P-ISSN 2586-9000
E-ISSN 2586-9027
Homepage : *https://tci-thaijo.org/index.php/SciTechAsia* Science & Technology Asia
Vol.28 No.3 July - September 2023 Page: [98-115]
Original research article

Prediction Models for Tourism Stock Market Trends during COVID-19 pandemic using News Sentiment Analysis with Data Mining: A Case Study

Thanaporn Wansri, Apapan Phaksuwan, Khemika Iamtae, Saifon Chaturantabut*

Department of Mathematics and Statistics, Faculty of Science and Technology,
Thammasat University, Pathum Thani 12120, Thailand

Received 15 July 2023; Received in revised form 19 September 2023
Accepted 22 September 2023; Available online 26 September 2023

ABSTRACT

This work introduces a process for predicting the trends of stocks in the tourism sector during COVID-19 pandemic using sentiment analysis of COVID19 news headlines with data mining techniques. The COVID-19 news headlines are first collected daily and analyzed via sentiment analysis to obtain their polarity based on naïve Bayes and neural network techniques. These polarity results are then used with the related stock historical data to predict the trend of the stock prices by K-nearest neighbor and decision tree classifications. In our numerical experiments, seven major stocks from the tourism and hotel business operated in Thailand are considered. Our proposed prediction models are shown to have accuracies ranging around 70%- 90%. The highest accuracy of about 90% is achieved when a neuron network is used in the sentiment analysis with the decision tree for predicting the stock trends.

Keywords: Data mining; Machine learning; Stock market; Sentiment analysis; Text mining; tourism industry; COVID-19

1. Introduction

Tourism is one of the most severely affected industries during the COVID-19 pandemic due to many factors, such as travel restriction, quarantine rules, and lockdown measures [1–4]. According to the World Tourism Organization 2021 [5], the revenue of the world-wide tourism industry decreased by 74% in 2020 during the COVID-19 pandemic. Since the COVID-

***Corresponding author:** saifon@mathstat.sci.tu.ac.th
doi: 10.14456/scitechasia.2023.49

19 vaccination started to be available, overall stock markets turned in a positive direction [6,7], including tourism businesses. As a result, more investors started to be interested in tourism stocks. To handle the risk of investing in tourism industry's stocks, it is important to investigate different suitable forecasting models.

Predicting stock markets is a challenging task due to their high volatility. The factors that may affect the markets include political, social, and economic issues. These factors are partially reflected through the news which can influence investors' sentiments. To make use of the news information, it is essential to have some efficient automated procedures that can collect and analyze textural data from a large volume of available news. This work aims to handle this issue by employing data mining techniques to predict stock trends based on news information and price historical data.

A number of existing works employed the time series and machine learning methods for predicting stock markets. The effect of quantitative easing on stock prices was investigated in [8] by using a structural time series model, which was estimated by maximum likelihood in a time-varying parametric framework, using the S&P 500 index as the dependent variable and the Fed's balance as an explanatory variable in addition to the unobserved components accounting for the behaviour of variables. A hybrid time-series model based on a feature selection method was proposed in [9] for forecasting the leading industry stock prices. In this proposed model, stepwise regression was first employed with multivariate adaptive regression splines and kernel ridge regression to select the key features used in the model. Then, this study constructed the forecasting model by using a genetic algorithm to optimize the parameters of support vector regression. The auto-regressive integrated moving average (ARIMA) models have been used to predict stock prices from New York Stock Exchange (NYSE) and Nigeria Stock Exchange (NSE) [10], Chinese stock prices [11] and Indian stock prices [12]. The results showed that ARIMA models have strong potential for short-term prediction. In [13], stock prices were predicted based on deep neural networks with financial product price data treated as a one-dimensional series that was generated by the projection of a chaotic system composed of multiple factors into the time dimension, and the price series was reconstructed using the time series phase-space reconstruction method. Other works on predicting stock prices using their historical data can be found in [14–16].

In practice, trends of stock prices can be heavily impacted by related financial news. Therefore, to increase the accuracy of stock prediction models, in addition to the stock historical information, we can incorporate the effect of news extracted from the process called sentiment analysis. Sentiment analysis can be considered as a natural language processing to classify the subjective information in source materials. It can be used to determine whether the attitude towards a particular topic or news is positive or negative. Recently, sentiment analysis has been introduced in predicting stock market behavior. In general, sentiment analysis can be categorized into lexicon based or machine learning based approaches. There are various lexicons that have been created and used to determine the sentiment for the text [17, 18]. The main benefit of using lexicon-based approaches is that it does not require any specific training data. However, since lexicon-based techniques use static lists, they may not be able to predict movement of stock markets accurately [19]. Alternative approaches based on machine learning methods are commonly used to overcome this issue. Stock price predictions have been performed in [20, 21] by using deep learning models. The work in [22] aimed to associate stock prices with web financial information time series based on support vector regression. In [23], the sentiment analysis on social media that incorporates part-of-speech tags into topic modeling was used for Iranian stock price movement prediction. A hybrid model combining a deep learning approach with a sentiment analysis model was proposed in [21] for stock price prediction by first employing a convolutional neural network (CNN) model for classifying the investors' hidden sentiments and then applying the long short-term memory (LSTM) neural network approach for analyzing the technical indicators from the stock market for the Shanghai Stock Exchange. Sentiment analysis has been used in [24] with machine learning approaches based on a support vector machine and generalized autoregressive conditional heteroskedasticity modeling and the results showed that investors' online opinions could have strong effect on value stocks relative to growth stocks. In [25], various deep-learning methods including support vector machines, linear regression, naïve Bayes and long short-term memory are used together with sentiment analysis for predicting stock prices.

Since the outbreak of the COVID-19 pandemic, there have been a number of literatures investigating its effect on the stock market. The prediction of U.S. oil markets during the COVID-19 was performed in [26] by using social media information with CNN. In [27], the impact of the COVID-19 pandemic on abnormal returns of tourism shares listed in the Shanghai and Shenzhen stock exchanges were investigated. The work in [28] empirically studied the market performance and response trends of Chinese industries to the COVID-19 pandemic using three main models for calculating abnormal returns: the average adjusted return rate model; the market index adjusted return rate model; and the market model. Economic losses in tourism for the city of Sorrento in Italy during the COVID-19 pandemic have been investigated by forecasting the loss of tourists and added value in 2020 using time-series analysis with autoregressive-integrated moving average (ARIMA) models [29]. Determinants of tourism stock returns in China during the COVID-19 have been studied in [30] using quantum computing with different deep learning prediction models. In [31], the impact of COVID-19 pandemic on logistics performance, economic growth and tourism industry of Thailand has been investigated by using ARIMA models. The majority of other works related to the impact of the COVID-19 on tourism are based on forecasting the number of incoming tourists or classifying customers' opinions in different countries [32–34]. This work aims to investigate the tourism stock trends during the COVID-19 by using machine learning and sentiment analysis with historical stock information. We consider the data of major tourism stocks in Thailand during the period when the COVID-19 vaccines were starting to be widely accessible. The main contributions and highlights of this work are summarized as follows.

- This work introduces stock trend prediction models that employ machine learning techniques and sentiment analysis with an efficient way to compute polarity scores.

- In this work, naïve Bayes and neural networks are applied for sentiment analysis to obtain the news polarities (positive/negative). These polarities are then used with the past information of these stock prices as inputs to perform K-nearest neighbor and decision tree to finally predict the stocks' future trends.
- Instead of using specific news directly related to the stock of interest, we use general news on COVID-19 in the sentiment analysis and train our classification model for determining its polarity based on the stock behavior. This approach can overcome the problem of calculating sentiment scores of the stocks whose companies may not appear directly on the news regularly.
- The proposed prediction models are shown to provide accuracies ranging around 70%- 90% for predicting daily trends of tourism stocks in Thailand during COVID19 when the vaccine started to be available.

The results of this study can help investors devise appropriate investment strategies in dealing with fluctuations of tourism stock prices.

The remainder of this paper is organized as follows. Section 2 provides the overview of the models used for predicting the tourism stock trends in this work. The details on sentiment analysis performed in this work are described in Section 3. Then, the approaches for forecasting the trend of stock prices are given in Section 4. Section 5 gives the information on data of tourism-related stock prices and financial news headlines on COVID19 used in our sentiment analysis. The numerical investigations are considered in Section 6. The conclusion and possible extensions are discussed in Section 7.

2. Prediction Model

This work mainly considers two independent/input variables, which are historical stock prices and sentiment score (polarity) from the related news. The dependent/output variable of interest represents the trend of the stock market. Therefore, there are two main parts for constructing the predicting model. The first part is the sentiment analysis using related COVID-19 news headlines.The second part is stock trend prediction using data mining techniques based on sentiment score and historical stock prices. The following diagram in Fig. 1 summarizes the overall process of performing trend prediction for daily stock price.

From Fig. 1, The process starts from collecting headlines of financial news that relates to COVID-19 in Thailand and collecting data related to stock prices. Then, we perform the sentiment analysis on the collected news headlines using naïve Bayes and neural network methods. The polarity results from this analysis are then used together with the historical stock data in classification techniques based on K-NN and decision tree to predict the trend of each stock.

3. Sentiment Analysis

The goal of sentiment analysis used in this work is to determine if the collected local COVID-19 news has positive or negative effect to the stock price of each company. We consider two classification methods, which are naïve Bayes and neural network, for determining the probability of sentiments based on the processed

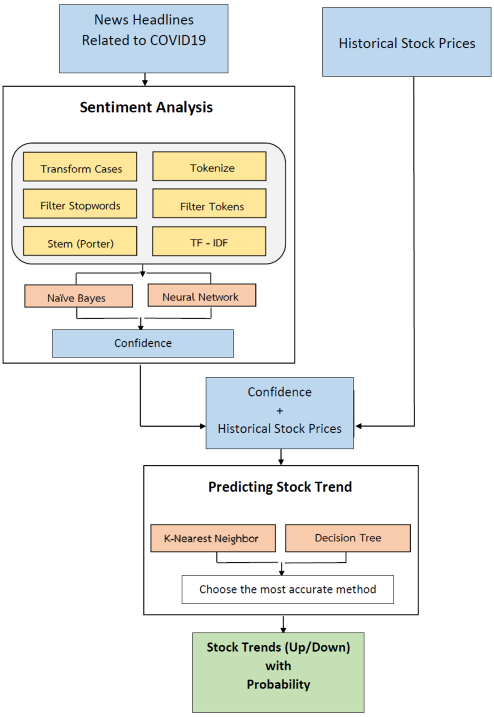


Fig. 1. The overview of the procedure for predicting the trend of the hotel stock prices.

data of the collected news headlines. In this work, we considered both cases that use the COVID-19 news on the same day and the previous five days. The latter case can be used to handle the situation when the current stock market might still be sensitive to news in the past.

3.1 Data preparation

To prepare the data for sentiment analysis, we perform initial preparation steps of text processing, such as tokenization, standardization, stop-word-removal, stemming, and filtering tokens.

- Transform cases: All uppercase letters are converted into lowercase in this step.
- Tokenize: This step splits the texts of a document into a series of words or tokens.
- Filter tokens: We can filters tokens or words by length. We set the minimum to 4 and the maximum to 25.
- Stemming: This step aims to reduce related forms of a common base form, e.g. reducing “fishing”, “fished”, “fish”, and “fisher” to the base word “fish”. Porter stemmer, one of the most popular stemmer, is used here.
- Filter stopwords: This step removes common English words such as ‘a’, ‘the’, etc.

After the preprocessing steps, we next extract the context of each word by generating **n-grams(terms)**. The term **n-gram** refers to a series of consecutive tokens or words of length n. This work uses $n = 2$ or bi-gram. Setting $n = 2$ means that a sequence of two-words for each document is generated. The effect of n used in n-gram can be found in [35,36].

Next,**term frequency–inverse document frequency (TF–IDF)**, which can be used to determine how relevant a word is to a document, is considered. TF–IDF is computed from the product of **t**erm frequency (TF) and**inverse document frequency (IDF)**. Consider a corpus that contains N documents. Term frequency (TF) is a measure that provides the frequency of a word (w) in a document (d). It is defined as the ratio of a word’s occurrence in a document to the total number of words in a document and can be computed by using the following formula

$$TF(w, d) = \frac{n_w}{n_d},$$

where n_w is the number of word w in the document d and n_d is the total number of words in the document d. Note that TF does not consider the importance of words. Some words, e.g. ‘of’, or ‘and’ can be most frequently present but are of little significance. Inverse Document Frequency (IDF) is used to measure the importance of a word. IDF provides weight to each word based on its frequency in the corpus D and it is calculated from the following formula

$$IDF(w, D) = \ln\left(\frac{N_D}{N_{Dw}}\right)$$

where N_D is the number of documents in the corpus D and N_{Dw} is the number of all documents in the corpus D that contain the word w. Finally, TF-IDF is computed from the product of TF and IDF as shown below

$$TF{-}IDF(w, d, D) = TF(w, d){*}IDF(w, D).$$

Notice that the TF value for a word w is large if the word appears many times in a given document d. Notice also that the IDF of a word w is zero when this word appears in all documents in the corpus and it gets small as the word is contained in many documents in the corpus. Therefore, IDF (and hence TF-IDF) gives more emphasis on the word that is rare in the corpus, while TF (and hence TF-IDF) provides more emphasis on the word that is more frequent in the document.

3.2 Naïve Bayes

The Naïve Bayes classifier is one of the most efficient methods for predicting the probability of the sentiment for text data [37]. It was shown to give accurate prediction with small processing time for a large data set [38]. It is a probabilistic classifier, which tells us the probability of the observation being in a class. For sentiment analysis, naïve Bayes classifier is used to identify the document to be either positive or negative based on TF-IDF values. It has been used to predict the polarity of each document because of its simplicity and speed [37].

Naïve Bayes classification technique is based on Bayes’ Theorem, which assumes the independence among predictors. From Bayes’s rule, we have

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

In sentiment analysis, for a document d, out of all classes $c \in C$ the classifier returns the class $\hat{c}$ which has the maximum probability given the document, i.e.

$$\hat{c} = \arg\max_{c\in C} P(c|d).$$

Using Bayes’ rule gives $\hat{c} = \arg\max_{c\in C} P(c|d) = \arg\max_{c\in C} \frac{P(c)P(d|c)}{P(d)}$. Note that, in practice, we can drop the denominator $P(d)$ because it is the same for all classes c. By using naïve Bayes assumption, we suppose that n features $f_1, f_2, ..., f_n$ in the document d are *independent*. I.e. a naïve Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. So, we have

$$\begin{aligned} P(d|c) &= P(f_1, f_2, ..., f_n|c) \\ &= P(f_1|c)\cdot P(f_2|c)\cdot ....\cdot P(f_n|c) \\ &:= \Pi_{i=1}^{n} P(f_i|c). \end{aligned}$$

That is, naïve Bayes classifier gives

$$\hat{c} = \arg\max_{c\in C} \frac{P(c)\Pi_{i=1}^{n} P(f_i|c)}{P(d)}.$$

Naïve Bayes classifier will be used to determine the polarity of the news headlines to be either positive or negative in this work.

3.3 Artificial neural network

An artificial neural network (or neural network) is a classification method that can handle the correlation/dependence between input variables [39]. The main advantage of neural networks lies in the fact that neural networks are data driven self-adaptive methods. They can adjust themselves to the data without any explicit specification of functional or distributional form for the underlying model.

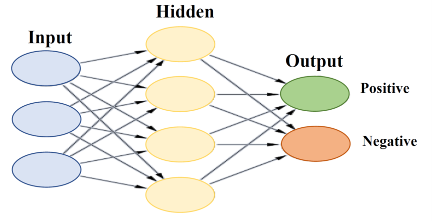


Fig. 2. Simplified diagram of neuron network for classifying input news headlines to be positive and negative.

Neural networks are generally represented by a network diagram as shown in Fig. 2, which consists of three layers: an input layer, a hidden layer and an output layer [40]. A neuron can be considered as a mathematical mapping that produces an output value in two steps. First, the neuron computes a weighted sum of its inputs and then applies an activation function in the hidden layer to this sum to derive its output. The activation function is usually a nonlinear function that is learned from the input data. For sentiment analysis, the nodes of the input layer come from the textural data and the output layer typically includes positive and negative nodes.

3.4 Training process and sentiment score

In general, when sentiment analysis of news or news headlines is used for predicting the trends of stock prices of any particular company, these news must be directly related to that company, e.g. containing the company's name. However, in practice, these related news may not be available, especially for small companies. In this work, we investigate the effect of COVID-19 news on the trends of stock prices for companies in the tourism sector. In our work, instead of using the news that directly contains each company's name, we use general news on COVID-19 and train our classification model for determining its polarity based on the stock behavior (e.g. prices are going up or going down). This approach can help us to overcome the problem of calculating the sentiment scores of the stocks whose companies may not appear directly in the news on a daily basis.

In this work, the predicted outputs from sentiment analysis are given in terms of confidence values for negative and positive sentiments, which are the probabilities for the news to be positive and negative, respectively. These confidence values can be converted to a value that indicates the polarity of the news, called "sentiment score" as follows. Suppose a given news has confidence values for being positive and being negative C_p and C_n, respectively. Then, the corresponding sentiment score s can be defined as

$$s = \begin{cases} C_p & \text{when} \quad C_p > C_n \\ -C_n & \text{when} \quad C_p \leq C_n. \end{cases} \tag{3.1}$$

Note that $C_p, C_n \in [0, 1]$ and $C_p + C_n = 1$. That is, for the news with confidence value for having positive polarity larger than confidence value for having negative polarity, the sentiment score is the same as the confidence value for being positive. For the news with larger confidence for having negative polarity, the sentiment score is the negative of the confidence value for being negative. Notice that the sentiment scores are in the range $[-1, 1]$.

4. Stock Trend Prediction

After obtaining sentiment scores from COVID-19 news headlines, we use these scores with historical stock data to predict the trend of each stock's prices. In this work, we investigate the prediction results from using K-nearest neighbor and decision tree approaches.

4.1 K-Nearest Neighbor

K-Nearest Neighbor (KNN) classifier is a method for classifying a new point based on its similarity to the available training points in the feature space. To measure the similarity, the Euclidean distance between the new point being classified and all the other points in the training set is computed. To label this new point, it looks at the K labeled points closest to the new point, which are its nearest neighbors, and the label of the new point is the class that most of these neighbors belong to. That is, the class label is assigned the same class as the majority of the nearest K instances in the training set. The KNN training is extremely fast, which is a great advantage for analyzing a large data set of stocks. The KNN model is also one of the best suited to evaluate financial data and it is less prone to data overfitting [41].

4.2 Decision tree

A decision tree is a tree-like structure that is used as a supervised classifier. It consists of a root node, decision nodes (internal nodes) and leaf nodes (terminal nodes) [42]. The root node is the topmost node in the tree diagram that has no incoming edges. The decision nodes contain conditions to split the data and they are related to the features. Each of the decision nodes has one incoming edge and has two or more outgoing edges. The leaf nodes correspond to class labels. Each of the leaf nodes has exactly one incoming node and no outgoing node.

In order to build a decision tree, we have to identify the attribute for the root node in each level, which can be done by various attribute selection measures. One of the most commonly used measures in the case of categorical variables is called "information gain" which can be computed from a quantity called "entropy." Entropy is a measure of disorder or impurity in the given data set. Suppose a data set has N classes. Let p_i be the probability of randomly choosing data from class i, for $i = 1, ..., N$. Let T be the set of training samples. The entropy is given by

$$E(T) = -\sum_{i=1}^{N} p_i \log_2(p_i).$$

Suppose A is an attribute with K different patterns. Let T_v be the subset of T with $A = v$, and $Values(A)$ is the set of all possible values of A, we can define the entropy of the set T for the attribute A as

$$E(T, A) = \sum_{v \in Values(A)} \frac{|T_v|}{|T|} E(T_v).$$

The information gain (IG) is then defined as

$$IG(S, A) = E(T) - E(T, A).$$

From the above, entropy measures impurity in the data and the information gain measures the expected reduction in entropy. The feature which has minimum impurity will be considered as the root node. Information gain is also a criterion for choosing the feature for splitting the tree at each step. Note that, besides the information gain, there are other measures for attribute selection, such the Gini index [43]. The resulting tree described above could generally over-fit the data, which would lead to inaccurate prediction for untrained data.

To overcome this issue, the pruning process has to be performed. This process removes the decision nodes starting from the leaf node such that the overall accuracy is not significantly changed.

Advantages of decision trees include providing interpretable predictive models and requiring no particular relationship between the responses. Decision trees are also independent of feature scaling and can handle large data sets [42,43].

5. Data Collection

There are two types of data used in this work: daily COVID-19 news and historical stock prices. We collect the headlines of local COVID-19 news in Thailand daily from reliable financial news websites, which are Thaienquirer.com and Nationthailand.com. For the stock prices, this work considers the following seven companies shown in Table 1 with the highest trading volumes in Tourism & Leisure sector from the Stock Exchange of Thailand (SET). We consider four attributes of the companies' stock prices, which are opening, high, low, and closing prices. These attributes are later used for future trend prediction. Note that, since these stock attributes are available on a daily basis, news headlines are also collected on a daily level. These data on stock prices and new headlines were collected during March 1, 2021 to August 31, 2021 when there were gradual roll-outs of COVID-19 vaccine in Thailand to help restore the tourism industry.

6. Numerical Experiments and Evaluation

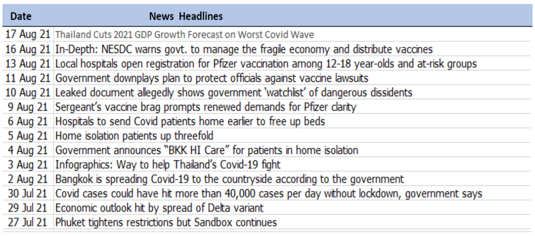
This section demonstrates the performance of the data mining methods and sentiment analysis described in the previous sections on predicting the behaviors of the stocks in Table 1. The numerical experiments for predicting the behavior of the stock prices in this work consist of two main steps. The first step is to classify the polarity of the news headline related to COVID-19 and the second step is to predict the trend of the stock prices. In the first step, for each tourism stock in Table 1, two classification models for predicting the polarity of the news headlines are constructed by using naïve Bayes and neural network methods. The results from each of these two models are then used in the second step for predicting the trend (up/down) of the stock prices for each company based on the KNN and decision tree classifiers.

Table 1. Example how to create tables.

Symbol	Quantity
CENTEL	Central Plaza Hotel Public Company Limited
DUSIT	Dusit Thani Public Company Limited
ERW	Erawan Group Public Company Limited
LRH	Laguna Resorts and Hotels Public Company Limited
MANRIN	Mandarin Hotel Public Company Limited
ROH	Royal Orchid Hotel
MINT	Minor International Public Company Limited

6.1 Sentiment analysis of COVID-19 news headlines

To perform sentiment analysis, we first collect the data of the COVID-19 news headlines as shown in Fig. 3 and apply naïve Bayes and neural network as described in the previous section with 80% of data for training and 20% for testing.



Date	News Headlines
17 Aug 21	Thailand Cuts 2021 GDP Growth Forecast on Worst Covid Wave
16 Aug 21	In-Depth: NESDC warns govt. to manage the fragile economy and distribute vaccines
13 Aug 21	Local hospitals open registration for Pfizer vaccination among 12-18 year-olds and at-risk groups
11 Aug 21	Government downplays plan to protect officials against vaccine lawsuits
10 Aug 21	Leaked document allegedly shows government 'watchlist' of dangerous dissidents
9 Aug 21	Sergeant's vaccine brag prompts renewed demands for Pfizer clarity
6 Aug 21	Hospitals to send Covid patients home earlier to free up beds
5 Aug 21	Home isolation patients up threefold
4 Aug 21	Government announces "BKK HI Care" for patients in home isolation
3 Aug 21	Infographics: Way to help Thailand's Covid-19 fight
2 Aug 21	Bangkok is spreading Covid-19 to the countryside according to the government
30 Jul 21	Covid cases could have hit more than 40,000 cases per day without lockdown, government says
29 Jul 21	Economic outlook hit by spread of Delta variant
27 Jul 21	Phuket tightens restrictions but Sandbox continues

Fig. 3. Examples of COVID-19 news headlines.

The predicted outputs from sentiment analysis are given in terms of confidence values for negative and positive sentiments, which are the probabilities for the news to be positive and negative, respectively. We compute the sentiment score for each news headline based on the formula given in Eq. (3.1). This work considers 3 cases of the headline news used for computing the sentiment scores. The first case uses one daily COVID-19 news, the second case uses multiple daily news, and third case uses multiple news in the previous 5 days. In the last two cases, the total number of multiple news can be different each day depending on the COVID-19 situation and the sentiment scores are the average value. In particular, if n news headlines are considered and s_j is the sentiment score of the news headline j for $j = 1, \ldots, n$, the the average sentiment score is simply $\frac{1}{n}\sum_{j=1}^{n} s_j$. To visualize the effect of COVID-19 news, we plot the corresponding average sentiment scores (and their corresponding 4-day moving average) obtained from naïve Bayes and neuron network with the prices of stocks related to travel and tourism sector in Thailand as shown in the first three plots in Figs. 5-11.

Notice that, from the first three plots of Figs. 5-11, it is hard to visually observe the relationship of the sentiment scores and the prices for each of these tourism stocks. The corresponding correlations between prices for these stocks and the sentiment scores from naïve Bayes and neural network models are given in Fig. 4, which shows that the ROH stock has the largest correlation when using naïve Bayes and, MANRIN has the largest correlation when using neural network. We next incorporate the sentiment scores with the historical information of these stocks to predict their future trends.

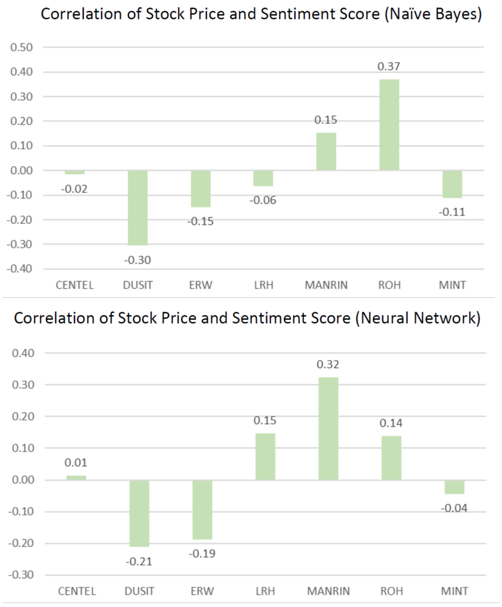


Fig. 4. Correlations of stock prices and the sentiment scores obtained from naïve Bayes and neuron network approaches.

6.2 Stock trend prediction

This section illustrates the predicted trends of tourism stock prices based on the sentiment analysis performed in the previous section together with the historical stock information. In particular, the inputs for the prediction models are sentiment scores, open, high, low and closing prices and the final output is the predicted trend of the stock prices. We apply the KNN and decision tree approaches to predict the stock trends for all 7 companies in Table 1. We use 80% of data for training and 20% for testing. Since there are 2 methods used in the sentiment analysis and 2 methods used in the stock trend prediction, there are total 4 combinations of the approaches to obtain the final outputs. The following abbreviations for these combinations are used in the displayed results.

BK: Sentiment analysis and trend prediction are performed via naïve Bayes and KNN, respectively.

BT: Sentiment analysis and trend prediction are performed via naïve Bayes and decision tree, respectively.

NK: Sentiment analysis and trend prediction are performed via neural network and KNN, respectively.

NT: Sentiment analysis and trend prediction are performed via neural network and decision tree, respectively.

The last plots in Figs. 5-11 illustrate the prediction accuracy of these 4 approaches for 3 cases of headline news mentioned earlier. Notice that the best approaches for predicting the stock trend for these companies are different and this may due to the fact that some of these stock prices are weakly correlated as shown in Fig. 12. The summary of the most accurate approach for each stock is shown in Fig. 13.

From Fig. 13, the best accuracies of these stocks are roughly ranging from 72% to 92%. In particular, CENTEL, DUSIT, ERW, LRH, MANRIN, ROH, and MINT attain the best accuracies of 78.26%, 86.96%, 72.73%, 73.91%, 81.25%, 92.86%, and 78.26%, respectively. Note that different stocks may attain their best accuracies by using different approaches. The highest accuracy of 92.86 % is achieved in the case of ROH stock when naïve Bayes and decision tree are used for sentiment analysis and trend prediction, respectively. Notice from Fig. 13 that 5 out of 7 stocks attain their best accuracies when using multiple daily news. In addition, most of these stocks achieve their best accuracies when neuron network is used for sentiment analysis and decision tree is used for predicting stock trends.

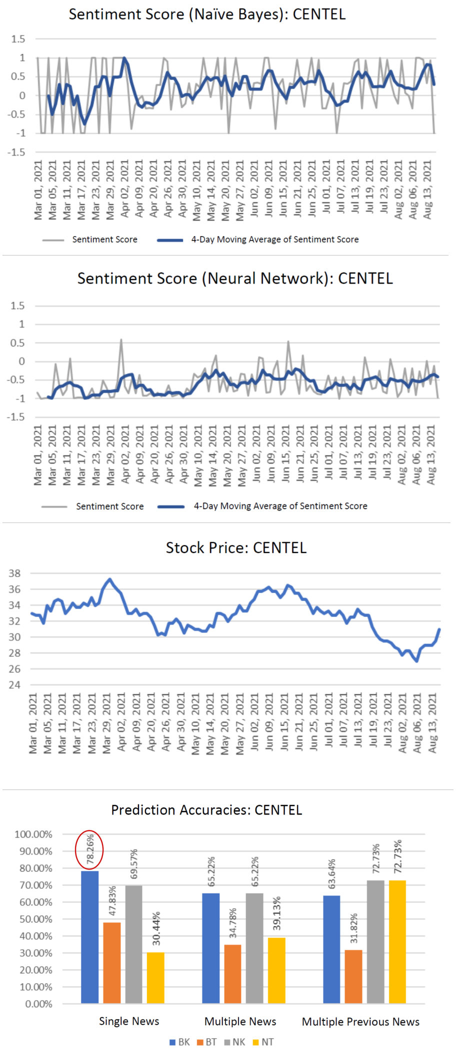


Fig. 5. The first three plots are the average sentiment scores obtained from naïve Bayes and neuron network and the plot of CENTEL stock prices, respectively. The last plot shows the accuracy of predicting stock trends from different approaches.

The averages of accuracies for all 7 companies are shown in Fig. 14, which demonstrates that the highest overall accuracy of 69.64 % is attained in the case of multiple news using neural network for sentiment analysis and decision tree for stock trend prediction.

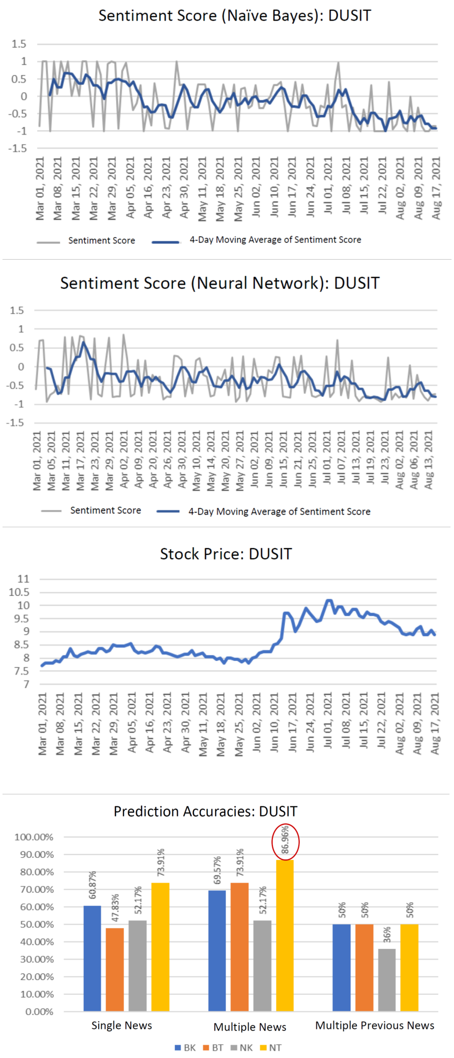


Fig. 6. The first three plots are the average sentiment scores obtained from naïve Bayes and neuron network and the plot of DUSIT stock prices. The last plot shows the accuracy of predicting stock trends from different approaches.

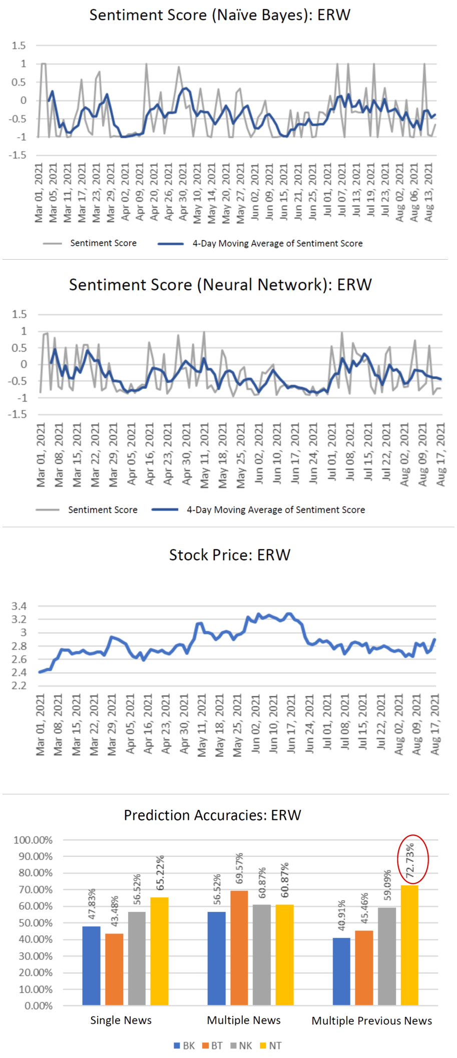


Fig. 7. The first three plots are the average sentiment scores obtained from naïve Bayes and Neuron Network and the plot of ERW stock prices. The last plot shows the accuracy of predicting stock trends from different approaches.

Note that, in addition to the above results displayed in this section, we have performed some numerical experiments that consider the sentiments of news headlines in the future (next 5 days) to predict the stocks' trends. The corresponding results are shown to be less accurate than the ones illustrated earlier in this section. E.g. for CENTEL, the prediction accuracies when using 4 approaches (i.e. BK, BT, NK, and NT) are ranging from 27.27% to 68.18% and, for MINT, the accuracies are ranging from 40.91% to 63.63%.

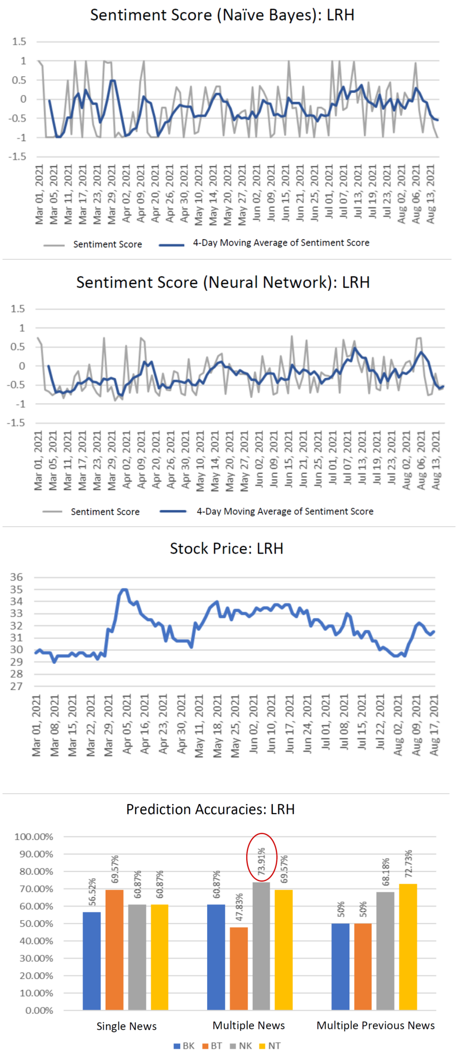


Fig. 8. The first three plots are the average sentiment scores obtained from naïve Bayes and neuron network and the plot of LRH stock prices. The last plot shows the accuracy of predicting stock trends from different approaches.

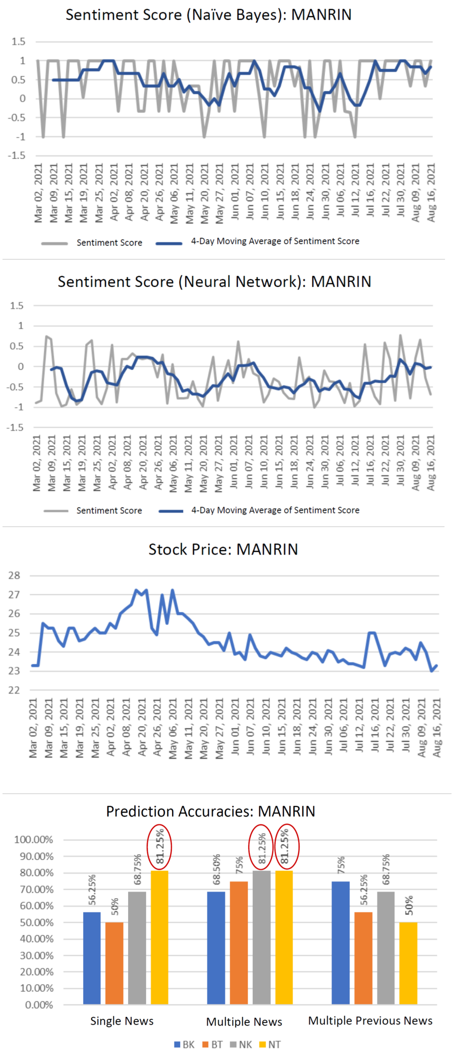


Fig. 9. The first three plots are the average sentiment scores obtained from naïve Bayes and neuron network and the plot of MANRIN stock prices. The last plot shows the accuracy of predicting stock trends from different approaches.

7. Conclusion

This work investigates the techniques for predicting the trends of tourism stocks based on news headlines related to COVID-19 pandemic and historical stock numeric data, i.e. open, high, and low prices. Naïve Bayes and neural networks are applied for sentiment analysis to obtain the news polarities (positive/negative). These polarities are then used with the past information of these stock prices as inputs to perform KNN and decision tree to finally predict the stocks' future trends. We consider the data of major tourism stocks in Thailand during the period when the COVID-19 vaccines were starting to be

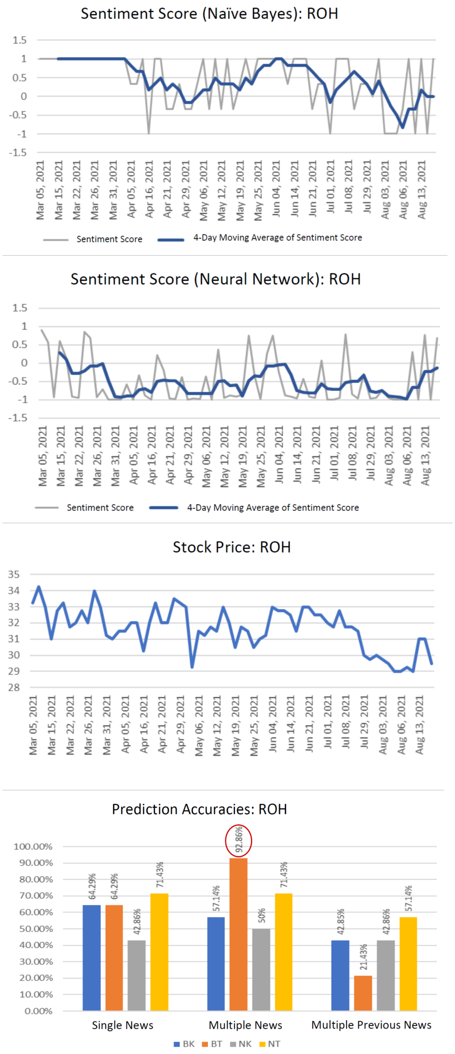


Fig. 10. The first three plots are the average sentiment scores obtained from naïve Bayes and neuron network and the plot of ROH stock prices. The last plot shows the accuracy of predicting stock trends from different approaches.

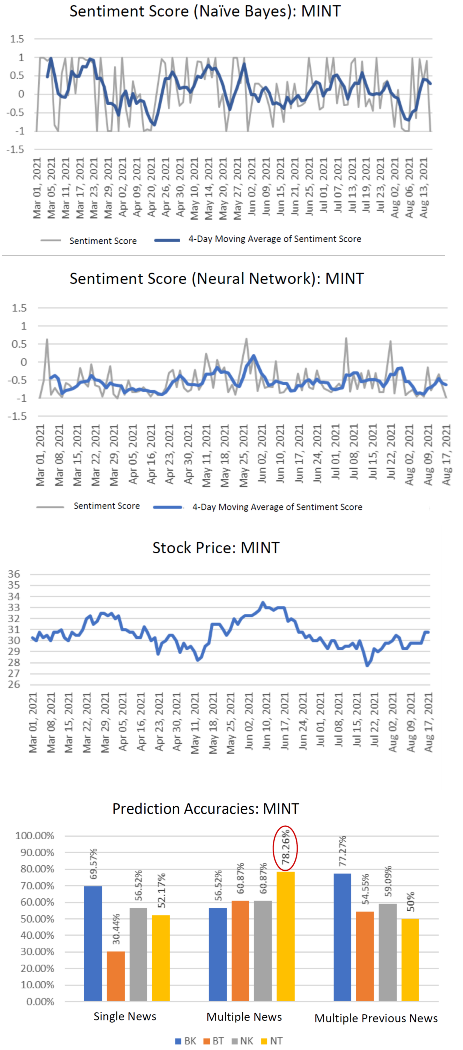
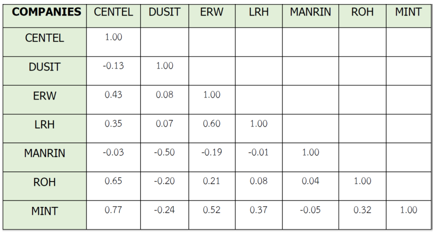


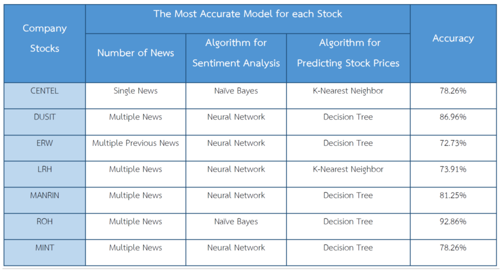
Fig. 11. The first three plots are the average sentiment scores obtained from naïve Bayes and neuron network and the plot of MINT stock prices. The last plot shows the accuracy of predicting stock trends from different approaches.

widely accessible. In our numerical experiments, we consider 3 cases of news: one daily COVID-19 news, multiple daily news, and multiple news in the previous 5 days. The results shows that the highest accuracies for the stocks considered in this work are ranging between 72.73 % and 92.86 %. On average, the best performance is achieved when naïve Bayes and decision tree are used together. It is important to note that these highest accuracies between 72.73 % and 92.86 % are obtained from different approaches depending on the stocks. These results are consistent with the previous works, e.g. [44]. Note also that the trend prediction for each of the



COMPANIES	CENTEL	DUSIT	ERW	LRH	MANRIN	ROH	MINT
CENTEL	1.00						
DUSIT	-0.13	1.00					
ERW	0.43	0.08	1.00				
LRH	0.35	0.07	0.60	1.00			
MANRIN	-0.03	-0.50	-0.19	-0.01	1.00		
ROH	0.65	-0.20	0.21	0.08	0.04	1.00	
MINT	0.77	-0.24	0.52	0.37	-0.05	0.32	1.00

Fig. 12. Correlation of Stock Prices.



Company Stocks	The Most Accurate Model for each Stock			Accuracy
	Number of News	Algorithm for Sentiment Analysis	Algorithm for Predicting Stock Prices	
CENTEL	Single News	Naïve Bayes	K-Nearest Neighbor	78.26%
DUSIT	Multiple News	Neural Network	Decision Tree	86.96%
ERW	Multiple Previous News	Neural Network	Decision Tree	72.73%
LRH	Multiple News	Neural Network	K-Nearest Neighbor	73.91%
MANRIN	Multiple News	Neural Network	Decision Tree	81.25%
ROH	Multiple News	Naïve Bayes	Decision Tree	92.86%
MINT	Multiple News	Neural Network	Decision Tree	78.26%

Fig. 13. Summary of the approach that gives the most accurate prediction for each stock.

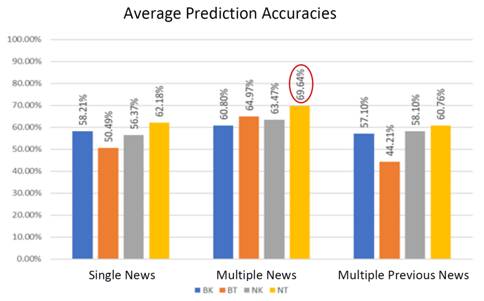


Fig. 14. Average of all accuracies from using different approaches for 7 companies in Table 1.

stocks shown in this work is derived from the sentiment analysis of general COVID-19 news headlines that do not necessarily relate to this particular stock company. Consequently, the results in this work may not be comparable with most of the previous works that considered the news containing the content related directly to the stock company of interest. Note that the stock companies considered in this work are mainly limited to hotel and resort business. This study therefore can be extended to other tourism-related industries, such as airlines and travel agencies. In addition, incorporating other existing machine learning techniques with sentiment analysis will be also taken into our future consideration.

Acknowledgements

The authors gratefully acknowledge the financial support provided by Faculty of Science and Technology, Thammasat University, Contract No. SciGR 12/2565.

References

[1] Yang Y, Zhang CX, Rickly JM. A review of early COVID-19 research in tourism: Launching the Annals of Tourism Research's Curated Collection on coronavirus and tourism. Annals of Tourism Research. 2021;91:103313.

[2] Gössling S, Scott D, Hall CM. Pandemics, tourism and global change: a rapid assessment of COVID-19. Journal of Sustainable Tourism. 2021;29(1):1–20.

[3] Sigala M. Tourism and COVID-19: Impacts and implications for advancing and resetting industry and research. Journal of business research. 2020;117:312–21.

[4] Duarte Alonso A, Kok SK, Bressan A, O'Shea M, Sakellarios N, Koresis A, et al. COVID-19, aftermath, impacts, and hospitality firms: An international perspective. International Journal of Hospitality Management. 2020;91:102654. Available from: https://www.sciencedirect.com/science/article/pii/S0278431920302061.

[5] Organisation WT. UNWTO world tourism barometer and statistical annex. 2021;7.

[6] Cong Nguyen To B, Khac Quoc Nguyen B, Van Thien Nguyen T, Thi Minh Nguyen P. Vaccine initiation rate and volatility in the international stock market during COVID-19. Bao and

Van Thien Nguyen, Tam and Thi Minh Nguyen, Phuong, Vaccine Initiation Rate and Volatility in the International Stock Market during COVID-19 (September 29, 2021). 2021.

[7] Khalfaoui R, Nammouri H, Labidi O, Jabeur SB. Is the COVID-19 vaccine effective on the US financial market? Public Health. 2021;198:177–9.

[8] Al-Jassar SA, Moosa IA. The effect of quantitative easing on stock prices: a structural time series approach. Applied Economics. 2019;51(17):1817–27.

[9] Tsai MC, Cheng CH, Tsai MI, Shiu HY. Forecasting leading industry stock prices based on a hybrid time-series forecast model. PloS one. 2018;13(12):e0209922.

[10] Ariyo AA, Adewumi AO, Ayo CK. Stock price prediction using the ARIMA model. In: 2014 UKSim-AMSS 16th international conference on computer modelling and simulation. IEEE; 2014. p. 106–12.

[11] Jarrett JE, Kyper E. ARIMA modeling with intervention to forecast and analyze Chinese stock prices. International Journal of Engineering Business Management. 2011;3(3):53–8.

[12] Mondal P, Shit L, Goswami S. Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices. International Journal of Computer Science, Engineering and Applications. 2014;4(2):13.

[13] Yu P, Yan X. Stock price prediction based on deep neural networks. Neural Computing and Applications. 2020;32(6):1609–28.

[14] Lahmiri S. Wavelet low-and high-frequency components as features for predicting stock prices with backpropagation neural networks. Journal of King Saud University-Computer and Information Sciences. 2014;26(2):218–27.

[15] Serletis A. Money and stock prices in the United States. Applied Financial Economics. 1993;3(1):51–4.

[16] Du J, Liu Q, Chen K, Wang J. Forecasting stock prices in two ways based on LSTM neural network. In: 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). IEEE; 2019. p. 1083–6.

[17] Strapparava C, Valitutti A, et al. Wordnet affect: an affective extension of wordnet. In: Lrec. vol. 4. Lisbon, Portugal; 2004. p. 40.

[18] Baccianella S, Esuli A, Sebastiani F. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10); 2010. .

[19] Li X, Xie H, Chen L, Wang J, Deng X. News impact on stock price return via sentiment analysis. Knowledge-Based Systems. 2014;69:14–23.

[20] Mohan S, Mullapudi S, Sammeta S, Vijayvergia P, Anastasiu DC. Stock price prediction using news sentiment analysis. In: 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService). IEEE; 2019. p. 205–8.

[21] Jing N, Wu Z, Wang H. A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. Expert Systems with Applications. 2021;178:115019.

[22] Liang X, Chen RC, He Y, Chen Y. Associating stock prices with web financial information time series based on support vector regression. Neurocomputing. 2013;115:142–9.

[23] Derakhshan A, Beigy H. Sentiment analysis on stock social media for stock

price movement prediction. Engineering Applications of Artificial Intelligence. 2019;85:569–78.

[24] Wu DD, Zheng L, Olson DL. A decision support approach for online stock forum sentiment analysis. IEEE transactions on systems, man, and cybernetics: systems. 2014;44(8):1077–87.

[25] Mehta P, Pandya S, Kotecha K. Harvesting social media sentiment analysis to enhance stock market prediction using deep learning. PeerJ Computer Science. 2021;7:e476.

[26] Wu B, Wang L, Wang S, Zeng YR. Forecasting the US oil markets based on social media information during the COVID-19 pandemic. Energy. 2021;226:120403.

[27] Liew VKS. Abnormal returns on tourism shares in the Chinese stock exchanges amid the COVID-19 pandemic. Available at SSRN 3863889. 2020.

[28] He P, Sun Y, Zhang Y, Li T. COVID–19's impact on stock prices across different sectors—An event study based on the Chinese stock market. Emerging Markets Finance and Trade. 2020;56(10):2198–212.

[29] Agovino M, Musella G. Economic losses in tourism during the COVID-19 pandemic. The case of Sorrento. Current Issues in Tourism. 2021:1–25.

[30] Pan WT, Huang QY, Yang ZY, Zhu FY, Pang YN, Zhuang ME. Determinants of Tourism Stocks During the COVID-19: Evidence From the Deep Learning Models. Frontiers in Public Health. 2021;9. Available from: https://www.frontiersin.org/articles/10.3389/fpubh.2021.675801.

[31] Laeeq Razzak Janjua FM, Sukjai P, Rehman A, Yu Z. Impact of COVID-19 pandemic on logistics performance, economic growth and tourism industry of Thailand: an empirical forecasting using ARIMA. Brazilian Journal of Operations & Production Management. 2021;18(2):e2021999.

[32] Sontayasara T, Jariyapongpaiboon S, Promjun A, Seelpipat N, Saengtabtim K, Tang J, et al. Twitter sentiment analysis of Bangkok tourism during COVID-19 pandemic using support vector machine algorithm. Journal of Disaster Research. 2021;16(1):24–30.

[33] Mishra RK, Urolagin S, Jothi J, Neogi A, Nawaz N. Deep learning-based sentiment analysis and topic modeling on tourism during Covid-19 pandemic. Frontiers in Computer Science. 2021;3(10.3389).

[34] Obembe D, Kolade O, Obembe F, Owoseni A, Mafimisebi O. Covid-19 and the tourism industry: An early stage sentiment analysis of the impact of social media and stakeholder communication. International Journal of Information Management Data Insights. 2021;1(2):100040.

[35] Jimenez M, Maxime C, Le Traon Y, Papadakis M. On the impact of tokenizer and parameters on n-gram based code analysis. In: 2018 IEEE International Conference on Software Maintenance and Evolution (ICSME). IEEE; 2018. p. 437–48.

[36] Tremblay A, Tucker BV. The effects of N-gram probabilistic measures on the recognition and production of four-word sequences. The Mental Lexicon. 2011;6(2):302–24.

[37] Pratiwi IYR, Asmara RA, Rahutomo F. Study of hoax news detection using naïve bayes classifier in Indonesian language. In: 2017 11th International Conference on Information & Communication Technology and System (ICTS). IEEE; 2017. p. 73–8.

[38] Banchs RE. Text mining with MATLAB®. Springer; 2013.

[39] Priddy KL, Keller PE. Artificial neural networks: an introduction. vol. 68. SPIE press; 2005.

[40] Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. vol. 2. Springer; 2009.

[41] Asghar MZ, Rahman F, Kundi FM, Ahmad S. Development of stock market trend prediction system using multiple regression. Computational and mathematical organization theory. 2019;25(3):271–301.

[42] Quinlan JR. Learning decision tree classifiers. ACM Computing Surveys (CSUR). 1996;28(1):71–2.

[43] Banfield RE, Hall LO, Bowyer KW, Kegelmeyer WP. A comparison of decision tree ensemble creation techniques. IEEE transactions on pattern analysis and machine intelligence. 2006;29(1):173–80.

[44] Hindrayani KM, Fahrudin TM, Aji RP, Safitri EM. Indonesian Stock Price Prediction including Covid19 Era Using Decision Tree Regression. In: 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI). IEEE; 2020. p. 344–7.