International Journal of Electrical and Computer Engineering (IJECE)
Vol. 11, No. 5, October 2021, pp. 4313~4324
ISSN: 2088-8708, DOI: 10.11591/ijece.v11i5.pp4313-4324

Predicting stock price movement using effective Thai financial probabilistic lexicon

Surinthip Sakphoowadon[1], Nawaporn Wisitpongphan[2], Choochart Haruechaiyasak[3]
[1,2]Faculty of Information Technology and Digital Innovation, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand
[3]National Electronics and Computer Technology Center, Pathum Thani, Thailand

Article Info

Article history:

Received Sep 11, 2020
Revised Apr 4, 2021
Accepted Apr 29, 2021

Keywords:

Event term
Probabilistic lexicon
Stock price prediction
Thai financial news articles
ThaiFinLex

ABSTRACT

Predicting stock price fluctuation during critical events remains a big challenge for many researchers because the stock market is extremely vulnerable and sensitive during such time. Most existing works rely on various numerical data of related factors which can impact the stock price direction. However, very few research papers analyzed the effect of information appearing in financial news articles. In this paper, a novel probabilistic lexicon based stock market prediction (PLSP) algorithm is proposed to predict the direction of stock price movement. Our approach used the proposed Thai financial probabilistic lexicon (ThaiFinLex) derived from Thai financial news and stock market historical prices. The PLSP development consists of three steps. Firstly, we constructed ThaiFinLex by extracting event terms from news articles and calculating their associated probability of increasing/decreasing values of stock prices. Then, event terms with bad prediction performance were filtered out. Finally, the stock price directions were predicted using the PLSP and the remaining effective event terms. Our results indicated that the proposed model can be used for predicting stock price movement. The performance is as high as 83.33% when PLSP is used to predict stocks from the financial sector.

Corresponding Author:

Surinthip Sakphoowadon
Faculty of Information Technology and Digital Innovation
King Mongkut's University of Technology North Bangkok
Bangkok 10800, Thailand
Email: surinthip_sa@hotmail.com

1. INTRODUCTION

For decades, data mining has been an important tool for investors to predict the stock prices' value and movement. According to the studies, stock prices can be affected by various factors: gold price, foreign exchange rate [1], crude oil price [2], other market indices, influential events and news articles that may directly or indirectly be related to the stock markets. Whilst most data sources used for predicting the stock price trend are quantitative, many research papers have also studied stock market trend prediction by analyzing financial news articles in order to improve prediction efficiency.

In general, the performance of stock price trend forecasting using quantitative data such as historical stock prices or stock market indexes is quite accurate [3]-[5]. However, these methods are not flexible enough for adapting to price fluctuation caused by critical events. This is because such approach, which relies solely on quantitative data, lacks human intuition, business knowledge, and does not take into account financial situation of companies on the stock market. To further explore the effect of news on the stock market price, several prediction models have factored in historical stock prices and financial news articles in

order to improve prediction performance [6], [7]. However, the existing techniques, to the best of our knowledge, still cannot achieve high prediction accuracy due to the way financial news articles were interpreted in the model. For instance, a prediction method using two major data sources: financial news articles and historical stock prices, proposed by Schumaker *et al.* [6], yielded only 58% accuracy.

According to our studies, several researchers tried to improve prediction performance by means of finding a suitable representation of news articles. For example, bag of words [6], [8], [9], noun phrases [6], [10], and named entities [6] were used as features to represent the information of the news articles. However, such features vaguely capture the impact of the companies' situation. Thus, many researchers have focused on other representations such as word couple [7], and event features [11]-[14]. The event feature was the latest feature representation of the event stated in the news articles. It is a more informative representation compared to other feature representations because it consisted of meaningful subject, verb, and object. On the other hand, Li *et al.* [15] proposed another technique, which exploits sentence-level summarization in stock price prediction. Nevertheless, the above news representation techniques did not yield satisfied prediction results because none of the results achieved higher than 75% accuracy.

Considering all of the aforementioned research, prediction accuracies obtained from most studies were still not at a satisfactory level. In addition, only a few studies focused on using the Lexicon concept. Therefore, this paper proposed a new predictive algorithm called probabilistic lexicon based stock market prediction (PLSP) along with the Thai financial probabilistic lexicon (ThaiFinLex). The goal of the PLSP is to outperform other existing models in predicting the directions of the stock prices. The ThaiFinLex, which stored event terms and probabilities of each event term, were derived from historical financial news articles learned from a one-year worth of data set. The proposed PLSP algorithm predicts directions of a certain stock price movement mentioned in related news articles by calculating the probability values of the relevant event terms collected in the ThaiFinLex. The results obtained from the prediction are stock price directions (up or down) at the day's end. The contribution of this work is to provide an efficient predictive model aiming to improve the accuracy of semantic-based stock price prediction. Moreover, the proposed PLSP algorithm can be applied to predicting gold price trends, foreign exchange rate trends, and crude oil price trends.

The scope of this study focused on two major data sources: the closing prices of the top 100 stocks in the stock exchange of Thailand (SET100) and relevant financial news articles. Data sets used in the experiment were divided into three groups. Each data set consisted of financial new articles and historical closing prices. The first data set consisting of trading information over a period of 1 year from March 2015 to February 2016 were used to generate a probabilistic lexicon. The second data set (March 2016 to February 2017) was used to analyze the efficiency of each event term and evaluate our proposed prediction model. The last data set (March 2017 to February 2018) was used for model evaluation. The results show that the proposed model has an overall accuracy of 75% and can achieve as high as 83.33% accuracy when predicting the stock price movement of companies in financial sector. The rest of this paper is organized as follows: Section 2 illustrates the proposed method for stock market price prediction. Section 3 describes the experimental results and discussion. Finally, the conclusion is presented in section 4.

2. RESEARCH METHOD

The proposed probabilistic lexicon based stock market prediction (PLSP) consists of four major systems that include data preparation system, Thai financial probabilistic lexicon (ThaiFinLex) generation system, effective event term analysis system, and model evaluation system, as shown in Figure 1.

2.1. Data preparation system

News articles and stock price historical data were collected and prepared before using them in the other steps. Data preprocessing system consists of many subsystems as follows:

2.1.1. Text preprocessing

This subsystem was designed to extract terms or words from selected Thai financial news articles, whereas the data were gathered over a period of three years (March 2015 to February 2018) from a targeted news website. For this study, we used www.kaohoon.com which is a website reporting news regarding companies listed in the stock markets. Because time was very important in our study, we labeled each news article with its corresponding actual affected date: date on which the news will affect the stock price. This can be done by considering the released time and released day of the news article and time during the opening of the stock market. Normally, the stock trading days are Monday to Friday. The trading market opens at 10:00 AM and closes at 4:30 PM. The affected dates of news articles, which were released after 4:30 PM of previous trading date $(d-1)$ until 4:30 PM of current trading date (d) were marked as being affected on date d. For example, all news articles which were released after 06/01/2016, 4:30 PM to 07/01/2016, 4:30

PM would be marked as being affected on 07/01/2016. However, the affected date of all news articles which were released between 08/01/2016, 4:30 PM to 11/01/2016, 4:30 PM were 11/01/2016 because the market closed during the weekend (09/01/2016 to 10/01/2016).

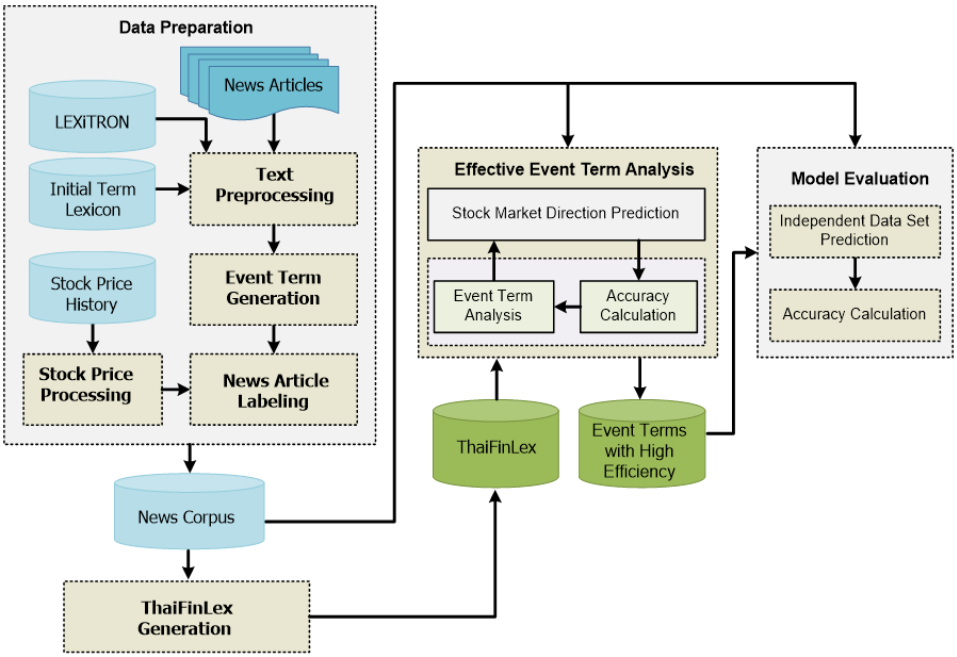

Figure 1. Stock market prediction using PLSP model

Afterward, terms or words within each article will be extracted using two lexicons for Thai language: the LEXiTRON [16] which is a widely used Thai dictionary, and the initial lexicon which was constructed manually by stock experts for this study. Each term stored in the initial lexicon is labeled as either "*clue*" or "*predicate*". A "clue" refers to a noun that represents the topic of an event mentioned in the articles, for instance, "company", "oil price", and "production cost". A "predicate" is a term that expresses more details about a clue, including a verb or an adjective such as "increase", "decrease", "invest", and "nice". In addition to classifying terms as clues or predicates, each term will be mapped to its corresponding synonym prior to storing in the initial lexicon. Table 1 presents examples of terms stored in the initial set of the lexicon.

Table 1. Examples of terms stored in the initial lexicon

Terms	Term Types	Synonyms
ต้นทุน (Cost)	Clue	ต้นทุน (Cost)
ต้นทุนการผลิต (Production Costs)	Clue	ต้นทุน (Cost)
กำไร (Profit)	Clue	กำไร (Profit)
กำไรสุทธิ (Net Profit)	Clue	กำไร (Profit)
เพิ่มขึน (Increased)	Predicate	เพิ่ม (Increased)
ขึน (rose)	Predicate	เพิ่ม (Increased)
แจ่ม (Nice)	Predicate	ดี (Good)
ดี (Good)	Predicate	ดี (Good)

Stop word removal is the last step of text preprocessing. The purpose of this process is to gather valuable words and eradicate terms that are less meaningful terms for predicting directions of stock price movements. Furthermore, Thai conjunction terms ("and", "or", "therefore", and "because") were included in the system because these words were used to detect boundaries in sentences or long compound words.

Unlike other studies, we also focused on the impact of the information that was embedded within each news article on the stock mentioned in the article. One news article can contain information about more than one stock whilst the information of one stock can be published in many news articles within one day. Hence, each article is also marked with associated stocks which were mentioned in the news.

2.1.2. Event term generation

The idea is similar to many existing works which focused on building a corpus using meaningful phrase extraction from document [7], [17]. After performing the stop word removal, the remaining terms were coupled to generate new meaningful terms. Each term is composed of a clue and a predicate. For this study, we refer to these coupled terms as an "event term". Note that a clue and its associated predicate may not necessarily be next to one another in a sentence. They can either be close to one another in the same sentence or be in a different sentence but in the same paragraph. Furthermore, if any event term appeared more than once in the same article, the generation function would consider its appearance once. Table 2 illustrates a few examples of event terms.

Table 2. Examples of event terms

Synonym of clue terms	Synonym of predicate terms	Event terms
ราคาน้ำมัน (Oil prices)	ขึ้น (Increased)	ราคาน้ำมันขึ้น (Oil prices increased)
ค่าใช้จ่าย (Expenditure)	ขึ้น (Increased)	ค่าใช้จ่ายขึ้น (Expenditure Increased)
ต้นทุน (Cost)	พุ่ง (jumped)	ต้นทุนพุ่ง (Cost jumped)

2.1.3. Stock price processing

Several studies used closing prices for stock market direction prediction [5], [18]-[20]. For this study, we used the rate of change (ROC) which was derived from the closing prices of each stock. Thus, the ROC of a stock i (s_i) traded on a date j (d_j) can be calculated as (1).

$$ROC_{s_i,d_j} = \frac{P_{s_i,d_j} - P_{s_i,d_{(j-1)}}}{P_{s_i,d_{(j-1)}}} * 100 \quad (1)$$

Where P_{s_i,d_j} denotes the closing price of a stock i on a trading date j and $P_{s_i,d_{(j-1)}}$ denotes the closing price of a stock i on the previous trading day.

2.1.4. News article labeling

Terms extracted from news articles related to a stock i and released on a trading date j were labeled as "up" or "down" depending on the ROC values. The changing status of the stock is "up" when ROC_{s_i,d_j} is greater than 0. On the other, if ROC_{s_i,d_j} is less than 0, the stock is labeled as "down". The data sets related to the stock i on the trading date j will be ignored from the experiments when ROC_{s_i,d_j} is equal to zero. The obtained result from the data pre-processing system is a news corpus. This news corpus will be used as input data in the remaining steps.

2.2. ThaiFinLex generation

There are three steps to generate ThaiFinLex: calculating the weight of each event term based on ROC, finding the total weight, and computing probability associated with each event term. One-year data set of Thai financial news articles collected from March 2015 to February 2016 were used for ThaiFinLex generation. In this study, we considered both the title and the content of the news articles.

2.2.1. Event term weighing

Let d_j be an affected date on which news articles containing a certain event term k (t_k) are released. In this work, an affected date j (d_j) has to be a trading date. Let $W^U_{t_k,s_i}$ and $W^D_{t_k,s_i}$ be the weights of an event term k relevant to stock i whose weight status are "up" and "down", respectively. The weight of an event term can be computed as (2), (3).

$$W^U_{t_k,s_i} = \sum_{j=1}^{n} (C^U_{t_k,s_i,d_j} * |(ROC)_{s_i,d_j}|) \text{ when } (ROC)_{s_i,d_j} > 0 \quad (2)$$

$$W^D_{t_k,s_i} = \sum_{j=1}^{n} (C^D_{t_k,s_i,d_j} * |(ROC)_{s_i,d_j}|) \text{ when } (ROC)_{s_i,d_j} < 0 \quad (3)$$

Where n denotes the total number of trading days from the data set. $C^U_{t_k,s_i,d_j}$ represents an indicator which will be equal to 1 if an event term k appears in any news articles related to a stock i on an affected date j when the status of the stock i on the date j is "up". $C^D_{t_k,s_i,d_j}$ denotes a value which will be equal to 1 if event

term k appears in any news articles related to a stock i on an affected date j when the status of the stock i on the date j is "down". Otherwise, $C^{U}_{t_k,s_i,d_j}$ and $C^{D}_{t_k,s_i,d_j}$ will be equal to 0.

2.2.2. The total weight calculation

The total weight is simply the summation of the weight calculated in the previous step across all stocks. Let $TW^{U}_{t_k}$ and $TW^{D}_{t_k}$ be the total weights of a certain event term k relevant to all stocks in SET100 whose total weight status are "up" and "down", respectively. The total weight of event term k can be computed as (4), (5).

$$TW^{U}_{t_k} = \sum_{i=1}^{r} W^{U}_{t_k,s_i} * \frac{CU}{CT} \quad (4)$$

$$TW^{D}_{t_k} = \sum_{i=1}^{r} W^{D}_{t_k,s_i} * \frac{CD}{CT} \quad (5)$$

Where r denotes the number of all stocks involved in this study. $W^{U}_{t_k,s_i}$ and $W^{D}_{t_k,s_i}$ represent the weights derived from (2) and (3), respectively. Let CU_{s_i} be the number of distinctive days that a certain event term k relevant to stock i appears in the articles, and the changing status of such stock i mentioned in the articles is "up". CU in (4) is the sum of CU_{s_i} of all stocks. Let CD_{s_i} be the number of distinctive days that a certain event term k relevant to a stock i appear in the news articles, and the changing status of a stock i mentioned in the articles is "down". CD in (5) derived from the summation of CD_{s_i} of all stocks. CT is equal to $CU + CD$.

2.2.3. Probability calculation

The probabilities associated with each event term are calculated and stored into our lexicon. Let $P^{U}_{t_k}$ and $P^{D}_{t_k}$ be the directional probabilities of each event term k which is relevant to stocks in SET100, when the directions are "up" and "down". The probabilities can be formulated as (6), (7).

$$P^{U}_{t_k} = \frac{TW^{U}_{t_k}}{TW^{U}_{t_k} + TW^{D}_{t_k}} \quad (6)$$

$$P^{D}_{t_k} = \frac{TW^{D}_{t_k}}{TW^{U}_{t_k} + TW^{D}_{t_k}} \quad (7)$$

Where $TW^{U}_{t_k}$ and $TW^{D}_{t_k}$ are the total weights derived from the (4) and (5), respectively. The directional probabilities of each event term in (6) and (7) can represent the directional probabilities of a certain event term relevant to the whole stock market in this study**.** The obtained result of the lexicon generation is a proposed probabilistic lexicon called ThaiFinLex, as shown in Figure 1. Table 3 shows event terms stored in the ThaiFinLex along with their directional probabilities**.**

Table 3. Examples of event terms in the ThaiFinLex

Event Terms	Directional probabilities	
	Up	Down
กำไรขึ้น (Profit increased)	0.71	0.29
กำไรลง (Profit decreased)	0.34	0.66
ต้นทุนขึ้น (Cost increased)	0.25	0.75
ค่าใช้จ่ายลง (Expenses decreased)	0.84	0.16

2.3. Effective event term analysis

Similar to the lexicon generation process, we considered both the titles and the contents of news articles during the same one-year period to analyze an effective event term. In particular, the data set (March 2016 to February 2017) consisting of 3091 distinctive news articles and the closing prices of 100 stock symbols during the same time period will be used as input data for our PLSP model. To find an effective event term, we have to analyze how accurate each event term is in PLSP**.**

2.3.1. Stock market direction prediction

In this step, the process focused on the prediction of each news article. The event terms were extracted from the related news articles, and then the corresponding event terms and their probability values

stored in ThaiFinLex were retrieved for calculation. Let a_v be a news article used to predict the price trend of the stock mentioned in the news. Let $SC^U_{a_v,s_i}$ and $SC^D_{a_v,s_i}$ be the total directional scores that stock i will be "up" or "down", given the occurrence of news article v (a_v) respectively. Then, the directional scores of each news article v related to a certain stock i consisting of m event terms can be computed as (8), (9).

$$SC^U_{a_v,s_i} = \prod_{k=1}^{m} P^U_{t_k} \quad (8)$$

$$SC^D_{a_v,s_i} = \prod_{k=1}^{m} P^D_{t_k} \quad (9)$$

Where m denotes the number of event terms that appeared in a news article v. $P^U_{t_k}$ and $P^D_{t_k}$ are the probabilities, stored in the ThaiFinLex, that the stock price will be up or down as a result of an event term k. Finally, let $P^U_{a_v,s_i}$ and $P^D_{a_v,s_i}$ be the probabilities that stock i will be "up" or "down" when considering the impact of all the event terms that appear in the news article v. By normalizing the directional scores in (8) and (9), $P^U_{a_v,s_i}$ and $P^D_{a_v,s_i}$ can be expressed as (10), (11).

$$P^U_{a_v,s_i} = \frac{SC^U_{a_v,s_i}}{SC^U_{a_v,s_i} + SC^D_{a_v,s_i}} \quad (10)$$

$$P^D_{a_v,s_i} = \frac{SC^D_{a_v,s_i}}{SC^U_{a_v,s_i} + SC^D_{a_v,s_i}} \quad (11)$$

PLSP predicts that the stock price will be up when $P^U_{a_v,s_i} > P^D_{a_v,s_i}$ and down when $P^U_{a_v,s_i} < P^D_{a_v,s_i}$.

2.3.2. Accuracy calculation

To calculate how accurate PLSP is, we simply use the confusion matrix (12).

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (12)$$

Where TP, true positive, denotes the number of times PLSP predicted that the stock price movement direction would be "up" and the actual direction was "up". True negative (TN) denotes the number of times PLSP predicts that the stock price direction would be "down" and the actual direction was "down". False positive (FP) denotes the number of times PLSP predicted that the movement direction would be "up" but the actual direction was "down". Finally, false negative (FN) denotes the number of times PLSP predicted that the stock price movement direction would be "down" but the actual direction was "up".

2.3.3. Event term efficiency calculation

A few events mentioned in the articles can significantly cause the stock market fluctuation. However, some event has less impact on stock price movements than the others. Therefore, the event terms used to predict the directions of stock price movements have different levels of impact on the movements. To improve the prediction performance, the efficiency of each event term k (EF_{t_k}) will be calculated using (13).

$$EF_{t_k} = \frac{CR_{t_k}}{CR_{t_k} + NCR_{t_k}} \quad (13)$$

Where CR_{t_k} is the total number of times that an event term k can correctly predict the trend of stock price movement when such event term was mentioned in news articles. Similarly, NCR_{t_k} is the total number of times that an event term k incorrectly predicts the stock price trends from input news articles. The prediction efficiency (EF) from (13) is a value between 0 and 1.

2.3.4. Event term analysis

As can be seen in Table 3, each event term *k* in ThaiFinLex has its associated directional probability $P^U_{t_k}$ and $P^D_{t_k}$ or probabilities that the stock price will be up and down as a result of an event term *k*. However, we noticed that many event terms in ThaiFinLex have less influence on the stock price movement. More specifically, their associated directional probabilities are between 0.4-0.6. Therefore, additional steps were taken to eliminate such event terms. The event terms with $P^U_{t_k}$ or $P^D_{t_k}$ values that are greater than or equal to 0.6 (>= 0.6) are included in the system. According to these criteria, 913 event terms which appeared in 1871 news articles were used as the input data set in the analysis.

The procedure for analyzing effective event terms is shown in Figure 2. In this step, EF level will be used to iteratively find event terms that yield high prediction accuracy. These effective event terms will be included in the PLSP prediction model. Five EF thresholds and six iterations were used in this study.

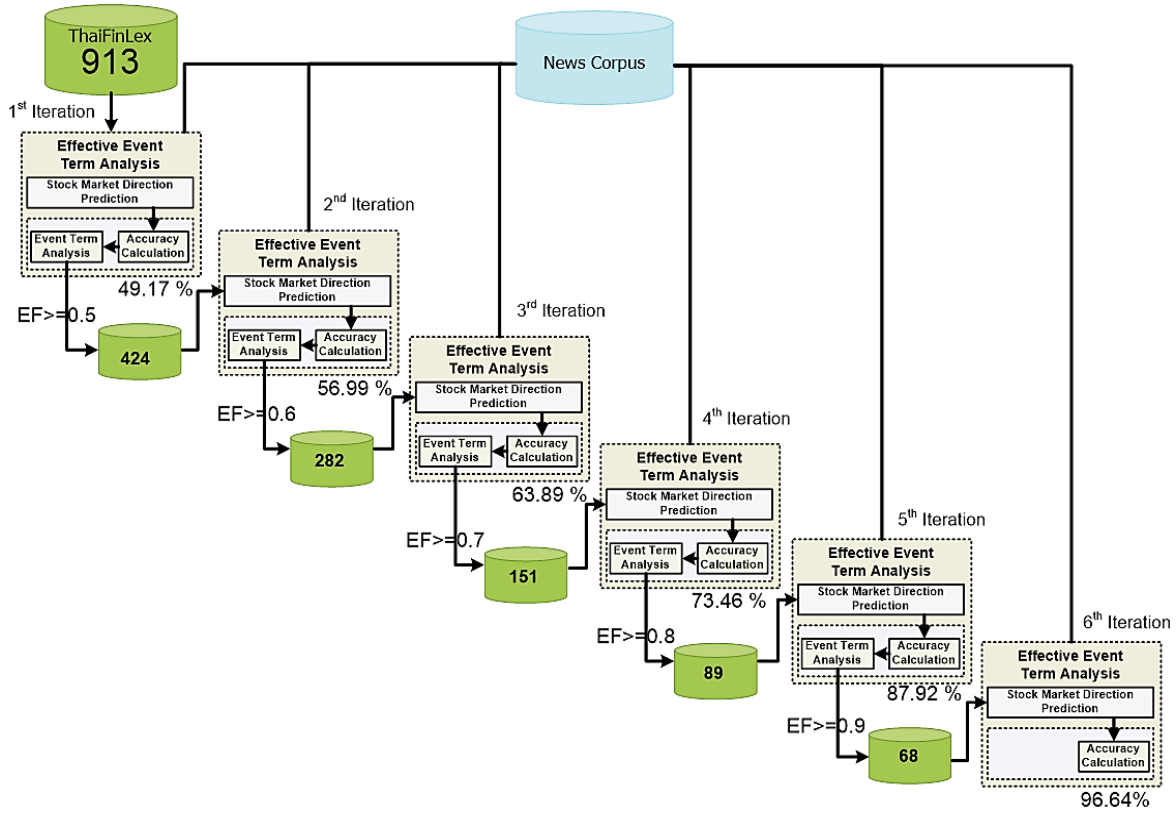

Figure 2. Five iterations of effective event term analysis

a. In the first iteration of the algorithm, the total number of selected event terms, 913 terms, from ThaiFinLex were further analyzed to obtain highly efficient event terms. A total of 1871 news articles, which contained those 913 terms, were used to predict the stock price direction using (10) and (11). The prediction accuracies were then calculated based on (12). The obtained result of this first iteration showed an accuracy of 49.17 %. EF values of all the 913 event terms were computed by using (13). After that, the event terms, whose EF values are greater than 0.5, will be chosen. As a result, 424 event terms met the criteria. For the next step, the chosen 424 event terms were then used as input event terms in the second iteration.
b. For the second iteration, 424 event terms derived from the first iteration and the 1579 related news articles were used as input data set. For the prediction result, the experiment yielded an accuracy of 56.99 % which was better than the first iteration. Similar to the first iteration, a new EF value of each event term was computed. Since the system calculates new EF values in every iteration, the EF values of the same event term changed due to the fact that ambiguous articles were excluded in every iteration. In the second iteration, the EF threshold used for selecting event terms is increased to 0.6. The total number of event terms which satisfied this condition was 282. This new set of the event terms will be used as input data set on the third iteration. The same process is repeated on every iteration. The numbers of the effective event terms obtained from the 3th, 4th, and 5th iterations were 151, 89, and 68, respectively.
c. In the last iteration (the 6th iteration), 68 event terms and 119 news articles which were obtained from the 5th iteration were used as the input data set. When considering this set of event terms, prediction accuracy was as high as 96.64 %. After the last iteration, we obtained a collection of the event terms with high prediction accuracy. This collection consisted of five data sets with different numbers of event terms, consisting of 424, 282, 151, 89, and 68 event terms. These five data sets were further used in the next steps.

Table 4 illustrates the overall results of the six iterations. The obtained results of event term analysis using different EF thresholds showed that the higher the threshold, the fewer the number of news articles and

the fewer number of event terms. Furthermore, the higher the EF threshold values, the higher the obtained accuracy. This is due to the fact that unrelevant event terms were filtered out during the selective process, so the remaining event terms used for the higher thresholds are usually ones that have higher impact on the stock price trend. Thus, better performance can be achieved when considering only the event terms with high efficiency. Table 5 shows examples of the event terms with high efficiency obtained from the final iteration.

Table 4. Prediction results of six iterations using the PLSP model

Iteration	EF Threshold	The number of news articles	The number of event terms	Accuracy (%)
1	-	1871	913	49.17
2	>=0.5	1579	424	56.99
3	>=0.6	997	282	63.89
4	>=0.7	471	151	73.46
5	>=0.8	207	89	87.92
6	>=0.9	119	68	96.64

Table 5. Examples of event terms with high efficiency

Event terms	Prediction efficiency (EF)
โครงการซื้อหุ้นคืน (Share buyback program)	0.9
ต้นทุนพุ่ง (Cost increased dramatically)	1.0
กำไรทะลัก (Very high profit)	1.0
ผลประกอบการเติบโต (Turnover increased)	1.0

3. RESULTS AND DISCUSSION

3.1. Comparison of the PLSP with the baseline models

After obtaining the highly efficient event terms as shown in Table 5, we compared the performance of PLSP against three different well-known models used for predicting stock prices: SVM [7], [8], [10], J48 [21], and BayesNet [22]. For this comparison, the input news articles data set (March 2016 to February 2017) used for evaluating the prediction performance was the same data set that was used for extracting the high efficiency event terms. Five sets of data with five different values of EF thresholds were then used as input for prediction.

The results, as shown in Table 6, clearly showed that the proposed PLSP with threshold of more than 0.5 performed better than the other three models. It is because the efficiency of the PLSP model depends on the efficiency of the event terms while the other three models ignored the use of event terms. Another reason is that the higher the EF threshold, the more relevant the event terms are used for prediction.

Table 6. Performance comparison of 4 models

EF Threshold	The total number of news articles	The total number of event terms	Accuracy (%)			
			BayesNet	J48	SVM (Linear kernel)	PLSP
>=0.5	1579	424	51.93	54.84	52.06	56.99
>=0.6	997	282	51.25	54.86	55.17	63.89
>=0.7	471	151	50.32	54.99	62.00	73.46
>=0.8	207	89	53.62	54.11	77.29	87.92
>=0.9	119	68	58.82	57.14	79.83	96.64

3.2. Model evaluation using an independent data set

In the previous section, we have evaluated the accuracy of the proposed PLSP model by considering event terms appearing in either title or content of the news articles. However, many studies suggested that the prediction model did not need to consider all of the contents in the news. In [11], the researchers showed that using only article titles for stock market prediction could achieve better results than using news contents. Radinsky *et al.* [23] confirmed that the article title information was useful for event prediction. Thus, many research papers predicted the stock markets by using the concise summary information of the news articles such as article titles [12], [13] or news-headlines [24]-[26]. To improve the prediction of the PLSP model, we further tested PLSP with article titles and first paragraphs of independent data set: 1653 Thai financial news articles collected over a period of 1 year (March 2017 to February 2018). Only event terms in the ThaiFinLex with high efficiency were used in this step. The proposed model (PLSP) was used as a prediction model and the confusion matrix was used to evaluate the performance. The obtained results of five different sets of event terms with different EF thresholds are shown in Table 7.

Table 7 shows the performance of PLSP using event terms with different efficiency levels. The best prediction result is obtained when PLSP is used with 0.7 threshold value. The model evaluation yielded the highest accuracy of 75%. This means that the data set with 151 event terms is the best data set for stock market prediction. However, the result of using EF threshold at 0.8 and 0.9 data set did not yield high accuracy compared with that of using EF threshold at 0.7. This is because the number of event terms with such high efficiency levels (EF=0.8, EF=0.9) is small, causing a smaller set of input articles. Hence, one mis-prediction amounted to a large percentage of inaccurate prediction. Furthermore, there were various unpredictable factors that affect the actual directions of the stock market apart from the event information in the news articles.

Table 7. Prediction performance of independent testing

EF Threshold	The total number of news articles related to all considering stocks	The total number of event terms	Accuracy (%)
>=0.5	712	424	57.44
>=0.6	441	282	60.77
>=0.7	148	151	75.00
>=0.8	59	89	71.12
>=0.9	42	68	69.05

Li *et al.* [27] studied stock movement prediction according to firm characteristics: trading volume, turnover, price-to-earnings (P/E) ratio, price-to-book (P/B) ratio, risk (β), and industry sector. They showed that stocks in some industries were more predictable than others. Hence, we further analyzed the prediction performance of PLSP with EF=0.7 (148 articles) when applied to different groups of industries. The stocks in this study were further categorized into seven industry groups. The prediction results are shown in Table 8.

Table 8. Prediction performance categorized by industries

Industries	The total number of news articles related to all considering stocks	Accuracy (%)
Agro & Food	9	66.67
Financials	18	83.33
Industrials	2	50.00
Property & Construction	34	73.53
Resources	32	75.00
Services	33	78.79
Technology	20	70.00

Table 8 illustrates the performance result using the PLSP with data set categorized by different industries. According to the results, the prediction accuracy of the financial industry was the highest among the 7 industry groups (83.33%). Interestingly, the prediction performance of the five groups: financials, property & construction, resources, services and technology exceeded 70% accuracy. The results show that stocks of these industries are highly affected by financial news articles.

The proposed model (PLSP) with efficient event terms can achieve 83.33% accuracy when predicting trends of stocks in the financial sector. We have shown that the efficiency of the PLSP is superior when compared with the other approaches [6], [7], [11], [12], [14], [15]. However, the result obtained by considering the whole market did not yield high accuracy. This is presumably due to the fact that the news article does not have much influence on certain industries. Thus, applying our prediction model to such industries can only result in poor performance.

According to our comprehensive study, predicting stock price direction using information from news articles has been shown to be less accurate than using quantitative data from the stock market: Stock return, price-earnings ratio (P/E), and trading volume. Likewise, the accuracy of PLSP is confined to a few limitations. First and foremost, it is apparent that ThaiFinLex is a static lexicon. More importantly, ThaiFinLex is generated from the 1-year worth of dataset (year 2016). Thus, what were used to be considered effective event terms may not be as effective in the next few years. On the other hand, there may be new terms in the future mentioned in news articles that may affect the direction of the stock price movements, but PLSP ignores such terms that do not exist in the ThaiFinLex. For example, blockchain and smart contract have just become buzzwords in Thailand in 2017. Therefore, ThaiFinLex does not include such words because it was developed based on a dataset in 2016.

The second limitation is the use of the Thai language in this experiment. Training the machine to understand the Thai language is a hard problem by itself. To list a few challenges, there is no space to separate each word from one another, therefore efficiency of the work is limited to the algorithm used for splitting sentences into word tokens. In addition, news articles often contain catchy or fancy phrases that are not typically used in everyday communication. Therefore, it is difficult to extract key relevant words. In this paper, we used the LEXiTRON and the initial term Lexicon to extract words and phrases. As a result, relevant event terms that were not included in the initial set of our dictionary corpus may be excluded from the experiment.

Another limitation is the reliability of rare event terms which do not appear in news articles that often. More specifically, the occurrence frequency of an event term found in the training dataset determines its reliability in predicting the stock price movement. Last but not least, it should be pointed out that the stock price direction can be affected by many other factors such as gold price, foreign exchange rate, oil price, and other social media data [28]-[30] which are not considered in this study. These limitations imply that the proposed model can be further improved in many ways, e.g., determining new event terms and updating ThaiFinLex with these terms automatically, considering PLSP prediction result as one of the many features that can be used to predict the directions of stock prices.

4. CONCLUSION

In this study, we proposed a novel algorithm to predict the directions of a certain stock price movement, called PLSP. We constructed the ThaiFinLex developed by using a statistical text mining concept. ThaiFinLex is a lexicon containing event terms and their corresponding trend prediction probabilities. The ThaiFinLex was developed by using terms appearing in Thai financial news articles and the historical stock prices. The proposed model along with the ThaiFinLex aims to improve the semantic-based stock prediction performance. To improve the prediction acurracy of the proposed model, high quality event terms were analyzed. Five prediction efficiency (EF) thresholds were used to analyze the event terms in ThaiFinLex to detect the effective event terms to be included in PLSP. By focusing on the data set of the event terms with high efficiency, we compared the predictive performance of the proposed PLSP model with the other popular models, including SVM, J48, and BayesNet. The obtained results showed that the PLSP model with threshold outperforms the other three models. To further evaluate PLSP model, we investigated the extracted event terms with high efficiency from the most relevant part of the articles: the title and the first paragraph of news articles. By using these event terms on a one-year worth of independent data set, we found that the accuracy is at most 75% when EF threshold is 0.7. However, when applying the same model to seven groups of industries, the prediction result of three industries: financials, resources, and services industry achieved greater than 75% accuracy. This result revealed that the highly efficient event terms in this study are suitable for predicting stock price trends of these three industries. In addition, the experimental results indicated that the proposed PLSP algorithm yielded a superior performance than the other previous works.

The PLSP algorithm can be applied to trend predictions of other assets: Gold price, foreign exchange rate, crude oil price, etc because these markets can be affected by sentiments in financial news articles, Twitter, and other social media data. Nevertheless, the limitation of the proposed PLSP is the use of the static lexicon (ThaiFinlex). In the future, we plan to integrate quantitative analysis models which are widely used models for stock market prediction, namely moving average convergence divergence (MACD) and relative strength index (RSI) in order to improve the prediction performance of PLSP. The size of ThaiFinLex should be increased by gathering data from various sources including financial news articles, and social media information. Moreover, the lexicon used for prediction should be updated automatically.

REFERENCES

[1] R. Rassetiadi, and Suharjito, "Foreign exchange prediction based on indices and commodities price using convolutional neural network," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 18, no. 1, pp. 494-501, 2020, doi: 10.11591/ijeecs.v18.i1.pp494-501.

[2] S. A. M. Almasani, V. I. Finaev, W. A. Abdo Qaid, A. V.Tychinsky, "The Decision-making Model for the Stock Market under Uncertainty," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7, no. 5, pp. 2782-2790, 2017, doi: 10.11591/ijece.v7i5.pp2782-2790.

[3] W. Dai, Jui-Yu Wuc, Chi-Jie Lud, "Combining nonlinear independent component analysis and neural network for the prediction of Asian stock market indexes," *Expert Systems with Applications*, vol. 39, no. 4, pp. 4444-4452, 2012, doi: 10.1016/j.eswa.2011.09.145.

[4] J. L. Ticknor, "A bayesian regularized artificial neural network for stock market forecasting," *Expert Systems with Applications*, vol. 40, no. 14, pp. 5501-5506, 2013, doi: 10.1016/j.eswa.2013.04.013.

[5] I. Svalina, V. Galzina, R. Lujić, G. Šimunović, "An adaptive network-based fuzzy inference system (ANFIS) for the forecasting: The case of close price indices," *Expert Systems with Applications*, vol. 40, no. 15, pp. 6055-6063, 2013, doi: 10.1016/j.eswa.2013.05.029.
[6] R. P. Schumaker and H. Chen, "Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFin Text System," *ACM Transactions on Information Systems*, vol. 27, no. 12, pp. 1-19, 2009, doi: 10.1145/1462198.1462204.
[7] M. Y. Kaya and M. E. Karsligil, "Stock price prediction using financial news articles," *2010 2nd IEEE International Conference on Information and Financial Engineering, Chongqing*, China, 2010, pp. 478-482, doi: 10.1109/ICIFE.2010.5609404
[8] X. Li, Chao Wang, Jiawei Dong, Feng Wang, Xiaotie Deng, Shanfeng Zhu, "Improving stock market prediction by integrating both market news and stock prices," *International Conference on Database and Expert Systems Applications-DEXA 2011*, vol. 6861, pp. 279-293, 2011.
[9] P. Samad, M. Sofianita, R. Shuzlina Abdul., "Analytics of stock market prices based on machine learning algorithms," *Indonesian Journal of Electrical Engineering and Computer Science, (IJEECS)*, vol. 16, no. 2, pp. 1050-1058, 2019, doi: 10.11591/ijeecs.v16.i2.pp1050-1058.
[10] M. Hagenau, M. Liebmann, D. Neumann, "Automated news reading: Stock price prediction based on financial news using context-capturing features," *Decision Support Systems*, vol. 55, no. 3, pp. 685-697, 2013, doi: 10.1016/j.dss.2013.02.006.
[11] X. Ding, Y. Zhang, T. Liu, J. Duan, "Using structured events to predict stock price movement: An empirical investigation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1415-1425, doi: 10.3115/v1/D14-1148.
[12] X. Ding, Y. Zhang, T. Liu, J. Duan, "Deep Learning for Event-Driven Stock Prediction," *Proceedings of the 24th International Joint Conference on Artificial Intelligenc*e, 2015, pp. 2327-2333,
[13] M. R. Vargas, B. S. L. P. de Lima and A. G. Evsukoff, "Deep learning for stock market prediction from financial news articles," *2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, Annecy, France, 2017, pp. 60-65, doi: 10.1109/CIVEMSA.2017.7995302.
[14] X. Zhang, S. Qu, J. Huang, B. Fang and P. Yu, "Stock market prediction via multi-source multiple instance learning," *IEEE Access*, vol. 6, pp. 50720-50728, 2018, doi: 10.1109/ACCESS.2018.2869735.
[15] X. Li, H. Xie, Y. Song, S. Zhu, Q. Li and F. L. Wang, "Does summarization help stock prediction? A news impact analysis," *IEEE Intelligent Systems*, vol. 30, no. 3, pp. 26-34, 2015, doi: 10.1109/MIS.2015.1.
[16] "LEXiTRON," 2016. [Online]. Available: http://lexitron.nectec.or.th.
[17] I. Veritawati, Ito Wasito, T. Basaruddi, "Text Preprocessing Using Annotated Suffix Tree with Matching Keyphrase," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 5, no. 3, pp. 409-420, 2015, doi: 10.11591/ijece.v5i3.pp409-420.
[18] M. Qiu and Y. Song, "Predicting the Direction of Stock Market Index Movement Using an Optimized Artificial Neural Network Model," *PloS One*, vol. 11, no. 5, pp. 1-11, 2016, doi: 10.1371/journal.pone.0155133.
[19] A. Nayak, M. M. Manohara Pai, Radhika M. Pai, "Prediction Models for Indian Stock Market," *Procedia Computer science*, vol. 89, pp. 441-449, 2016, doi: 10.1016/j.procs.2016.06.096.
[20] B. M. Henrique, Vinicius Amorim Sobreiro, Herbert Kimura, "Stock price prediction using support vector regression on daily and up to the minute prices," *Journal of Finance Data Science*, vol. 4, pp. 183-201, 2018, doi: 10.1016/j.jfds.2018.04.003.
[21] G. Domeniconi, G. Moro, A. Pagliarani, R. Pasolini, "Learning to predict the stock market Dow Jones index detecting and mining relevant tweets," *Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 2017, pp. 165-172, doi: 10.5220/0006488201650172.
[22] Y. Zuo and E. Kita, "Stock price forecast using Bayesian network," *Expert Systems with Applications*, vol. 39, no. 8, pp. 6729-6737, 2012, doi: 10.1016/j.eswa.2011.12.035.
[23] K. Radinsky, S. Davidovich, S. Markovitch, "Learning Causality for News Events Prediction," *Proceedings of the 21st International Conference on World Wide Web*, 2012, pp. 909-918, doi: 10.1145/2187836.2187958.
[24] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, D. C. L. Ngo, "Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment," *Expert Systems with Applications*, vol. 42, no. 1, pp. 306-324, 2015, doi: 10.1016/j.eswa.2014.08.004.
[25] R. Yadav, A. V. Kumar, A. Kumar, "News-based supervised sentiment analysis for prediction of futures buying behavior," *IIMB Management Review*, vol. 31, no. 2, pp. 157-166, 2019, doi: 10.1016/j.iimb.2019.03.006.
[26] B. Kavšek, "Using Words from Daily News Headlines to Predict the Movement of Stock Market Indices," *Managing Global Transitions*, vol. 15, pp. 109-121, 2017, doi: 10.26493/1854-6935.15.109-121.
[27] Q. Li, T. Wang, P. Li, L. Liu, Q. Gonga, and Y. Chenb, "The effect of news and public mood on stock movements," *Information Sciences*, vol. 278, pp. 826-840, 2014, doi: 10.1016/j.ins.2014.03.096.
[28] T. M. Nisar and M. Yeung, "Twitter as a tool for forecasting stock market movements: A short-window event study," *The Journal of Finance and Data Science*, vol. 4, pp. 101-119, 2018, doi: 10.1016/j.jfds.2017.11.002.
[29] Z. H. Kilimci and R. Duvar, "An Efficient Word Embedding and Deep Learning Based Model to Forecast the Direction of Stock Exchange Market Using Twitter and Financial News Sites: A Case of Istanbul Stock Exchange (BIST 100)," *IEEE Access*, vol. 8, pp. 188186-188198, 2020, doi: 10.1109/ACCESS.2020.3029860.
[30] B. Kaushik, H. Hemani, P. Vigneswara Ilavarasan "Social media usage vs. stock prices: an analysis of Indian firms," *Procedia Computer Science*, vol. 122, pp. 323-330, 2017, doi: 10.1016/j.procs.2017.11.376.

BIOGRAPHIES OF AUTHORS

Surinthip Sakphoowadon received her M.S. degree in Computer Information Systems from Assumption University, Thailand, in 2001. She is currently pursuing a Ph.D. degree in Information Technology at King Mongkut's University of Technology North Bangkok, Thailand. Her research interests are text mining and expert systems.

Nawaporn Wisitpongphan is an Assistant Professor in the Faculty of Information Technology and Digital Innovation and also a director of the Research Center of Information and Communication Technology at King Mongkut's University of Technology North Bangkok (KMUTNB), Thailand. She received her B.S., M.S., and Ph.D., in Electrical and Computer Engineering from Carnegie Mellon University in 2000, 2002, and 2008, respectively. Prior to joining KMUTNB, she was a researcher at General Motor Research Center, Warren, Michigan. While her expertise is in ad-hoc network and vehicle-to-vehicle communication, her current research interest is on smart environment and information technology management.

Choochart Haruechaiyasak received his Ph.D. degree from the Department of Electrical and Computer Engineering, University of Miami, in 2003. He has over 17-year experience working as a researcher at the National Electronics and Computer Technology Center (NECTEC), National Science and Technology Development Agency (NSTDA). Since 2020, he has co-founded and become the CEO of AI9, an AI startup company which focuses on developing solutions based on speech-to-text and text analytics technology.