical indicators*. in *Proceedings of the 8th International Conference on Computer and Communications Management*. 2020.
17. Hochreiter, S. and J. Schmidhuber, *Long short-term memory.* Neural computation, 1997. **9**(8): p. 1735-1780.
18. Devlin, J., et al., *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805, 2018.
19. Sun, C., et al. *How to fine-tune bert for text classification?* in *China National Conference on Chinese Computational Linguistics*. 2019. Springer.
20. Olah, C. and S. Carter, *Attention and augmented recurrent neural networks.* Distill, 2016. **1**(9): p. e1.
21. Staudemeyer, R.C. and E.R. Morris, *Understanding LSTM--a tutorial into Long Short-Term Memory Recurrent Neural Networks.* arXiv preprint arXiv:1909.09586, 2019.

22. Goldberg, Y. and O. Levy, *word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method.* arXiv preprint arXiv:1402.3722, 2014.
23. Hiew, J.Z.G., et al., *BERT-based financial sentiment index and LSTM-based stock return predictability.* arXiv preprint arXiv:1906.09024, 2019.
24. Malkiel, B.G., *The efficient market hypothesis and its critics.* Journal of economic perspectives, 2003. **17**(1): p. 59-82.
25. Wu, D., et al., *Stock prediction: an event-driven approach based on bursty keywords.* Frontiers of Computer Science in China, 2009. **3**(2): p. 145-157.
26. Xie, B., et al. *Semantic frames to predict stock price movement*. in *Proceedings of the 51st annual meeting of the association for computational linguistics*. 2013.
27. Ding, X., et al. *Using structured events to predict stock price movement: An empirical investigation*. in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.
28. Ding, X., et al. *Knowledge-driven event embedding for stock prediction*. in *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*. 2016.
29. Chang, C.Y., et al. *Measuring the information content of financial news*. in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2016.
30. Peng, Y. and H. Jiang, *Leverage financial news to predict stock price movements using word embeddings and deep neural networks.* arXiv preprint arXiv:1506.07220, 2015.
31. Luss, R. and A. d'Aspremont, *Predicting abnormal returns from news using text classification.* Quantitative Finance, 2015. **15**(6): p. 999-1012.
32. Skuza, M. and A. Romanowski. *Sentiment analysis of Twitter data within big data distributed environment for stock prediction*. in *2015 Federated Conference on Computer Science and Information Systems (FedCSIS)*. 2015. IEEE.

33. Sehgal, V. and C. Song. *Sops: stock prediction using web sentiment*. in *Seventh IEEE international conference on data mining workshops (ICDMW 2007)*. 2007. IEEE.

34. Lavrenko, V., et al. *Mining of concurrent text and time series*. in *KDD-2000 workshop on text mining*. 2000. Citeseer.

35. Xiong, G. and S. Bharadwaj, *Asymmetric roles of advertising and marketing capability in financial returns to news: Turning bad into good and good into great.* Journal of Marketing Research, 2013. **50**(6): p. 706-724.

36. Shi, L., et al., *Deepclue: Visual interpretation of text-based deep stock prediction.* IEEE Transactions on Knowledge and Data Engineering, 2018. **31**(6): p. 1094-1108.

37. Pennington, J., R. Socher, and C.D. Manning. *Glove: Global vectors for word representation*. in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.

38. Bojanowski, P., et al., *Enriching word vectors with subword information.* Transactions of the Association for Computational Linguistics, 2017. **5**: p. 135-146.

39. Ling, W., et al., *Finding function in form: Compositional character models for open vocabulary word representation.* arXiv preprint arXiv:1508.02096, 2015.

40. Kim, Y., et al. *Character-aware neural language models*. in *Thirtieth AAAI conference on artificial intelligence*. 2016.

41. Wehrmann, J., et al. *A character-based convolutional neural network for language-agnostic Twitter sentiment analysis*. in *2017 International Joint Conference on Neural Networks (IJCNN)*. 2017. IEEE.

42. Othan, D., Z.H. Kilimci, and M. Uysal. *Financial sentiment analysis for predicting direction of stocks using bidirectional encoder representations from transformers (BERT) and deep learning models*. in *Proc. Int. Conf. Innov. Intell. Technol.* 2019.

43. Tantisantiwong, N., et al., *Capturing investor sentiment from big data: The effects of online social media on set50 index.* CM Research Innovation, 2020. **2020**(4): p. 1-42.

44. Hu, Z., et al. *Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction*. in *Proceedings of the eleventh ACM international conference on web search and data mining*. 2018.

45. Zhai, Y., A. Hsu, and S.K. Halgamuge. *Combining news and technical indicators in daily stock price trends prediction*. in *International symposium on neural networks*. 2007. Springer.

46. Sezer, O.B. and A.M. Ozbayoglu, *Algorithmic financial trading with deep convolutional neural networks: Time series to image conversion approach.* Applied Soft Computing, 2018. **70**: p. 525-538.

47. Sutheebanjard, P. and W. Premchaiswadi. *Determining the time period and amount of training data for Stock Exchange of Thailand index prediction*. in *2010 2nd IEEE International Conference on Information and Financial Engineering*. 2010. IEEE.

48. Gupta, B.B. and Q.Z. Sheng, *Machine learning for computer and cyber security: principle, algorithms, and practices*. 2019: CRC Press.

49. Tzimas, M., et al., *Inference and Prediction of Nanoindentation Response in FCC Crystals: Methods and Discrete Dislocation Simulation Examples.* arXiv preprint arXiv:1910.07587, 2019.

50. Kingma, D.P. and J. Ba, *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980, 2014.

51. Ioffe, S. and C. Szegedy. *Batch normalization: Accelerating deep network training by reducing internal covariate shift*. in *International conference on machine learning*. 2015. PMLR.

52. Vaswani, A., et al. *Attention is all you need*. in *Advances in neural information processing systems*. 2017.

VITA

NAME	กิตติศักดิ์ ปรัชญาชูวงศ์
DATE OF BIRTH	16 กันยายน 2537
PLACE OF BIRTH	จังหวัดชลบุรี
INSTITUTIONS ATTENDED	วศ.บ. (เกียรตินิยมอันดับหนึ่ง) วิศวกรรมเครื่องกล มหาวิทยาลัยเกษตรศาสตร์ (พ.ศ. 2556-2560)
HOME ADDRESS	941 หมู่ 2 ถนนเทพารักษ์, เทพารักษ์, เมือง, สมุทรปรการ 10270