Artificial Neural Network Stock Price Prediction Model under the Influence of Big Data

Review of Integrative Business & Economics Research

Sakda Panwai
Don Muang Tollway Public Company Limited

ABSTRACT

Stock prices are highly nonlinear and random. Traditional time-series methods such as ARIMA and GARCH models are normally used. These models are effective only when the time-series is stationary, which is a restricting assumption and subject to model errors. The time-series models usually require the series to be log-transformed. This study applies a machine learning algorithm called Artificial Neural Network (ANN) to predict stock prices under the influence of several conditions. The model parameters include price-related, volume traded-related and Social Media effects. The proposed ANN models were tested to estimate and predict stock prices using two datasets from Stock Exchange of Thailand (SET). A 5-year dataset was used for model development (training and testing). A 1-month dataset was set aside for model validation only. The models were tested with/without Social Media effects. The trained/tested models produced a R-square of 0.98 whereas the validated models achieved a R-square of 0.63-0.69. It is important to note that the proposed model shows its robustness of prediction capability, showing a significant improvement by 16%-20%. The use of ANNs in predicting selected stocks is described and its robustness and capability in predicting stock prices are reported.

Keywords: Machine Learning, Artificial Neural Network (ANN), Stock Pricing Models

1. INTRODUCTION

1.1 Research Background

The Thai stock market in the Stock Exchange of Thailand (SET) has unique characteristics; a number of factors influencing the prices of stocks traded in SET are different from other markets. An example of the factors that influence the Thai stock market is foreign stock indexes, the value of the Thai Baht, the price of oil, the price of gold, the Minimum Loan Rate (MLR) and many others are found in Tantinakom (1996), Khumyoo (2000), Chotasiri (2004), Chaereonkithuttakorn (2005), Rimcharoen (2005), Sutheebanjard (2009) Worasucheep (2007), Phaisarn Sutheebanjard and Wichian Premchaiswadi (2010). Some research studies applied factors to forecast stock prices: trading value, trading volume, interbank overnight rate, inflation, net trading value of investment, value of the Thai Baht, price earnings ratio, the Dow Jones index, the Hang Seng index, the Nikkei index, the Straits Times index and the Kuala Lumpur Stock Exchange Composite index.

In 2000, Khumpoo (2000) used the Dow Jones index, the price of gold, the Hang Seng index, the exchange rate of the Japanese yen and the Thai baht, the Minimum Loan Rate (MLR), the Nikkei index, the price of oil, the Straits Times Industrial index and the Taiwan weighted index. In 2004, Chotasiri (2004) used the interest rate of Thailand and the USA, the exchange rate of USD, JPY, HKD and SGD, the stock exchange indexes of USA, Japan, Hong Kong and Singapore, the consumer price index, and the price of oil. In 2005, Chaereonkithuttakorn (2005) used the United States stock indices including the Nasdaq index, the Dow Jones index and the S&P 500 index. In 2005, Rimcharoen et al. (2005) used the Dow Jones index, the Nikkei index, the Hang Seng index, the price of gold and the Minimum Loan Rate (MLR). In 2007, Worasucheep (2007) used the Minimum Loan Rate (MLR), the exchange rate of the Thai Baht and the US dollar, daily effective over-night federal fund rates in the USA, the Dow Jones index and the price of oil. In 2008, Chaigusin, et al. used the Dow Jones index, the Nikkei index, the Hang Seng index, the price of gold, the Minimum Loan Rate (MLR) and the exchange rate of the Thai Baht and the US dollar. The common factors that researchers used to predict the SET index are summarized in Table 1. A comprehensive review can be also found in Phaisarn Sutheebanjard and Wichian Premchaiswadi (2010).

Those models were developed to achieve stock price patterns which is non-linear behavior regarding some certain factors, and to overcome limitations from the conventional models, like ARMA, ARIMA. Comprehensive study and review are referred to Ratnadip Adhikari, R. K. Agrawal (2013).

A number of applied science in engineering technology and applications using Artificial Intelligence found in Dia, H. (2001), Panwai, S. and Dia, H. (2005a), Panwai, S. and Dia, H. (2005b), Panwai, S. and Dia, H. (2006), Panwai, S. and Dia, H. (2007), Panwai, S. (2007), Dia, H. and Panwai, S. (2009a), Dia, H. and Panwai, S. (2009b), Dia, H. and Panwai, S. (2011), Wang, J., Indra-Payoong, N., Sumalee, A., and Panwai, S. (2014), Dia, H. and Panwai, S. (2014a), Dia, H. and Panwai, S. (2014b) have been scholastically recognized as AI-Based Behavioural Model. An intensive review for Artificial Intelligence (AI), Artificial Neural Network (ANN) and Fuzzy Logic, model development, model calibration and validation can be found in Panwai, S. (2007).

The machine learning in finance has been becoming increasingly important. Especially, machine learning algorithms are being used as an investment advice, trading on stock exchanges and gathering crucial information that might affect markets and investments, and will be an automated financial system. Related Works in AI applications in Finance were reported in Mojtaba Sedighi, Hossein Jahangirnia, Mohsen Gharakhani and Saeed Farahani Fard (2019) and Ratnadip Adhikari, R. K. Agrawal (2013).

However, no model takes into account the effects of data qualitative and quantitative. Moreover, issue of Social Media impacts on the stock price is quite new and there is a gap for prediction improvement. Unlike, this study applies Artificial Intelligence (AI) technique to learn stock market behavior using stock data (e.g. prices in different periods, volume traded, liquidity) and Social media data to understand how the stock price movement during certain events via Social Media. The data parameters will be described in the following sections.

Table 1 Impact Factors to Stock Exchange of Thailand Index (SET) Prediction (Phaisarn Sutheebanjard and Wichian Premchaiswadi (2010))

Index	Tanti-nakom (1996)	Khum-yoo (2000)	Chota-siri (2004)	Chaereon-kithuttakorn (2005)	Rim-charoen (2005)	Wora-sucheep (2007)	Suthee-banjard (2009)
Nasdaq Index				✓			
Dow Jones Index	✓	✓	✓	✓	✓	✓	✓
S&P 500 Index				✓			
Nikkei Index	✓	✓	✓		✓		✓
Hang Seng Index	✓	✓	✓		✓		✓
Straits Times Index	✓	✓	✓				
USD		✓	✓			✓	
JPY		✓	✓				
HKD			✓				
SGD			✓				
Gold price		✓			✓		
Oil price		✓	✓			✓	
MLR		✓			✓	✓	✓

Remarks: USD is the exchange rate of Thai Baht and the US dollar.
JPY is the exchange rate of Thai Baht and Japanese Yen.
HKD is the exchange rate of Thai Baht and Hong Kong dollar.
SGKD is the exchange rate of Thai Baht and Singapore dollar.

1.2 Gap Analysis: AI Models and Conventional Financial Models

Table 2 below presents gap analysis to compare between ANN and the conventional models. A number of aspects were discussed in other respective studies. It has been found that some areas of research are of interest e.g. qualitative and quantitative effects, Social Media effects, for this research study.

Table 2 Gap Analysis

Model Parameters/Description	Artificial Intelligent Models	Conventional Models
Models	ANN, Fuzzy Logic, Deep Learning, Panwai, S. (2007), Ratnadip Adhikari, R. K. Agrawal (2013), Mojtaba Sedighi, Hossein Jahangirnia, Mohsen Gharakhani and Saeed Farahani Fard (2019)	Autoregressive Moving Average (ARMA) Models Autoregressive Integrated Moving Average (ARIMA) Models Ratnadip Adhikari,R. K. Agrawal (2013)

Application	Applied Science, Engineering, Finance	Statistics, Finance
Data pattern	Non-linear behavior	Linear behavior
Accuracy Improvement	A number of networks, functions can be constructed and improved. Panwai, S. (2007)	Its complexity and limited increase in accuracy over less sophisticated methods (Sutheebanjard, P. and Premchaiswadi, W. (2016)
Model Interpretation	Black-Box	✓
Data hunger	✓	Limitation
Generalization	✖	✖
Deep Learning and on-line prediction	✓	✖
Big Data (e.g. Social Media effects)	✖	✖
Qualitative and quantitative impacts	✖	✖

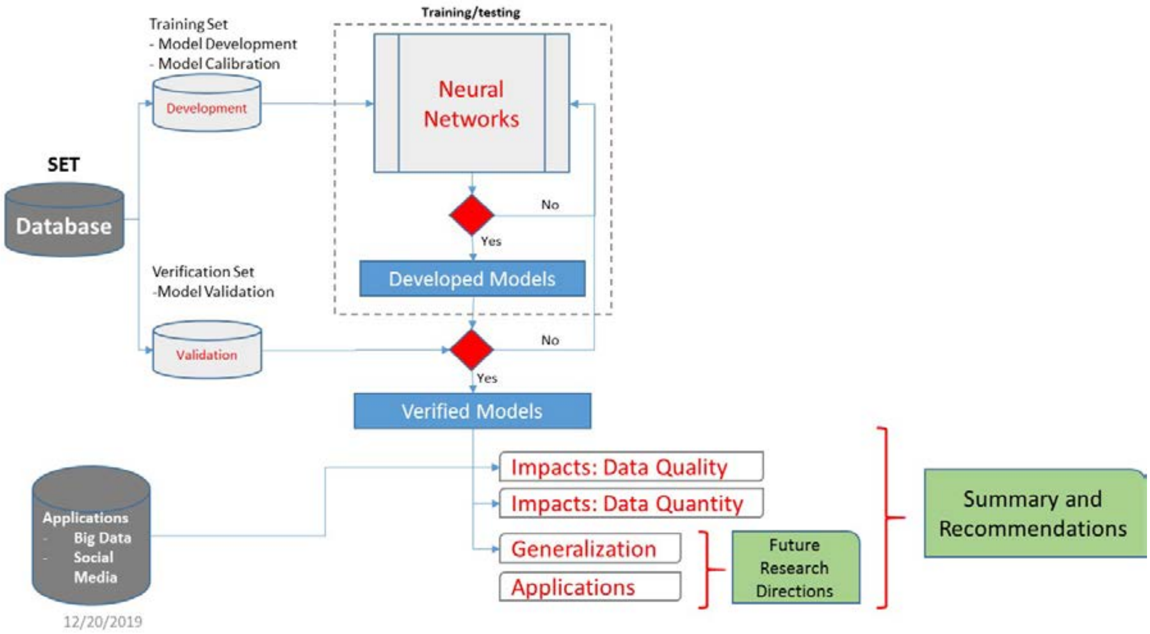
2. RESEARCH OBJECTIVES

This study aims to conduct Artificial Neural Network Stock Price Prediction Models to predict stock prices and to capture stock price behavior in SET Index. The primary objectives of the study are:

- To describe the datasets for model development, calibration and validation.
- To describe limitations in this study.
- To make a descriptive review of conventional models and Artificial Intelligence.
- To develop stock pricing models using an application of Artificial Neural Networks.
- To demonstrate impacts of data qualitative and quantitative and Social Media effects based on the developed models.
- To conclude findings in this study, and recommendation for future research study.

3. MODEL DEVELOPMENT FRAMEWORK

Figure 1 presents the study framework, which includes database, ANN models, model development, calibration, validation, generalization. SET database was gathered from SETSMART (to be explained later). The data is separated into two sets: one is used for model development while the other is set aside for model validation only. The contribution of this study is to apply Big Data technique to investigate the impacts related to Social Media (i.e. Twitter, Facebook, and etc.). Qualitative and quantitative data are also modeled. The models do not only take into account stock data, but also learn the Social Media effects. The methodology of this study is described next.



be applied for predicting the selected stocks price in the specified future time span. In conjunction with and without Social Media effects, the validated models can lean those effects, and then are discussed and reported.

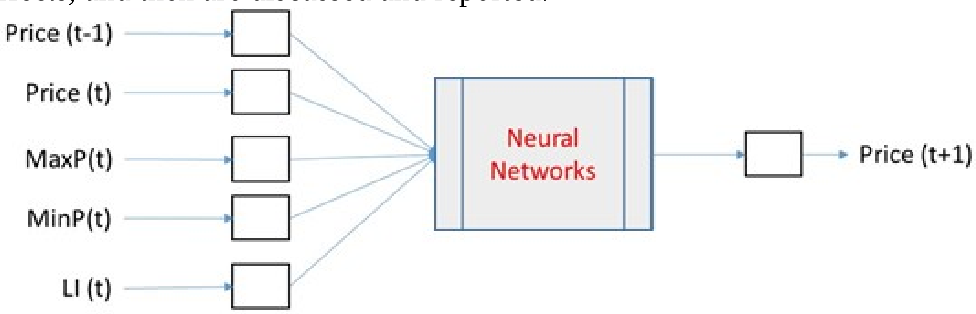


Figure 2 ANN Stock Price Prediction Model

3.1 Initial Stock Price Prediction Model Parameters

Table 3 below presents model parameters

Table 3 Model Parameters

Model input	Description
$P_{(t)}$:	Current stock price at time *t* , SETSMART data represented by “Open”
$P_{(t-1)}$:	Previous stock price at time *t-1*; SETSMART data represented by “Prior”
$MaxP_{(t)}$:	Previous maximum stock price at time *t;* SETSMART data represented by “Max”
$MinP_{(t)}$:	Previous minimum stock price at time *t;* SETSMART data represented by “Min”
$V_{(t)}$:	Current volume trade at time *t;* SETSMART data represented by “Total Volume”
$V_{(t-1)}$:	Previous volume trade at time *t-1;* SETSMART data represented by “Total Volume”
LI :	Liquidity Index is a ratio between Total volume traded by Outstanding shares.

SE:	Social Media (i.e. Twitter) Effect will be 0 represents No information 1 represents "Positive" information -1 represents "Negative" information
Model Output	
$P_{(t+1)}$:	Predicted stock price time *t+1*; SETSMART data represented by "Close"

3.1.1 Price-related

$P(t)$ represents open price;
$P(t-1)$ represents prior price;
$MaxP_{(t-1)}$ represents prior maximum stock price;
$MinP_{(t-1)}$ represents prior minimum stock price;
$P_{(t+1)}$ represents predicted close stock price;

3.1.2 Volume-traded

$V_{(t)}$ represents current total volume traded;
$V_{(t-1)}$ represents prior total volume traded;

Only v*olume traded parameters* will mislead the modeling results, as shown in the study conducted by Tomasz Kozdraj (2009). Therefore, this study applies *liquidity index* which derived from Volume traded per number of outstanding shares. This financial technical approach will eliminate the sizing differences effects. It is important to note here that the effect on the volume traded and liquidity will be discussed in Recommendation and Future Research Direction Section.

3.1.3 Social Media Effect

Social Media (i.e. Twitter) is denoted as "No information" or "Positive" or "Negative". These parameters were extracted using Python Script. Then a simply fuzzy rules: "No information" or "Positive" or "Negative" will be made in order to transfer the rules into the training data set. Other Social Medias are useful to enhance the proposed models and will be discussed in Recommendation and Future Research Direction Section.

4. SET DATA AND STATISTICS

4.1 Data Collection

Database from SET will be separately used for model development and model verification. Stocks in two different industries are selected: Siam Commercial Bank Public Company Limited (SCB.BK) which is Financial Sector and Italian-Thai Development Public Company Limited (ITD.BK) which is Property & Construction Sector, the data was used for this research study only. The two representative stocks were chosen to demonstrate feasibility of using Artificial Neural Networks (ANNs) in predicting stock price.

A 5-year data collection via SETSMART from 17/11/2014 to 15/11/2019 (1,221 observations) was applied for model development. Model training (977 observations) and testing (244 observations) were included in this process. A 1-Month data from

16/11/2019 – 15/12/2019 (17 observations) was set aside and applied for model verification only.

4.2 Dataset Profile and Descriptive Statistics

Table 4 and Table 5 present descriptive statistics for SCB.BK whereas Table 6 and Table 7 show descriptive statistics for ITD.BK. The results of the two stocks variations of MaxP and MinP values show highly correlated with P(t). In practice, it is very difficult to get MaxP and MinP during the trading day, unless when SET is closed. To reduce this complexity, only P(t) is used to present stock price for the trading day.

Table 4 Descriptive statistics (Quantitative data): SCB

Statistic	P(t-1) Prior	P(t) Open	MaxP High	MinP Low	P(t+1) Close	LI Liquidity
Number of observations	1221	1221	1221	1221	1221	1221
Minimum	106.500	107.500	110.500	104.500	106.500	0.037
Maximum	197.500	197.000	199.000	196.000	197.500	3.445
1st Quartile	133.500	134.000	135.000	132.500	133.500	0.137
Median	144.500	144.500	145.500	143.000	144.500	0.196
3rd Quartile	154.000	154.000	155.000	153.000	154.000	0.284
Mean	145.015	145.072	146.250	143.715	144.959	0.234
Variance (n-1)	255.977	254.126	254.056	253.566	255.218	0.029
Standard deviation (n-1)	15.999	15.941	15.939	15.924	15.976	0.171

Table 5 Correlation matrix (Pearson): SCB

Variables	Prior	Open	High	Low	Close	Liquidity
Prior	**1**	**0.998**	**0.996**	**0.995**	**0.992**	**-0.213**
Open	**0.998**	**1**	**0.998**	**0.997**	**0.994**	**-0.212**
High	**0.996**	**0.998**	**1**	**0.997**	**0.998**	**-0.196**
Low	**0.995**	**0.997**	**0.997**	**1**	**0.997**	**-0.224**
Close	**0.992**	**0.994**	**0.998**	**0.997**	**1**	**-0.206**
Liquidity	**-0.213**	**-0.212**	**-0.196**	**-0.224**	**-0.206**	**1**

Note: Values in bold are different from 0 with a significance level alpha=0.05

Table 6 Descriptive statistics (Quantitative data): ITD

Statistic	Prior	Open	High	Low	Close	Liquidity
No. of observations	1221	1221	1221	1221	1221	1221
Minimum	1.640	1.650	1.660	1.610	1.630	0.263
Maximum	9.450	9.500	9.600	9.300	9.450	106.451
1st Quartile	2.840	2.840	2.860	2.800	2.820	1.875
Median	4.460	4.480	4.500	4.420	4.460	4.007
3rd Quartile	6.950	7.000	7.100	6.850	6.950	10.319

Mean	4.899	4.906	4.973	4.831	4.895	8.232
Variance (n-1)	4.714	4.728	4.894	4.526	4.721	114.670
Standard deviation (n-1)	2.171	2.174	2.212	2.127	2.173	10.708

Table 7 Correlation matrix (Pearson): ITD

Variables	Prior	Open	High	Low	Close	Liquidity
Prior	**1**	**1.000**	**0.999**	**0.999**	**0.998**	**0.598**
Open	**1.000**	**1**	**0.999**	**0.999**	**0.999**	**0.596**
High	**0.999**	**0.999**	**1**	**0.999**	**0.999**	**0.610**
Low	**0.999**	**0.999**	**0.999**	**1**	**0.999**	**0.586**
Close	**0.998**	**0.999**	**0.999**	**0.999**	**1**	**0.601**
Liquidity	**0.598**	**0.596**	**0.610**	**0.586**	**0.601**	**1**

Note: Values in bold are different from 0 with a significance level alpha=0.05

Figure 5 presents Scattergram Data Plot for SCB.BK while Figure 6 presents Correlation Scatter Plot for SCB.BK. The similar results have been found in Table 4 and Table 5. In addition, LI: Liquidity Index has a negative corrected with stock price.

Figure 7 describes Scattergram Data Plot for ITD.BK whereas Figure 8 presents Correlation Scatter Plot for ITD.BK. The similar results have been found in Table 6 and Table 7. However, LI: Liquidity Index has a positive corrected with stock price.

4.3 Big Data

Besides financial data, information from other sources could be thought of "shock effect". In this study, Big Data from Social Media (i.e. Twitter) was used to feed into the ANN stock price prediction models.

For the purpose of research study only, Trump's phenomenon is an obvious case to present stock price movement, then this phenomenon had been closely monitored using Python Script. Then the data was classified either "No information" or "Positive" or "Negative" effect to the stock price. This study was conducted to understand how the developed models response to those data or events.

Information of Donald J. Trump's Twitter was fed into Python Script. Details can be found in Appendix, demonstrating how the Python Script gathers the information and classify into the defined "No information" or "Positive" or "Negative" effect. This data was daily monitored due to the fact that limitation of Standard user of Twitter unless a premium account service is subscribed. However, the study applies data availability to test the feasibility of the developed model. The information is set up as the following rules:

"1" if it is positive effect, stock daily return above a certain level.

"-1" if it is negative effect, stock daily return lower than a certain level.

"0" if it is between the two thresholds or has no information.

These three different rules are fed as an input parameter. The threshold can be adjusted for fine-tuning the proposed models. This limitation found in constructing the rules will be reported in Recommendation and Future Research Direction.

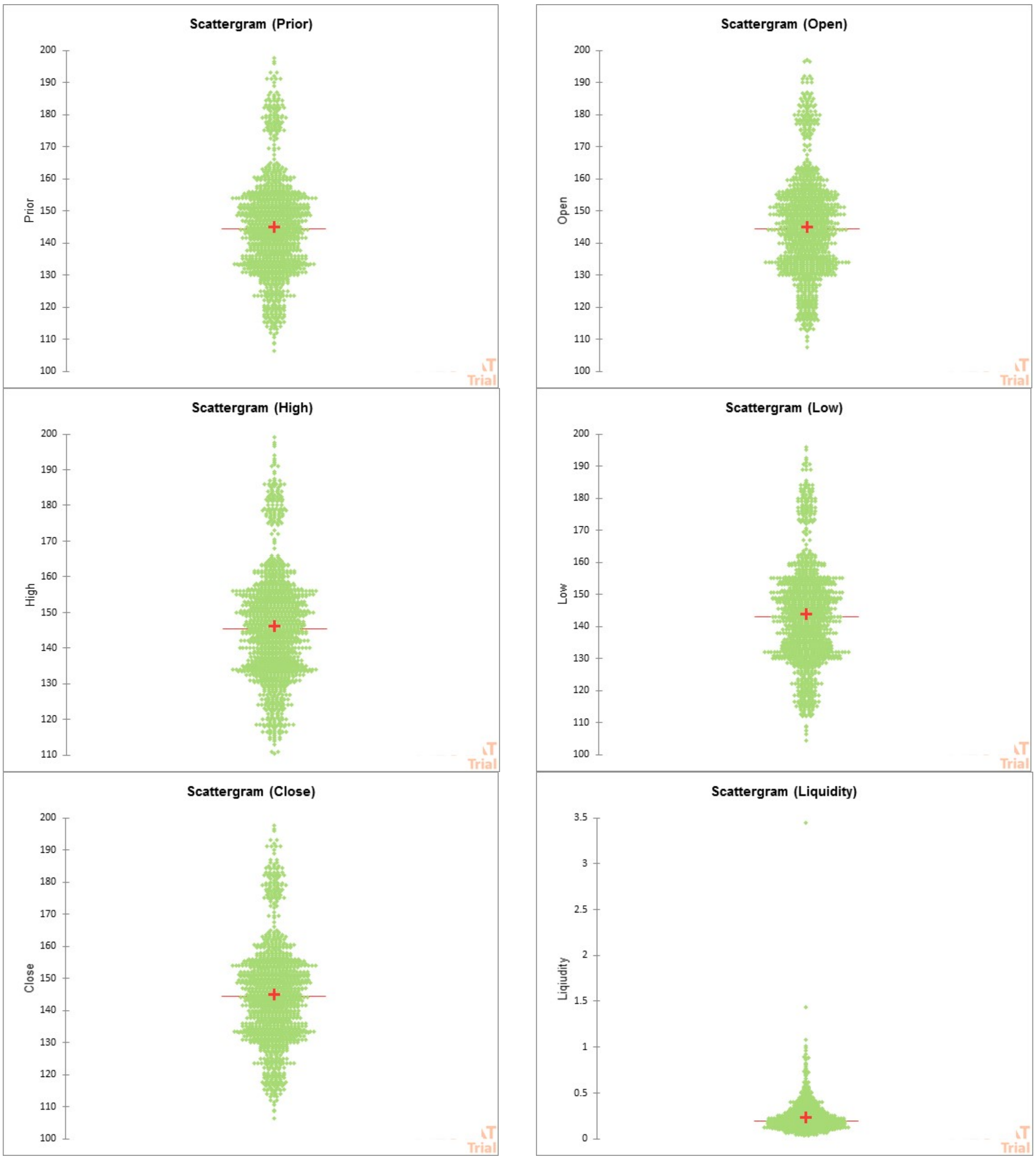


Figure 3 Scattergram Data Plot for SCB

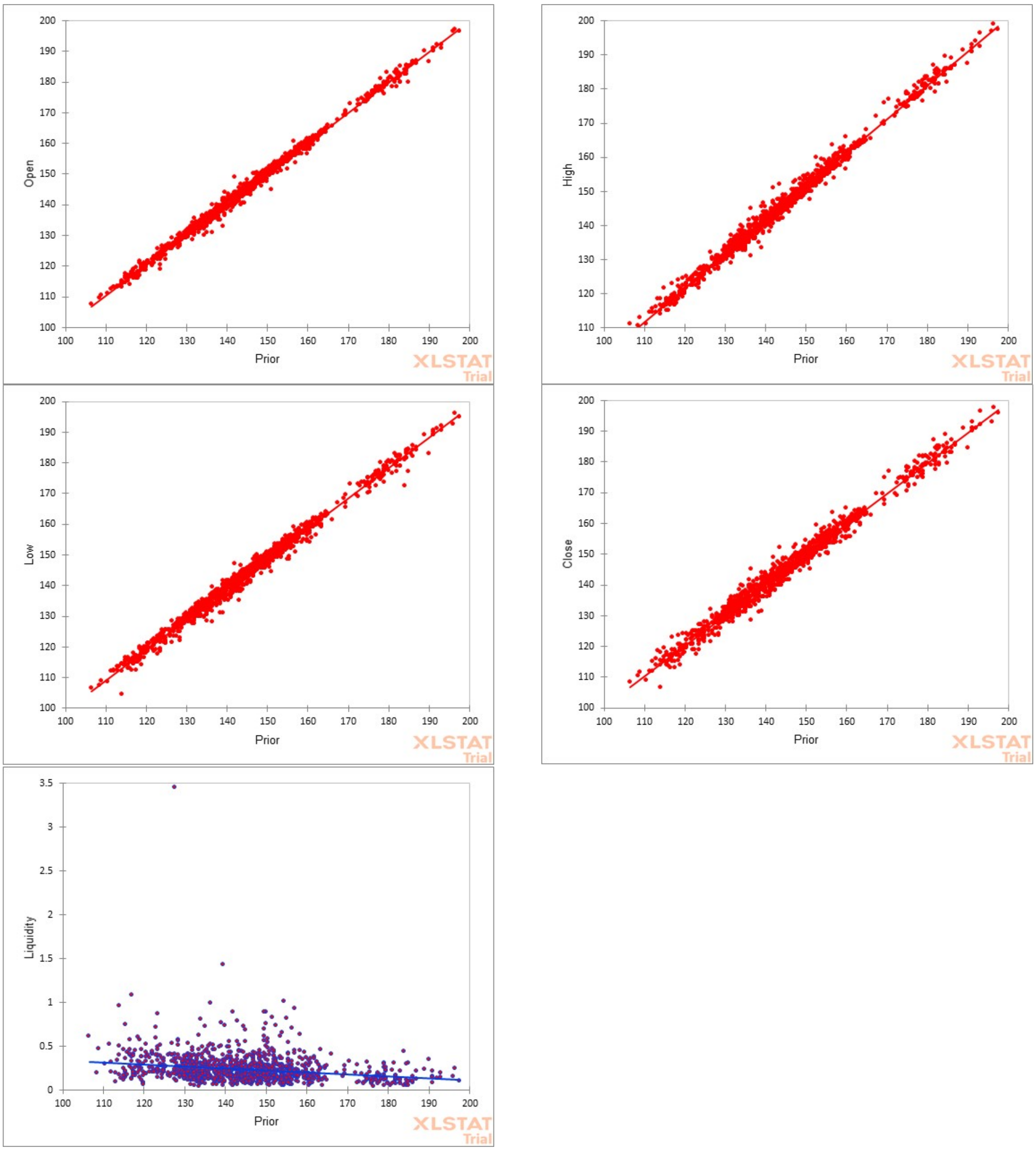


Figure 4 Correlation Scatter Plot for SCB

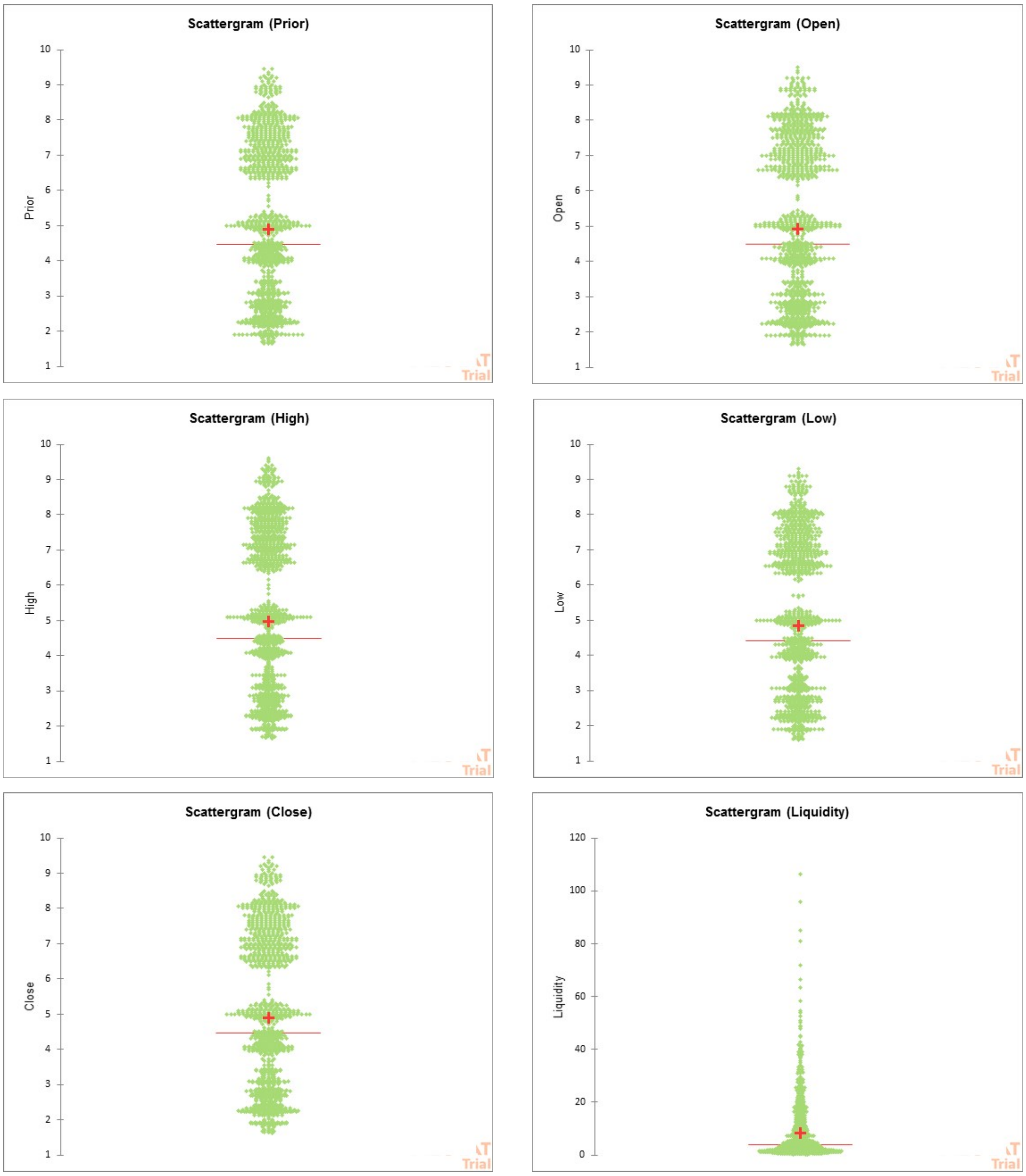


Figure 5 Scattergram Data Plot for ITD

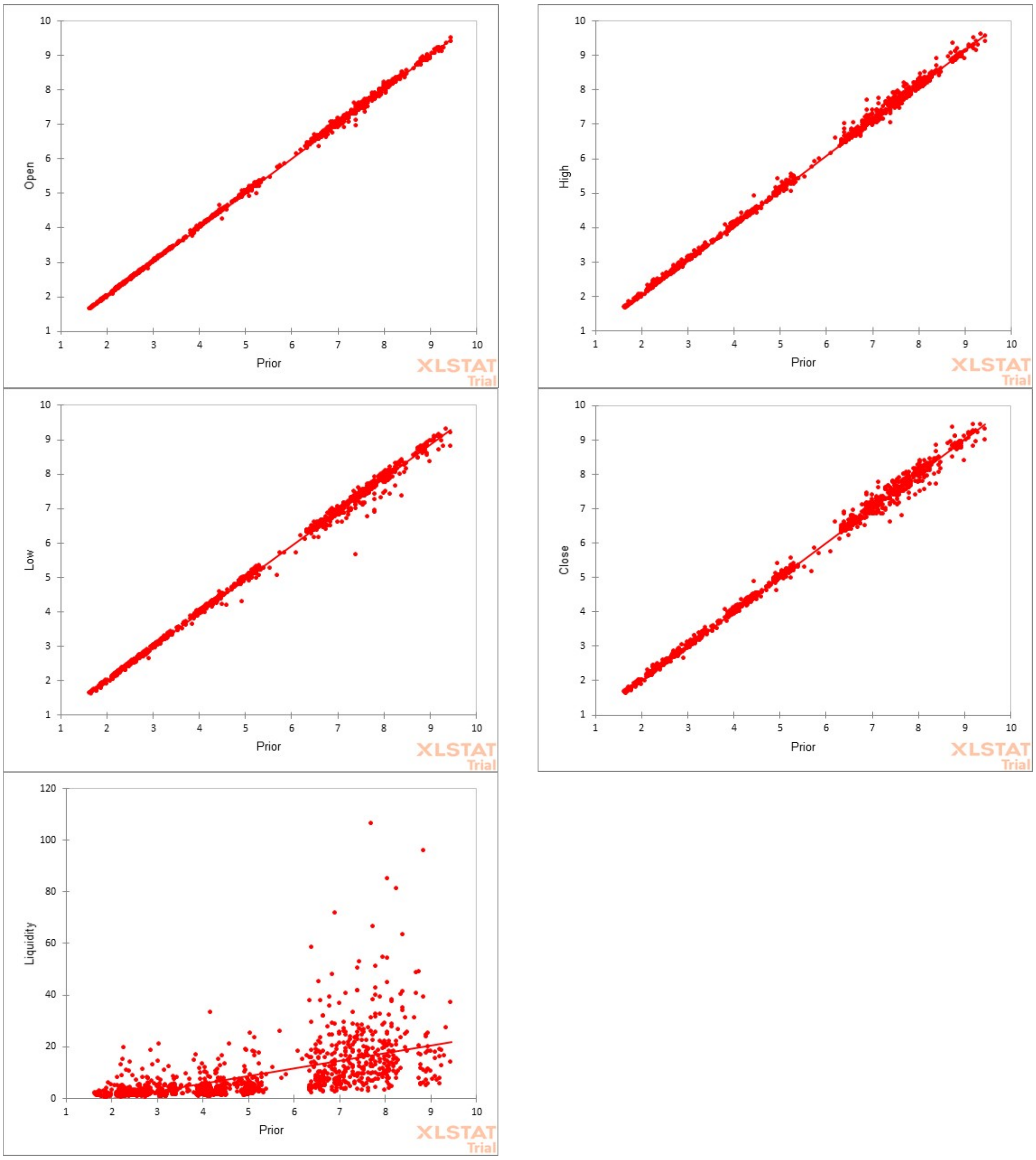


Figure 6 Correlation Scatter Plot for ITD

5. ARTIFICIAL NEURAL NETWORKS (ANNS) FUNDAMENTAL

Fundamental of a standard three-layered feed-forward neural network has been reported in Dia (1996) and applications in using ANN in reactive agents and cognitive agents found in Panwai (2007). Readers are referred to those respective papers.

6. ANN STOCK PRICE PREDICTION MODEL DEVELOPMENT

Figure 7 presents Stock Price Prediction Model architecture using Artificial Neural Networks. In previous section, it has been found that MaxP, MinP and P(t) are highly positive correlated, then for simplicity MaxP and MinP are eliminated and this will be discussed in Recommendation and Future Research Direction. The final architecture consists of Price (t-1), Price (t), LI(t), SE(t). They are fed into the ANN as input data in the Input Layer. The ANN takes input data to proceed into the Hidden Layer. For simplicity, the study applied standard NeuralTools functions to search for best-fit transfer functions and optimal number of hidden nodes in the Hidden Layer. Details of fine-tuning ANN Architecture found in Panwai (2007). After ANN processing with weight transferred into the Output Layer, the output which Price (t+t) can be reached. A number of trainings have been performed to receive the best model. Goodness-of-fit measured by Root Mean Square Error and R-square, which are used as key indicator.

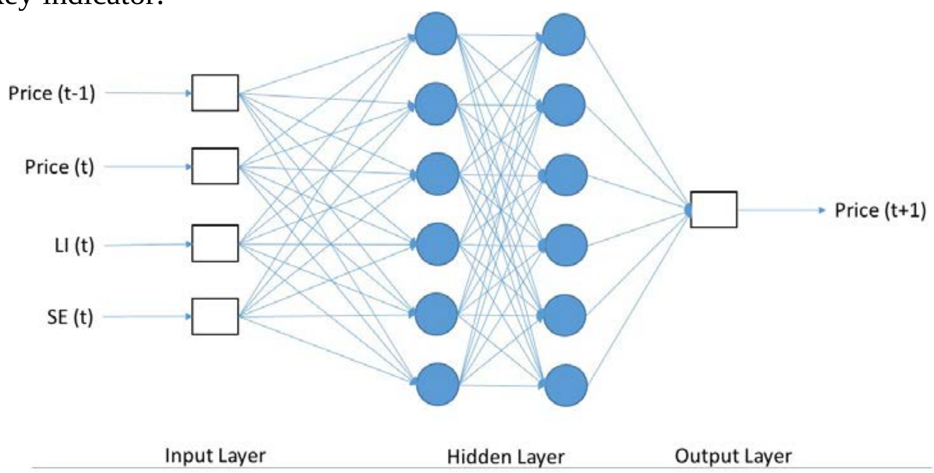


Figure 7 ANN Stock Price Prediction Model Architecture

6.1 Input

P(t-1) represents prior price; it is a set of real number consisting {0 to N}.

P(t) represents open price; it is a set of real number consisting {0 to N}.

LI represents Liquidity index for the selected Stock i, it is a set of float number consisting {0 to N}.

Social Effect can be classified into three categories: "No Information" is represented by "0", "Positive" is represented by "1" and "Negative" is represented by "-1". These parameters are extracted using Python Script. It is a set of {-1, 0, 1}.

6.2 ANN Process

There are two types of ANN Networks employed in this study:

PN/GNN Net with a category dependent variable. A probabilistic Neural Network will be trained. If the dependent variable is numeric, a generalized regression Neural Networks will be trained. PN and GNN Networks operate in a similar way. Every training case is represented by an element of the nets (a node A). A prediction for a case with an unknown dependent variable value is obtained by interpolation from training cases, with neighboring cases given more weight. Optional interpolation parameters are found during training.

MLF – Multi-Layer Feed Forward Network consists of an input layer of nodes, one or two layer of hidden layer. By selecting zero nodes, the second layer is eliminated, it is seldom needed for better prediction accuracy. Given by NeuralTools software, it can be auto-configuration the based on training data. If possible, use the more time-consuming Best Net Search to find the optimal configuration.

6.3 Output

$P_{(t+1)}$ represents predicted close stock price; it is a set of real number consisting {0 to N}.

7. INITIAL FINDINGS

A pilot test was purposely investigated to get preliminary results and to find out model performance and indicators. The findings will be discussed next.

7.1 Initial model results

The goodness-of-fit is described by R-Square and Mean Squared Error (MSE) which were used as a key indicator for both model calibration (trained/tested) and validation. The initial ANN model was conducted to demonstrate model performance. For this purpose, only SCB.BK dataset was used. Input parameters include P(t-1), P(t) and LI whereas output is only P(t+1). A 5-year SCB.BK dataset from 17/11/2014 - 15/11/2019 was applied to model stock price prediction, ANN embed in @RISK Software in MS Excel was applied.

Based on obtainability, ANN configuration used in this study is PN/GNN Net with a category dependent variable. The main advantage of PN/GNN Network is that, unlike MLF network, they do not require any configuration. At the same time their prediction accuracy is generally compatible to those MLF networks, learning process is quicker than MLF.

Neural Network simulation during learning process to achieve acceptable output (price level) is shown in Figure 8. It also presents a display of model results consisting of observed data parameters plotted against predicted parameter (testing), which randomly selected for testing (or cross-validation). The outcome is predicted price and can be interpreted as Good/Bad prediction as shown in the figure below.

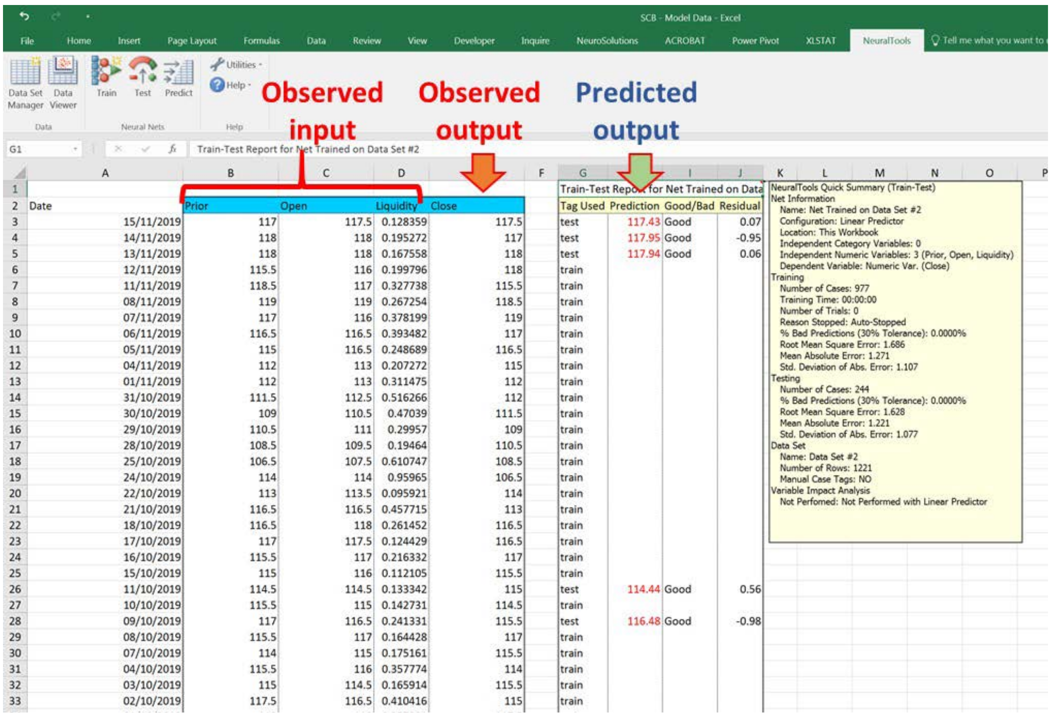
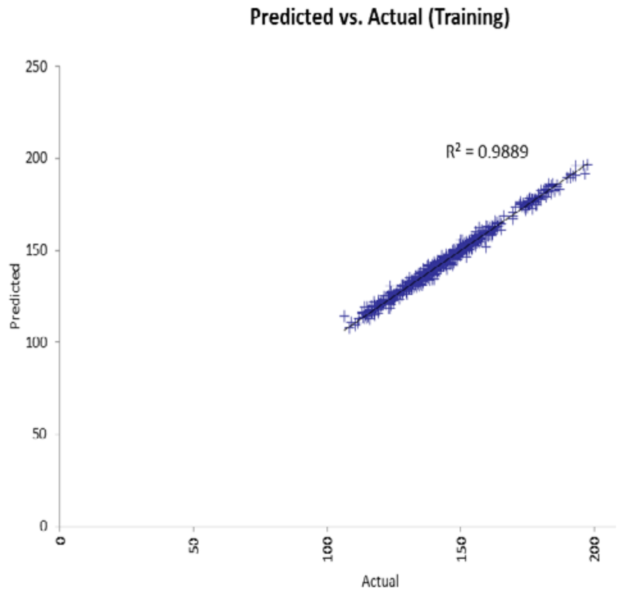


Figure 8 Initial results form NeuralTools

The initial performance are shown in Table 8. R-Square of 0.9889 was achieved. A lower RMSE of testing model than training model has shown its robustness, no over-training was found. Figure 9 presents to Figure 11 also describe the similar model results. Figure 9 presents model R-Square of model training (0.9889) and testing (0.9892). Figure 10 and Figure 11 describe model Residual for both training and testing models. The initial results present a merit of using ANN for this application.

Table 8 Prediction Measure of Performance

Neural Network	
R-Square (Training)	0.9889
Root Mean Sq. Error (Training)	1.671
Root Mean Sq. Error (Testing)	1.663



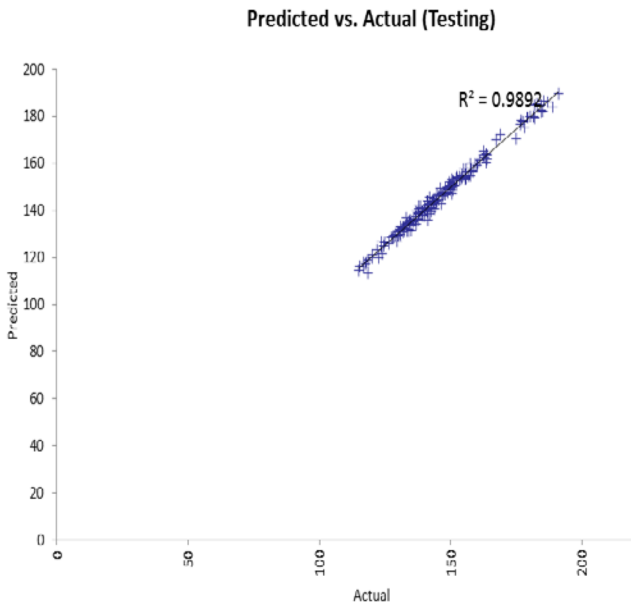


Figure 9 R-Square training model vs. testing model from NeuralTools

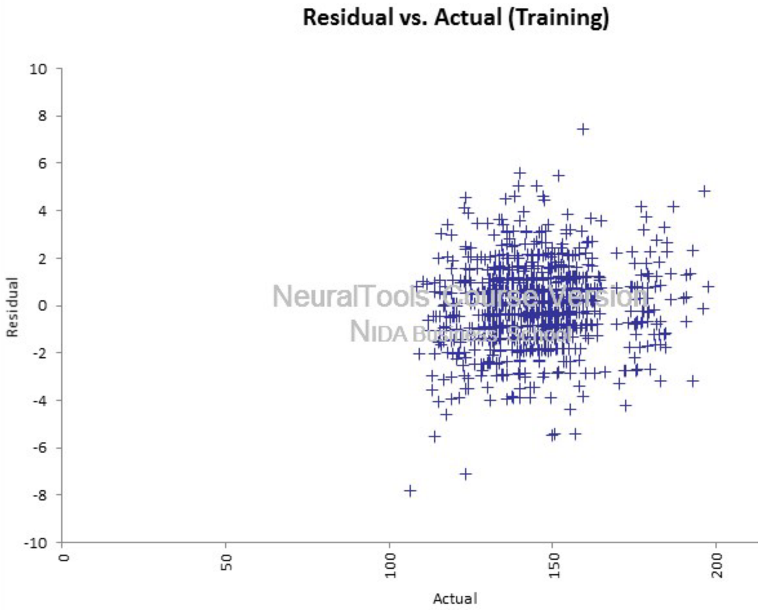


Figure 10 Residual training (Actual vs. Predicted) from NeuralTools

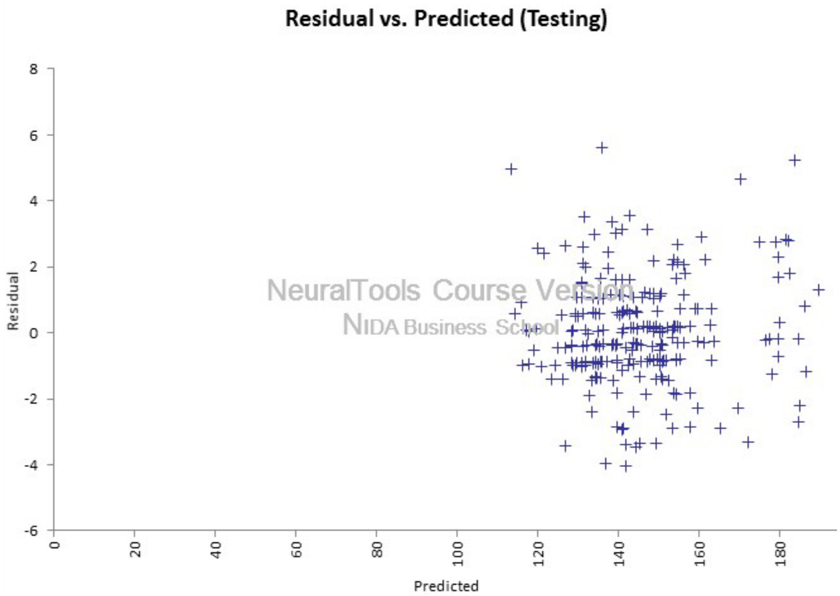


Figure 11 Residual Testing (Actual vs. Predicted) from NeuralTools

Train and Test: Effect on Data Sizes

To avoid over-training effect, the data set for training and testing (or cross-validation) has to be well-defined. A study done by Panwai, S. (2007) recommended that a combination of 70 percent for training and 30 percent for testing is a good structure. The initial results conducted in this study presented in Table 9 and Figure 12 confirmed the author's recommendation. Thirty percent of dataset for testing (cross-validation) presents the lowest RMSE. Therefore, this proportion between training dataset and testing dataset was applied in this study.

Table 9 RMSE: Data Size for Cross-Validation

10%	20%	30%
1.484	1.667	1.614
1.600	1.667	1.717
1.719	1.718	1.717
1.753	1.747	1.789
2.568	1.952	1.828

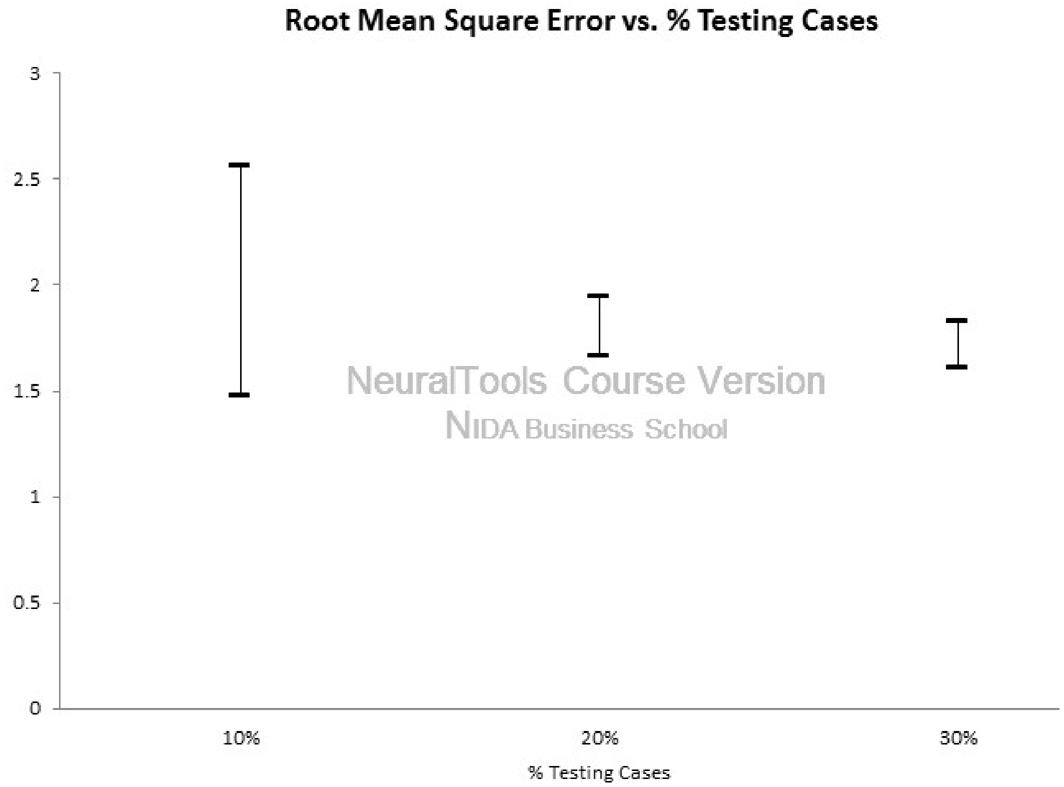


Figure 12 Error Plot of RMSE of Data Size for Cross-Validation (10% - 30%)

8. MODEL REFINEMENT AND SHOCK EFFECTS

The proposed ANN stock price prediction models were trained and tested using 70% of dataset and 30% of dataset, respectively. The data points were automatically randomly-selected. A number of trains and tests were performed. The models with the highest R-Square representing a-goodness-of-fit, then were selected as the best model.

The trained/tested models were then validated using the validation dataset. These were separated into two conditions: (1) under no influence of Social Media Effects and (2) under the influence of Social Media Effects.

8.1 No Social Effect Information

At this stage, the proposed models were trained and tested in NeuralTools embedded in MS Excel. Functions such as dataset management, statistics data view, training, testing and prediction were used to set up the process. A limit number of ANN architectures were tested. The study applied the best search algorithm to get the best combination of transfer functions and a number of nodes in hidden layer.

As a result, trained model for SCB.BK produces R-Square of 0.989, and performed well for tested model too. The R-Square of the tested model is 0.989. The same tendency was found for ITD.BK. R-Square for the trained model is 0.9974 while the tested model performs R-Square of 0.9968. These findings are presented in Table 10 and Figure 13 to Figure 14.

Table 10 also presents prediction measure of performance. RMSEs are ranged around 1.677-1.668 for SCB.BK and 0.1107-0.1230 for ITD.BK.

Table 10 Prediction Measure of Performance – No Social Effect Information

Neural Net – Results	SCB.BK - Neural Net	ITD.BK - Neural Net
R-Square (Training)	0.9890	0.9974
Root Mean Sq. Error (Training)	1.677	0.1107
Root Mean Sq. Error (Testing)	1.668	0.1230

Figure 13 and Figure 14 also depict the good model performance for SCB.BK and ITD.BK, respectively. The dataset of 5-consecutive years for training (977 observations) and testing (244 observations) are considerably sufficient. The proposed models embedded ANN performed well in this environment. Its capability of mapping data pattern to predict the outcome in the next step showed acceptable results.

Thereafter, the tested models were validated using dataset for validation (18 observations) from 18 November 2019 – 13 December 2019. The validated SCB.BK model produced R-Square of 0.6987 whereas the validated ITD.BK model gave 0.6369 shown in Figure 15. The validation of the two stocks are quite understandable. As a model assumption, (1) ANN will learn and forecast what they have been trained and (2) during the training period, no information about Social Media effect is fed into the models. Nevertheless, it is obvious that the proposed ANN prediction models show its robustness of prediction capability.

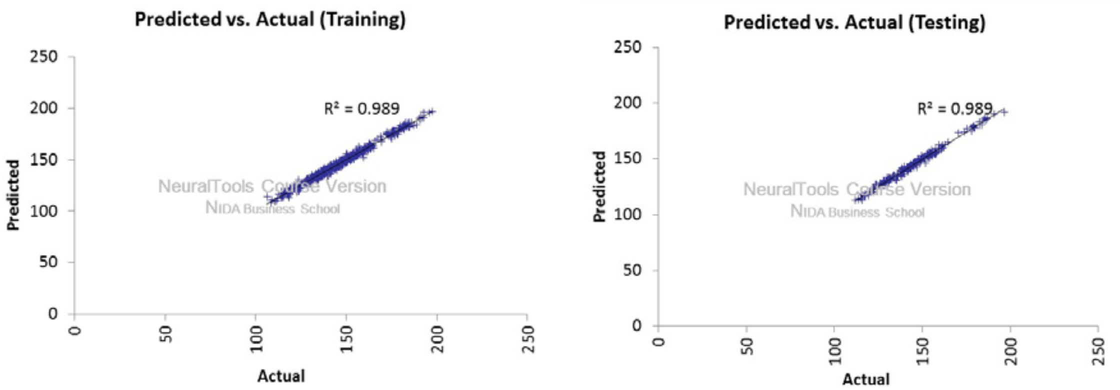


Figure 13 R-Square training model vs. testing models – SCB.BK

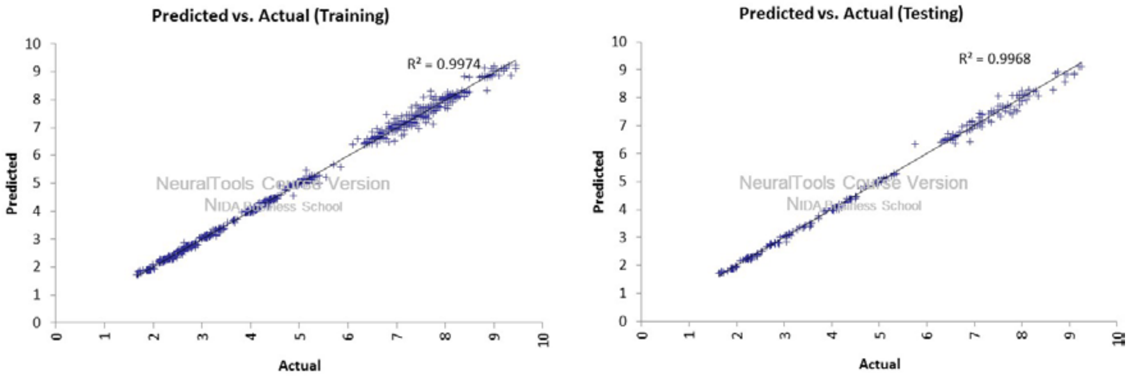
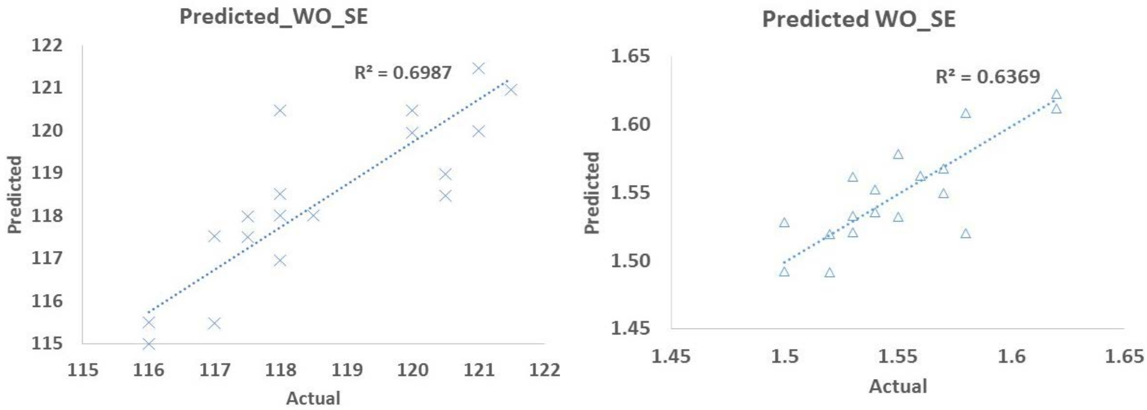


Figure 14 R-Square training model vs. testing models – ITD.BK



SCB.BK Model Validation ITD.BK Model Validation

Figure 15 Validation results using No Social Media Effect parameters

8.2 Social Media Effect Information

To improve the model performance, the effects of Social Media was used as input parameter and fed into the ANN architecture. The trained models were re-trained again. A same past 5-year dataset was used to train, and also applied for searching an indicator to trigger to ask for the Social Media data needed.

A set of {-1, 0, 1} in relation with a pre-defined stock return threshold were fed into the past 5-year dataset to allow the ANN prediction model to learn those events.

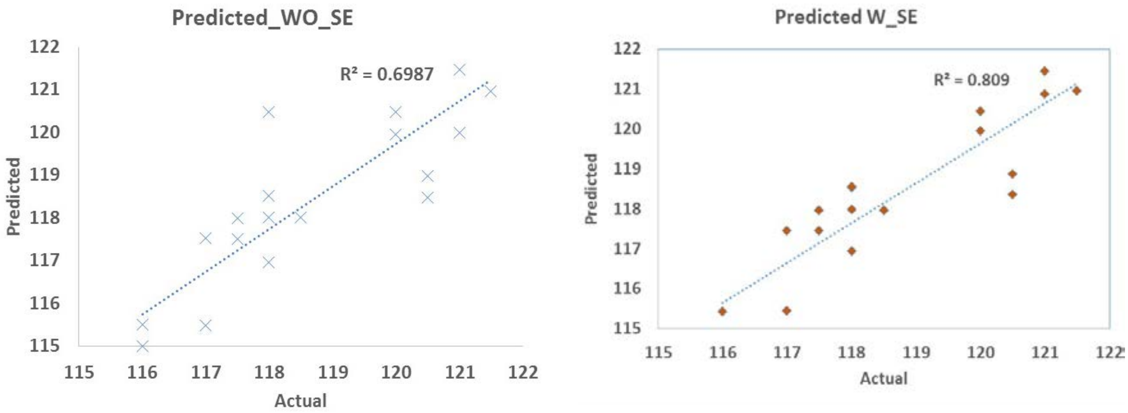
The dataset for validation (18 observations) from 18 November – 13 December 2019 was re-constructed to incorporate with Social Media effects. A simple rule-based was used to demonstrated the effect of BigData or Social Media. A set of {-1, 0, 1} can be obtained as follows:

Simple Rule-Based Search

If (Stock return > x%, calls Social Media data,
searches for set of POSITIVE "xxx" and set of NEGATIVE "yyyy" then
if "xxx", then 1
if "yyy", then -1 otherwise 0

Figure 16 illustrates the results of the well-trained/tested SCB.BK models that were validated using validation dataset. It is very important to note here that without

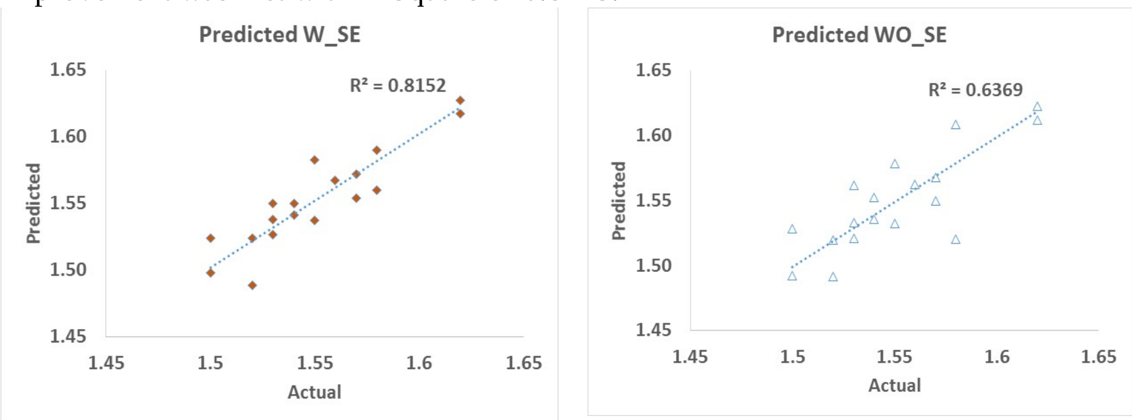
Social Media Effect information meaning that the proposed ANN models had not been learnt in the training stage. The proposed ANN model was able to predict SCB stock price with R-Square of 0.6987. After the process of transferring data from Twitter and extracting into the rules {-1, 0, 1}, model improvement was achieved with R-Square of 0.809.



Without Social Media Effects Parameters With Social Media Effects Parameters

Figure 16 SCB - Model results improvement using Social Media Effect parameters into ANN Stock Price Model

Figure 17 presents the results of the well-trained/tested ITD.BK models that were fed into the validation dataset. Without Social Media Effect information, the model R-Square was 0.6369. After the process of transferring data from Twitter and extracting the rules {-1, 0, 1} and set as input parameter in the ANN models, model improvement was met with R-Square of 0.8125.



Without Social Media Effects Parameters With Social Media Effects Parameters

Figure 17 ITD - Model results improvement using Social Media Effect parameters into ANN Stock Price Model

9. CONCLUSION

The study has conducted the Artificial Neural Network models to predict price for the selected stocks. Siam Commercial Bank Public Company Limited (SCB.BK) which is Financial Sector and Italian-Thai Development Public Company Limited (ITD.BK) which is Property & Construction Sector. They are in Stock Exchange of Thailand (SET).

Data collection made via SETSMART was gathered from 17/11/2014 - 15/11/2019 (1,221 observations or trading days). For general ANN models, careful selection of data set is very important and plays a key role for achievement. This study did not analyze this effects, but rather use a number of established works such as Panwai, S. (2007), Ali SORAYAEI, Zahra ATF and Masood GHOLAMI (2016). Based on the achievements, 70% data was used to training network and remaining 30% was set aside for cross-validation. The same results were also found in this study.

ANN architectures, functions and parameters can be selected to fine-tune the model performance, and it is not scope of this study. This study used NeuralTools embedded in MS Excel and applied a default best search algorithm. The study randomly set up dataset of 70% for training (977 observations) and 30% for validation (244 observations) to conform the mentioned findings. A 1-Month data from 16/11/2019 – 15/12/2019 (18 observations) was set aside and applied for model verification only.

The proposed models were then trained and tested under no influence of Social Media effect. The SCB.BK trained/tested models produced R-Square of 0.989 and 0.989, respectively whereas The ITD.BK trained/tested models gave R-Square of 0.9974 and 0.9968. The models were then validated using the validation dataset. The SCB.BK model produced R-Square of 0.6987 whereas the ITD.BK model gave R-Square of 0.6369. According to the fact that ANN model can learn and forecast whatever they have been trained before and (2) during the training period, no information about Social Media effect was fed into the models. Nevertheless, it is important to note here that the proposed ANN prediction models illustrated its robustness of prediction capability.

The models were then re-trained/tested again under the influence of Social Media effect conditions. A set of {-1, 0, 1} in relation with a pre-defined stock return threshold were fed into the training dataset to allow the ANN prediction models to learn those events. The dataset for validation was re-constructed to incorporate with Social Media effects. A simple rule-based was used to demonstrated the effect of BigData or Social Media effect. A set of {-1, 0, 1} can be obtained as "Negative" or "No information" or "Positive": After the process of transferring data from Twitter and extracting into the rules {-1, 0, 1}, model improvement was met with R-Square of 0.809 (SCB.BK) and 0.8125 (ITD.BK). The improvement ranged from 16%-20%.

The study has demonstrated the use of ANN application for stock price prediction. Researchers or students in particular field can be beneficial to develop and enhance the prediction with cares as discussed in the next section.

10. RECOMMENDATION AND FUTURE RESEARCH DIRECTION

The study showed a process-oriented tasks of model development. The findings in this research showed limitations and demonstrations of using Artificial Neural Networks in predicting selected stocks. The tasks or processes that were found and recommended for future research direction are:

1) This study used data available from SETSMART. Time (t) in this study represent one day. P(t-1) is a prior price which means one day earlier whereas P(t) is open price representing current time step. P(t+1) represents close price at the end of the day. The stock price prediction at the end period will be very precise. However, the proposed models can be used to fit with a smaller time-slice such as hourly basis.

2) Other model parameters should be studied in deep detailed analysis to understand their behavior such MaxP and MinP during the trading day. In reality, one cannot find out MaxP and MinP before the market close. Order imbalance issue remains for future research direction. The order imbalance will affect stock price prediction. Liquidity can be measured in various ways. These impacts are proposed for future research direction.
3) With a smaller time slice data, those parameters and effects can be investigated. Other sources of SET data should be then considered to smaller time *(t)* period to improve the proposed models and to deal with the dynamic nature of Social Media effects or BigData. Moreover, this is recommended for further study.
4) Pre-defined rule-based construction is a state-of-the-art and a time-consuming task. This study applied trial-and-error basis, a few number of thresholds were tested until the models obtained a better performance. The study showed about 16% to 27% improvement. The SCB.BK model's R-Square (trained was 0.9890, tested constituted 0.9968 and validation without Social Effect provided 0.6987) achieved 0.8090, and improved by 16% while the ITD.BK model's R-Square (trained was 0.9974, tested constituted 0.9890 and validation without Social Effect provided 0.6369) achieved 0.8125, and improved by 27%. However, more rules in relation with various thresholds can be constructed using fuzzy logic concept. This approach can be used to obtain the best performance. Other sources of data or BigData will be useful to enhance model development. This is also aimed for future research or commercial study.
5) Industry-related issue detailed analysis of industry difference is of interest. The future study should collect all companies in SET and analyze on sector-by-sector basis to find out in which sector is very sensitive to Social Media effects.
6) Online data stream approach: the study has demonstrated the off-line BigData i.e. Twitter. Online-testing with wider range of BigData will be recommended for future research direction. Some industries are more sensitive to Social Media effect, but some other may not. The proposed models would be a more accuracy and this could be a commercial concern.

ACKNOWLEDGEMENT

I would like to express my appreciation to all those who provided me the possibility to complete this report and master degree. A special gratitude I give to my boss and my workplace, Don Muang Tollway Public Company Limited in providing financial support.

REFERENCES

[1] Ali SORAYAEI, Zahra ATF and Masood GHOLAMI (2016), Prediction stock price using artificial neural network , Bulletin de la Société Royale des Sciences, Vol. 85, 2016, p. 991 – 998

[2] C. Khumyoo, "The Determinants of Securities Price in the Stock Exchange of Thailand," Master Thesis in Economics, Ramkhamhaeng University, Bangkok, Thailand, 2000.

[3] C. Worasucheep, "A New Self Adaptive Differential Evolution: Its Application in Forecasting the Index of Stock Exchange of Thailand," Evolutionary Computation, 2007. CEC 2007, 2007.
[4] Dia (2001). An object-oriented neural network approach to short-term traffic forecasting, Vol. 131, Issue 2, P.253-261, Publishing North-Holland
[5] Dia, H. and Panwai, S. (2014) *Intelligent Transport Systems: Neural Agent (Neugent) Models of Driver Behaviour*. LAP LAMBERT Academic Publishing, Deutschland, Germany
[6] Dia, H. and Panwai, S. (2014) *Intelligent Mobility for Smart Cities: Driver Behaviour Models for Assessment of Sustainable Transport*. (2014). Fourth International Conference on Big Data and Cloud Computing *(BdCloud 2014)*, 3-5 December 2014, Sydney, Australia.
[7] Wang, J., Indra-Payoong, N., Sumalee, A., and Panwai, S. (2014). Vehicle Reidentification With Self-Adaptive Time Windows for Real-Time Travel Time Estimation. *IEEE Transactions on Intelligent Transportation Systems*, *15*(2).
[8] Dia, H. and Panwai, S. (2011) *Neural Agent (Neugent) Models of Driver Behavior for Supporting ITS Simulations*. International Journal of ITS Research, Vol.9, Issue 1, January 2011.
[9] Dia, H. and Panwai, S. (2009) *Evaluation of Discrete Choice and Neural Network Approaches for Modelling Driver Behaviour*. Transportmetrica (August 2009), pp.1-22.
[10] Dia, H. and Panwai, S. (2009). *Models of Driver Behaviour for Supporting Vehicle Telematics and Intelligent Transport Systems Simulations*. Proceedings of the 10th Intelligent Transport Systems Asia Pacific Forum & Exhibition 2009, 8-10 July 2009, Queen Sirikit National Convention Center, Bangkok, Thailand
[11] Dia, H. and Panwai, S. (2007) *Modelling Drivers' Compliance and Route Choice Behaviour in Response to Travel Information*. Special issue on Modelling and Control of Intelligent Transportation Systems, Journal of Nonlinear Dynamics, vol. 49 no. 4, September (Springer).
[12] K. Chaereonkithuttakorn, "The Relationship between the Stock Exchange of Thailand Index and the Stock Indexes in the United States of America," Master Thesis in Economics, Chiang Mai University, Chiang Mai, Thailand. 2005.
[13] Mojtaba Sedighi, Hossein Jahangirnia, Mohsen Gharakhani and Saeed Farahani Fard (2019), A Novel Hybrid Model for Stock Price Forecasting Based on Metaheuristics and Support Vector Machine, MDPI, Received: 2 May 2019; Accepted: 20 May 2019; Published: 22 May 2019
[14] Phaisarn Sutheebanjard and Wichian Premchaiswadi (2016), Stock Exchange of Thailand Index prediction using Back Propagation Neural Networks, Second International Conference on Computer and Network Technology
[15] Phaisarn Sutheebanjard and Wichian Premchaiswadi (2010). Stock Exchange of Thailand Index prediction using Back Propagation Neural Networks, Second International Conference on Computer and Network Technology, IEEE Computer Society.
[16] P. Sutheebanjard and W. Premchaiswadi, "Factors Analysis on Stock Exchange of Thailand (SET) Index Movement," The 7th International Conference on ICT and Knowledge Engineering, ICTKE2009, Bangkok, Thailand, December, 2009.
[17] Panwai, S. (2007) Modelling Driver Behaviour under the Influence of Traffic Information. Ph.D Thesis, The University of Queensland, Australia.

[18] Panwai, S. and Dia, H. (2007) *Neural Agent Car Following Model.* IEEE Transactions on Intelligent Transportation Systems vol. 8 no. 1, pp. 60-70.

[19] Panwai, S. and Dia, H. (2006). *Development of Agent-Based Dynamic Route Choice Driver Behaviour Models.* Proceedings of the Agents in Traffic and Transportation Workshop, Conference on Autonomous Agents and Multi Agent Systems, 2006, Hakodate, Japan, pp. 70-79.

[20] Panwai, S. and Dia, H. (2005). *A Reactive Agent-Based Neural Network Car Following Model*. Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems (ITSC2005), 13-16 September 2005, Vienna, Austria (ISBN: 0-7803-9215-9/05), pp. 326-331. [Acceptance Rate: 67 per cent]

[21] Panwai, S. and Dia, H. (2005). *Development and Evaluation of a Reactive Agent-Based Car Following Model*. Proceedings of the Intelligent Vehicles and Road Infrastructure Conference (IVRI '05), 16th and 17th February 2005, Melbourne, Australia (ISBN 0-908556-79-9).

[22] Ratnadip Adhikari, R. K. Agrawal (2013), An Introductory Study on Time Series Modeling and Forecasting, LAP Lambert Academic Publishing, Germany, 2013.

[23] S. Rimcharoen, D. Sutivong and P. Chongstitvatana, "Prediction of the Stock Exchange of Thailand Using Adaptive Evolution Strategies," Tools with Artificial Intelligence, 2005. ICTAI 05. 17th 2005

[24] S. Chotasiri, "The Economic Factors Affecting the Fluctuation of The Stock Exchange of Thailand Index," Master Thesis in Economics, Chiang Mai University, Chiang Mai, Thailand, 2004.

[25] S. Chaigusin, C. Chirathamjaree and J. Clayden, "The Use of Neural Networks in the Prediction of the Stock Exchange of Thailand (SET) Index," Computational Intelligence for Modelling Control & Automation. 2008.

[26] Tomasz Kozdraj (2009). Using Artificial Neural Networks to Predict Stock Prices, ACTA Universitatis Lodziensis, Folia Oeconnomica 225, 2009, p. 280-293

[27] T. Tantinakom, "Economic Factors Affecting Stock Exchange of Thailand Index," Master Thesis in Economics, Chiang Mai University, Chiang Mai, Thailand, 1996.

APPENDIX

Appendix 1 Python Scripting to fetch Donald J. Trump's Twitter Messages

```
pip install python-twitter
pip install pandas

ıport twitter
ıport pandas as pd
›m pandas.io.json import json_normalize

### Entry
__name__ == '__main__':
 # Read Configuration
 config = {
    "twitter": {
        "access_token": "963642876335968257-bDjYdKDgAUGzi3dLH0gnkBY5t7Y2jpU",
        "access_token_secret": "EguCJE6hyZtniaxWmlzHgfgCGdz4hDpE9lCOHvfsEAtBz",
        "consumer_key": "HIwe8dRe6pObAxqdA5HVaXQnq",
        "consumer_secret": "KmNF8akJIAQj3g8faATns7yBkkw97eAA5qtAOnc6yx8hQbzU9o"
    }
 }

 # TWEETS
 output_file = f'tw-trump.csv'

 # Twitter
 access_token = config['twitter']['access_token']
 access_token_secret = config['twitter']['access_token_secret']
 consumer_key = config['twitter']['consumer_key']
 consumer_secret = config['twitter']['consumer_secret']
 print('Load Twitter config: complete')

 # Authentication
 api = twitter.Api(consumer_key=consumer_key, consumer_secret=consumer_secret,
                   access_token_key=access_token, access_token_secret=access_token_secret,
                   tweet_mode='extended')
 print(api.VerifyCredentials())

 # Begin process
 tweets = api.GetUserTimeline(screen_name='realDonaldTrump', count=100)
 tweets_dict = list(map(lambda x: x.AsDict(), tweets))
 tweets_df = json_normalize(tweets_dict)
 tweets_df['created_at'] = pd.to_datetime(tweets_df['created_at'],format='%a %b %d %H:%M:%S +0000 %Y') + pd.to_timedelta(7, unit='h')
 tweets_df.to_csv(output_file, index=False, encoding='utf8')
```

Appendix 2 Twitter Message Data using Python Script

created_at	favorite_count	full_text
11/12/2019 10:07	17947	https://t.co/ruQBK6gNLL
11/12/2019 08:58	43595	After years of rebuilding OTHER NATIONS, we are finally rebuilding OUR NATION. In everything we do, we are putting AMERICA FIRST! #KAG2020 https://t.co/sS0Y01MJYd
11/12/2019 08:55	41365	Day after day, we are exposing the depravity, dishonesty and sickness of the corrupt Washington establishment โ€” and with your help, we are going to complete the mission and DRAIN THE SWAMP! #KAG2020 https://t.co/SM5hocqoNi
11/12/2019 08:47	62028	THANK YOU PENNSYLVANIA! With your help, your devotion, and your drive, we are going to keep on working, we are going to keep on fighting, and we are going to keep ON WINNING! We are ONE movement, ONE people, ONE family, and ONE GLORIOUS NATION UNDER GOD! https://t.co/g64HD9yL9N
11/12/2019 06:45		RT @SecretarySonny: Very encouraged by todayโ€□s breakthrough on #USMCA โ€“ the agreement is a big win for America, especially for our farmersโ€ฆ
11/12/2019 06:24		RT @MikeKellyPA: Promises made, promises kept! USMCA is a big win & will further boost America's economy. Thank you to @POTUS @realDonaldTโ€ฆ
11/12/2019 06:22		RT @WaysandMeansGOP: Ways and Means Republicans, @POTUS, and @USTradeRep Amb. Lighthizer fought hard and delivered on their promise for anโ€ฆ
11/12/2019 06:22		RT @RepArrington: After a year of needless delay by @SpeakerPelosi & Democrat leadership, we are finally ready to deliver a win for Americaโ€ฆ
11/12/2019 06:19		RT @ChuckGrassley: Renegotiating NAFTA was a central campaign promise of Pres Trump and 2day he delivered a historic win for the American pโ€ฆ
11/12/2019 06:16		RT @USTradeRep: Statement from United States Trade Representative Robert Lighthizer https://t.co/1kqSHU3hDq https://t.co/MYZ5C5NQg1
11/12/2019 06:15		RT @RepLaHood: The announced agreement on #USMCA is great news! I applaud @realDonaldTrump & @USTradeRep for negotiating a strong agreementโ€ฆ
11/12/2019 06:14		RT @RepFredKeller: I congratulate President @realDonaldTrump and @HouseGOP leadership in reaching a deal on the #USMCA. The new trade dealโ€ฆ
11/12/2019 06:14		RT @RepMeuser: (1/2) The #USMCA now looks like it will finally come to the House floor for a vote next week. This agreement, as negotiatedโ€ฆ
11/12/2019 06:13		RT @SenatorFischer: Pleased to hear that @realDonaldTrumpโ€□s administration & House Democrats have reached a deal on #USMCA! This is a majorโ€ฆ
11/12/2019 06:12		RT @RoyBlunt: ๐□□บ๐□□ธ๐□□จ @POTUS and House Democrats have announced an agreement to move #USMCA forward. My statement here โคต๏ธ□ https://t.co/Age7TDWfโ€ฆ
11/12/2019 06:08		RT @RepAndyBarr: More than a year after President Trump negotiated the North America trade deal, itโ€□s good to finally see the unnecessary pโ€ฆ
11/12/2019 06:07		RT @PatrickMcHenry: American workers have waited long enough, the time to pass the #USMCA is now. From enabling our economy to continue toโ€ฆ
11/12/2019 06:05		RT @WaysandMeansGOP: Itโ€□s time for the U.S. Congress to pass USMCA as soon as possible, without further delay, to unlock the benefits of thโ€ฆ
11/12/2019 06:03		RT @RepSmucker: The USMCA is a big win for Pennsylvania and the people of my district! Thanks to @realDonaldTrump and Republican policies,โ€ฆ
11/12/2019 05:57		RT @RepLeeZeldin: Huge win for President Trump getting USMCA over the finish line, but most importantly, it's a huge win for the American wโ€ฆ