# The Relationship of Corporate News and Stock Price Forecasting

Sukanchalika Boonmatham
*Department of Information Technology*
*Faculty of Information Technology* and Digital Innovation
*King Mongkut's University of Technology North Bangkok*
Bangkok, Thailand
sukanchalika.b@email.kmutnb.ac.th

Phayung Meesad
*Department of Information Technology Management*
*Faculty of Information Technology* and Digital Innovation
*King Mongkut's University of Technology North Bangkok*
Bangkok, Thailand
pym@kmutnb.ac.th

***Abstract*—Finding a relationship between corporate news and the stock price is a challenging task. In this research, we try to build machine learning models that capture the relationship of news and stock prices of several companies. In this work, eight companies were selected randomly from Industry Group Index and Sectoral Index. Corporate news articles from the eight selected companies were collected along with their stock prices. Two of traditional machine learning models and two deep learning models were used in this study for comparison purpose. The models were based on Support Vector Machine (SVM), Multilayer Perceptron (MLP), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). Using news articles as inputs, the models were trained to classify stock prices into two classes: Up and Down of the stock closing price. For classification performance, Accuracy, Precision, and Recall were used. The results showed that GRU had highest precision and recall values with 0.67 and 0.62, respectively. GRU model also had an average accuracy score higher than other models with 0.65.**

***Keywords—Text Mining, Deep Learning, Classification Model***

## I. Introduction

News presentation to publicize facts or events that occurred in the corporate to both internal and external persons to be informed. To build business confidence, especially corporate news related to stock price indexes, which are tools to identify price levels and stock market trends. Reflecting the impact on the corporate so that the stakeholders can be informed of the trend or direction that occurred in the corporate.

In this research, we apply two traditional machine learning classification: SVM [1] and MLP [2], as well as two deep learning classification: LSTM [3] and GRU [4] for forecasting relationship of corporate news to stock price. The data sets used in the experiment are eight corporation news messages that were selected by random from Industry Group Index and Sectoral Index (Agro & Food Industry, Financials, Property & Construction, Services, Consumer Productions, Industrials, Resources, and Technology) together with the stock closing price index data. To forecasting classify the stock closing price of the next day. The predictive performance of classification modeling is compared.

## II. Literature Review

### A. Corporation News

A corporation news release is the reporting of facts or events that occur in a corporate. Corporation news affects the image, feeling, good attitude to the organization. Public relations of corporate news communicate both inside and outside the corporate, reflecting the stakeholders of the operations of corporate. Currently, corporate news tends to be offered through the website or social media. This makes it quick and easy to search even more.

### B. Traditional Machine Learning for Classification

There are many machine learning techniques can be applied for classification. To name a few are Support Vector Machine (SVM) and Multilayer Perceptron Neural Network (MLP).

SVM [1] is an algorithm that solves data classification problems. It is used for data analysis and data classification. By using the coefficient determination of the equation to create a line separating the data that is input into the process for the system to learn by focusing on the dividing line to best identify the data group. This algorithm can efficiently classify data with many dimensions and classify ambiguous data.

MLP [2] is a model of an artificial neural network (ANN) that simulates the function of the human brain for classification. MLP's working model begins with the model receiving input and then the model calculates the data and stores it in the hidden layer.

### C. Deep Learning for Classification

Deep learning is a kind of Artificial Neuron Networks (ANN), based on concepts and techniques of the human brain that allows computers to recognize, think, and remember as a human neural network. Deep learning is a popular algorithm created for machine learning [5] that can be processed in parallel. Deep learning can be used in a diversity of applications such as health informatics, translational bioinformatics, medical imaging, pervasive sensing, medical informatics, and public health [6].

The complexity of deep learning techniques varies. Each deep learning has its own learning process. The learning process of machine learning models is a mechanism for adjusting parameters. Usually, it runs multiple epochs to reduce predictive error values.

Some other popular Deep Learning techniques include Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). RNNs consists of two important parts: the input data at that node and the hidden state, which stores the results from previous node calculations. The two data are combined and the result is the input to the next node, which often encounters problems with data that is too long [7].

Long Short-Term Memory and Gated Recurrent Unit (GRU) are two popular architectures of RNN. LSTM is a solution to RNN's long sequential data. LSTM is popularly applied to time series data [3]. There is a work procedure that has a cell state that stores the state of each node so that it can be reversed from the previous state. In addition, there is a gate to control the flow of input data and a gate to control output data. LSTM consists of 3 main functions as shown in Fig. 1.

- Forget Gate has a new input data control function that works together with the previous hidden state to decide whether to keep this data or not.
- Input Gate has the function of controlling the new input data works together with previous hidden states in deciding whether to update data or not. The state update uses the input modulation gate to control the update and write to each node.
- Output gate has the function of controlling the data provided by the input gate to decide whether to store this data or forward it to the next node.

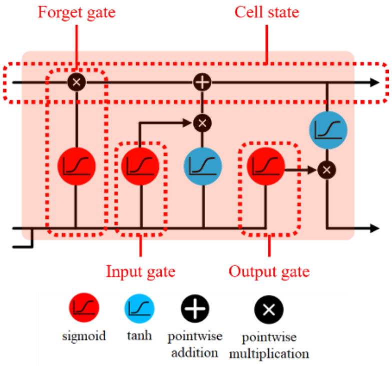

Fig. 1. Long Short-Term Memory (LSTM)

The Gated Recurrent Unit (GRU) [4] has an algorithm similar to LSTM, in which each node uses forget gate and input gate into update gate and combines cell state and hidden state as shown in Fig. 2. The results are as effective as the LSTM algorithm; however, the number of GRU parameters used are less to create sequential data sets and the architecture is easy to use. GRU forgets the characteristics of unimportant data and will not be lost when the long-term is propagated [8].

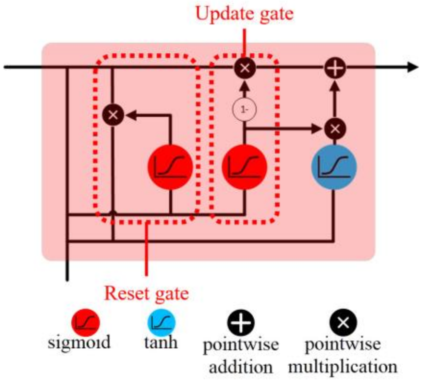

Fig. 2. The Gated Recurrent Unit (GRU)

### D. Data Classification

Data classification is the creation of a model or data classifier in order to predict the class of data. There are many machine learning models based on traditional machine learning models such as SVM and MLP as well as deep learning approaches such as DNN, CNN, LSTM, and GRU.

### E. Evaluation of classification

Evaluation of model performance obtained by forecasting and classification data [9] as follow:

- Accuracy is related to the ability of the data separator to be able to correctly identify the unseen data. The accuracy can be evaluated by using one or more sets of data that are separate from the training dataset, as in:

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total} \quad (1)$$

- Precision is the fraction of relevant instances among the retrieved instances (also called positive predictive value), as in:

$$Precision = \frac{True\ Positive}{Actual\ Results} \quad (2)$$

- Recall is the fraction of the total amount of relevant instances that were retrieved (also known as sensitivity), as in:

$$Recall = \frac{True\ Positive}{Predicted\ Results} \quad (3)$$

## III. Research Methodology

### A. Overview

This research objective is to find the relationship between corporate news and stock price forecasting. The data used in this experiment is the corporate news messages and stock price of the Industry Group Index of Stock Exchange of Thailand (SET). The machine learning models used are LSTM and GRU from deep learning methods and SVM and MLP from traditional methods. The models are built for forecasting the test data to classify two classes of stock prices: Up and Down. The model evaluation based on Accuracy, Precision, and Recall measures to indicate the performance of the trained models.

### B. Data Preparation

The data used in this experiments were the corporate news messages. The corporate was selected by a random method from each the Industry Group Index and Sectoral Index that is a tool indicates the price level and trend of the stock market i.e. CPF, SCB, LPN, BEC, TSR, IVL, PTTEP, and TRUE. Then, web crawling technique is used to retrieve news related to each organization from http://www.kaohoon.com [10]. Several corporate news retrieved are ranked as follows: CPF: 5331, SCB: 6101, LPN: 2329, BEC: 2932, TSR: 252, IVL: 5636, PTTEP: 6393, and TRUE: 5551.

After that, the corporate news messages were selected only in Thai and the date matched the closing price of the index in each corporate. The total number of news was respectively as follows: CPF: 1372, SCB: 1880, LPN: 336, BEC: 471, TSR: 188, IVL: 1538, PTTEP: 2135, and TRUE: 1363. Moreover, the news messages were fed to the cleaning process for removing some unwanted texts such as HTML tags, special symbols, spaces, quotation marks, or URLs.

Next, the texts were put into word tokenization process for creating a corpus. Thai word tokenization was based on newmm engine in the module word_tokenize from PythaiNLP package. The total number of word tokens that were tokenized were as follows: CPF:406628, SCB: 673310, LPN: 111456, BEC: 188897, TSR: 61972, IVL: 518538, PTTEP: 741613, and TRUE: 462480. After tokenization process, the words were converted into word indexes, which assign an index to each word arranged according to the word frequency.

Next, the datasets were prepared to the same sequence of lengths by setting the maximum length of the datasets and pre-padding with the number 0 for every data set. This dataset then was used to create models and evaluate further, as shown in Fig. 3.

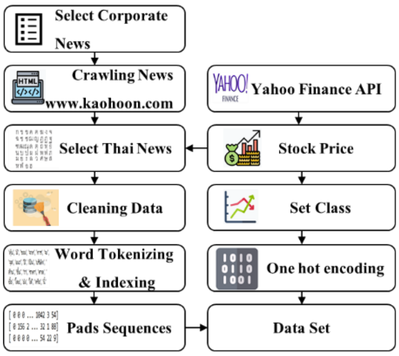

Fig. 3. Data preparation process.

For the stock closing price data, which is loaded from the Yahoo Finance API and there are classify the class into 2 classes, which are the stock price up and the stock price down. Based on the closing price today, it will affect the news that happens the next day. In the case that the previous day has a closing price of the stock less than the closing price of the day, the next day will set a class of "up". On the other hand, in the case that the previous day has a closing price of the stock greater than the closing price of the day, the next day will set a class of "down". After that, the stock closing price data is converted to binary format using one-hot encoding method.

### C. Training and Testing Datasets

The combination of the datasets obtained from the previous step, using the relationship of the stock closing date and the news date message occurs at the same time. After that, the data are split into 80 percent of all datasets as a training set and the 20 percent as a testing set to prepare for the next step. The training set and testing set have the following datasets: CPF: 774/194, SCB: 1,211/303, LPN: 208/53, BEC: 316/80, TSR: 121/31, IVL: 1,038/260, PTTEP: 1,480/371, and TRUE: 861/216.

### D. Forecasting and Classification model

From the experiment, two deep learning techniques: LSTM and GRU, and two traditional techniques: MLP and SVM were used to create models for forecasting data. For the forecasting model using deep learning techniques in the experiment, the parameters of LSTM and GRU algorithm are specified as shown in Fig.4.

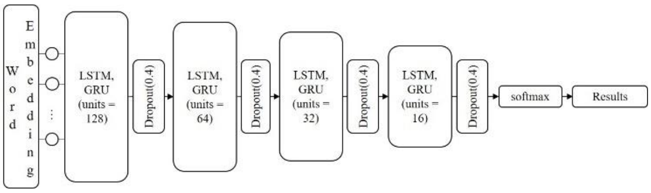

Fig. 4. Sequential Deep Learning Model (LSTM, GRU).

Forecasting models using deep learning techniques in which each cell consists of input gate, output gate, and forget gate. That framework as shown in Fig. 4. consist of four LSTM and GRU layers and set the parameters in each layer in the following order (dimension units = 128, 64, 32, and 16 with dropout=0.4) In addition, dropout technique is used to avoid overfitting [12] as it will drop connections between each node and not allow to send data to the next output. The sigmoid function is used as the activation function is set to be Softmax, which is a function that takes a summary of all processing from input to output by the probability that no more than 1.

### E. Performance evaluation

Evaluation of model performance obtained by forecasting data and classification data base on Accuracy, Precision, and Recall scores.

### F. Analyze and Summarize

For the final step of the experiment, the results of the performance tests of the model are applied to analyze the relationship between news corporate to stock forecasting.

## IV. Experimental Results

The experiments on the text corporate news by 8 random from each industry group include CPF, SCB, LPN, BEC, TSR, IVL PTTEP, and TRUE. That the datasets imported to the classification and prediction model are divided into training and testing datasets with all 4 models, LSTM, GRU, SVM, and MLP. The performance evaluation for the forecasting models measured by Precision, Recall, and Accuracy. The results show that the classification and prediction of the closing price class of the stock gave similar precision and recall values. The highest precision and recall values for LSTM, GRU, SVM, and MLP model which are BEC (0.67, 0.62), TSR (0.67, 0.62), TSR (0.58, 0.58) and BEC (0.54, 0.53), respectively. Based on the LSTM model provides the top 3 accuracy values that BEC, TSR, and PTTEP, which BEC has the best accuracy (0.63). The GRU model provides the top 3 accuracy values that TSR, BEC, and PTTEP, which TSR has the best accuracy (0.65). The SVM model provides the top 3 accurate values that TSR, BEC, and CPF, which TSR and BEC are the best accuracies (0.58). The MLP model provides the top 3 accurate values that BEC, LPN, and PTTEP, which BEC and LPN have the best accuracy (0.53) as shown in TABLE I.

For the experimental results, by comparing the classification and forecasting models in all 4 models, it was found that the GRU model gives an average accuracy more than other models. TSR stock which gives the accuracy is 0.65 the most. The accuracy-based 4 models are represented by a radar graph as Fig. 5.

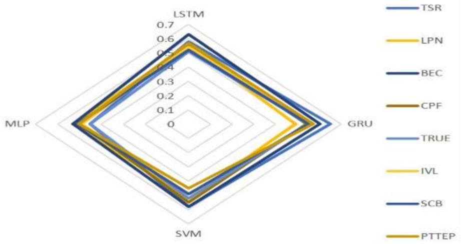

Fig. 5. The accuracy-based 4 models with LSTM, GRU, SVM, and MLP

Considering the experimental results from the detailed news text, it was found that there were a lot of corporate news articles crawling from the websites. The texts were taken into word tokenization process for creating a Thai corpus using libraries with PythaiNLP. Some of the words tokenized are incorrect. For example, the word “ธนาคารทหารไทย : thanakhan thahan thai” (Thai Military Bank) was tokenized to “ธนาคาร | ทหาร | ไทย : thanakhan | thahan | thai” (Thai Military Bank) or abbreviations as the word “ป.ต.ท. : por tor thor” (Petroleum Authority of Thailand) was tokenized to “ป. | ต. | ท. : por | tor | thor” (Petroleum Authority of Thailand). These words are the name of the same corporate should be the same. Therefore, words that have the same meaning or the same company should be added to the Thai corpus to increase the efficiency of the learning process. Another problem found was that crawling corporate news text from websites found many misspellings, such as “วิเคราะห์ : wi khraw” (analyze), spelling as “วิเครา : wi khrao” (no meaning in Thai) or “บริษัท : baaw ri sat” (company) spelling as “ริษัท : ri sat” (no meaning in Thai) etc. This results in incorrect words that causes the accuracy as low as 40-60%. For these problems, the researcher will seek for further solutions.

## V. Conclusion

In this research we focused on finding a relationship between corporate news and the stock price. Text mining and machine learning models were implemented to learn the relationship of news and stock prices. Corporate news from eight companies randomly selected from Industry Group Index and Sectoral Index were studied. Corporate news articles were collected along with their stock prices. Two of traditional machine learning models and two deep learning models were used: SVM, MLP, LSTM, and GRU. The models were trained to classify stock prices into two classes: Up and Down of the stock closing price. Classification performance were based on Accuracy, Precision, and Recall. The results showed that GRU had highest precision and recall values with 0.67 and 0.62, respectively. GRU model also had an average accuracy score higher than other models with 0.65.

## References

[1] C. D. A. Vanitha, D. Devaraj, and M. Venkatesulu, “Gene Expression Data Classification Using Support Vector Machine and Mutual Information-based Gene Selection,” *Procedia Comput. Sci.*, vol. 47, pp. 13–21, Jan. 2015, doi: 10.1016/j.procs.2015.03.178.

[2] C. Zhang *et al.*, “A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification,” *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 133–144, Jun. 2018, doi: 10.1016/j.isprsjprs.2017.07.014.

[3] S. Boonmatham and P. Meesad, “Time Series Analysis of Stock Prices based on Deep Learning,” in *Proceedings of the 6th Joint Symposium on Computational Intelligence (JSCI6)*, Dec. 2018, p. 2, [Online]. Available: http://sites.ieee.org/thailand-cis/event/jsci6/.

[4] M. Phi, “Illustrated Guide to LSTM’s and GRU’s: A step by step explanation,” *Medium*, May 01, 2020. https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21 (accessed May 07, 2020).

[5] L. Ren, J. Cui, Y. Sun, and X. Cheng, “Multi-bearing remaining useful life collaborative prediction: A deep learning approach,” *J. Manuf. Syst.*, vol. 43, pp. 248–256, Apr. 2017, doi: 10.1016/j.jmsy.2017.02.013.

[6] D. Ravì *et al.*, “Deep Learning for Health Informatics,” *IEEE J. Biomed. Health Inform.*, vol. 21, no. 1, pp. 4–21, Jan. 2017, doi: 10.1109/JBHI.2016.2636665.

[7] D. Wei *et al.*, “Research on Unstructured Text Data Mining and Fault Classification Based on RNN-LSTM with Malfunction Inspection Report,” *Energies*, vol. 10, no. 3, p. 406, Mar. 2017, doi: 10.3390/en10030406.

[8] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling,” *ArXiv14123555 Cs*, Dec. 2014, Accessed: Nov. 14, 2018. [Online]. Available: http://arxiv.org/abs/1412.3555.

[9] S. Saxena, “Precision vs Recall,” *Medium*, May 13, 2018. https://towardsdatascience.com/precision-vs-recall-386cf9f89488 (accessed May 06, 2020).

[10] “Online stock business news: Independent speaker of the capital market,” *Online stock business news*. https://www.kaohoon.com/ (accessed Jan. 16, 2020).

[11] M. Iliadis, L. Spinoulas, and A. K. Katsaggelos, “Deep fully-connected networks for video compressive sensing,” *Digit. Signal Process.*, vol. 72, pp. 9–18, Jan. 2018, doi: 10.1016/j.dsp.2017.09.010.

[12] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araujo, and J. Santos, “Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches [Research Frontier],” *IEEE Comput. Intell. Mag.*, vol. 13, no. 4, pp. 59–76, Nov. 2018, doi: 10.1109/MCI.2018.2866730.

TABLE I. Performance Evaluation Results

| **Model** | | **LSTM** | | | **GRU** | | | **SVM** | | | **MLP** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **stock** | **class** | **precision** | **recall** | **accuracy** | **precision** | **recall** | **accuracy** | **precision** | **recall** | **accuracy** | **precision** | **recall** | **accuracy** |
| **TSR** | UP | 0.58 | 0.25 | 0.58 | 0.68 | 0.64 | 0.65 | 0.59 | 0.62 | 0.58 | 0.47 | 0.44 | 0.45 |
| | Down | 0.69 | 0.93 | | 0.66 | 0.59 | | 0.57 | 0.53 | | 0.44 | 0.47 | |
| | Avg. | 0.64 | 0.58 | | 0.67 | 0.62 | | 0.58 | 0.58 | | 0.45 | 0.45 | |
| **LPN** | UP | 0.64 | 0.47 | 0.55 | 0.57 | 0.43 | 0.49 | 0.55 | 0.57 | 0.52 | 0.50 | 0.53 | 0.53 |
| | Down | 0.48 | 0.65 | | 0.43 | 0.57 | | 0.48 | 0.46 | | 0.54 | 0.52 | |
| | Avg. | 0.57 | 0.55 | | 0.51 | 0.49 | | 0.52 | 0.52 | | 0.52 | 0.53 | |
| **BEC** | UP | 0.80 | 0.74 | 0.63 | 0.58 | 0.58 | 0.60 | 0.57 | 0.45 | 0.58 | 0.59 | 0.37 | 0.53 |
| | Down | 0.54 | 0.52 | | 0.62 | 0.62 | | 0.58 | 0.69 | | 0.46 | 0.67 | |
| | Avg. | 0.67 | 0.62 | | 0.60 | 0.60 | | 0.57 | 0.57 | | 0.54 | 0.53 | |
| **CPF** | UP | 0.47 | 0.51 | 0.55 | 0.45 | 0.49 | 0.54 | 0.47 | 0.51 | 0.55 | 0.40 | 0.44 | 0.49 |
| | Down | 0.62 | 0.58 | | 0.60 | 0.57 | | 0.62 | 0.57 | | 0.56 | 0.53 | |
| | Avg. | 0.56 | 0.55 | | 0.54 | 0.54 | | 0.55 | 0.55 | | 0.50 | 0.49 | |
| **TRUE** | UP | 0.55 | 0.46 | 0.51 | 0.57 | 0.61 | 0.55 | 0.56 | 0.40 | 0.52 | 0.48 | 0.44 | 0.45 |
| | Down | 0.49 | 0.57 | | 0.52 | 0.48 | | 0.49 | 0.65 | | 0.43 | 0.47 | |
| | Avg. | 0.52 | 0.51 | | 0.54 | 0.55 | | 0.53 | 0.52 | | 0.46 | 0.45 | |
| **IVL** | UP | 0.50 | 0.47 | 0.54 | 0.51 | 0.44 | 0.55 | 0.45 | 0.44 | 0.49 | 0.45 | 0.43 | 0.49 |
| | Down | 0.56 | 0.60 | | 0.57 | 0.64 | | 0.52 | 0.53 | | 0.52 | 0.54 | |
| | Avg. | 0.54 | 0.54 | | 0.54 | 0.55 | | 0.49 | 0.49 | | 0.49 | 0.49 | |
| **SCB** | UP | 0.52 | 0.55 | 0.52 | 0.54 | 0.66 | 0.55 | 0.49 | 0.46 | 0.49 | 0.51 | 0.50 | 0.51 |
| | Down | 0.53 | 0.50 | | 0.57 | 0.45 | | 0.49 | 0.51 | | 0.51 | 0.52 | |
| | Avg. | 0.52 | 0.52 | | 0.55 | 0.55 | | 0.49 | 0.49 | | 0.51 | 0.51 | |
| **PTTEP** | UP | 0.45 | 0.43 | 0.57 | 0.43 | 0.38 | 0.57 | 0.35 | 0.49 | 0.45 | 0.37 | 0.44 | 0.50 |
| | Down | 0.65 | 0.67 | | 0.64 | 0.68 | | 0.57 | 0.43 | | 0.60 | 0.53 | |
| | Avg. | 0.57 | 0.57 | | 0.56 | 0.57 | | 0.49 | 0.45 | | 0.51 | 0.50 | |