Received: 9 November 2018 | Revised: 10 May 2019 | Accepted: 16 May 2019

DOI: 10.1002/for.2616

RESEARCH ARTICLE

WILEY

Using social media mining technology to improve stock price forecast accuracy

Jia-Yen Huang | Jin-Hao Liu

Department of Information Management, National Chin-Yi University of Technology, Taichung City, Taiwan, ROC

Correspondence
Jia-Yen Huang, Department of Information Management, National Chin-Yi University of Technology, No. 57, Sec. 2, Zhongshan Rd., Taiping Dist., Taichung City 41170, Taiwan, ROC.
Email: jygiant@ncut.edu.tw

Abstract

Many stock investors make investment decisions based on stock-price-related chip indicators. However, in addition to quantified data, financial news often has a nonnegligible impact on stock price. Nowadays, as new reviews are posted daily on social media, there may be value in using web opinions to improve the performance of stock price prediction. To this end, we use logistic regression to screen the chip indicators and establish a basic stock price prediction model. Then, we employ text mining technology to quantify the unstructured data of social media opinions on stock-related news into sentiment scores, which are found to correlate significantly with the change extent of the stock price. Based on the findings that the higher the sentiment scores, the lower the prediction accuracy of the logistic regression model, we propose an improved prediction approach that integrates sentiment scores into the logistic regression model. Our results show that the proposed model can improve the prediction accuracy for stock prices, and can thus provide a new reference for investment strategies for stock investors.

KEYWORDS

chip indicators, logistic regression model, prediction accuracy, sentiment scores, text mining

1 | INTRODUCTION

In order to make the best investment decisions, stock investors generally collect related information from sources such as TV media, newspapers, magazines and the Internet. However, faced with various information, investors often cannot identify which information is most important. How to build up high-yield investment strategies with the help of quantitative indicators and stock price forecasting tools is obviously a topic of concern to investors.

Stock price fluctuations are dynamic, nonlinear, nonstationary, and carry a lot of noise, which makes the stock price difficult to forecast. Stock price forecasting has been a research topic of wide concern in academia and financial domains. Early research has mainly been based on random walk theory (RWT) and efficient market hypothesis (EMH). RWT argues that stock price fluctuations are random and, hence, the next step of the stock price is as irregular as a person walking on a square. EMH suggests that designing a system based on any information to predict stock price changes is impossible because all information has already been reflected in the existing stock prices.

According to previous EMH research, stock prices are mainly driven by new information rather than current and past prices. Since the news is unpredictable, the stock price will follow the random walk model and cannot exceed 50% accuracy prediction (Qian & Rasheed, 2007). However, many studies have shown that stock prices are not random, but can indeed be predicted to a certain extent. News may be unpredictable, but early metrics can be extracted from online social media (blogs, Twitter, etc.) to predict changes in various economic and business indicators (Gallagher & Taylor, 2002), and to explore the relationship between financial news and stock price changes (Schumaker, Zhang, Huang, & Chen, 2012).

The decision-making process of investors is influenced by both rationality and emotions. Many financial analysts and investors count on rational approaches to predict stock prices, such as machine learning prediction methods. For example, some use a time series model based on financial theory to predict stock prices. Compared with other methods, an artificial neural network (ANN) is usually chosen as a stock price forecasting tool. However, these methods of using machine learning prediction models have shortcomings because the stock market is always affected by system uncertainties and other unknown factors (G. Zhang, Patuwo, & Hu, 1998). In order to improve the prediction accuracy, some scholars have proposed that, in addition to the use of financial indicators to build machine learning prediction models, other relevant factors should be considered (Y. Wang & Wang, 2016). For example, Tsaih, Hsu, and Lai (1998) proposed using mixed stock price and technical indicators to build an artificial intelligence trading model through machine learning. This can overcome the limitations of using a single method and can achieve long-term stable profitability. Adebiyi, Ayo, Adebiyi, and Otokiti (2012) argued that using technical variables and fundamental indicators to construct a stock forecasting model can make stock forecasting more accurate.

Stock price change is determined based on the decisions of many investors. For the same news, each investor may have a different interpretation, which in turn leads to different investment decisions (Mittermayer, 2004). With the popularity of the Internet, many investors post their views on stocks of concern on investment-related social media, where they also see other's views to consider as an investment reference (Rechenthin, Street, & Srinivasan, 2013). Recently, more researchers have paid attention to incorporating social network emotional factors responses to stock-market-related news into prediction models (Bollen, Mao, & Zeng, 2011; Yang, Mo, & Liu, 2015; X. Zhang, Shi, Wang, & Fang, 2017).

The aim of this study is to build a high-precision stock forecasting model to assist stock investors in making investment decisions. Since both the financial-based indicators and investors' opinions on social media have valuable reference to investors, this study takes both of these aspects into consideration in the construction of a stock price prediction tool.

The remainder of this paper is organized as follows. Section 2 reviews the development of social media mining with a special focus on its application to stock prediction. Section 3 provides the framework of this study. The details of constructing a stock price prediction model based on the chip indicators are described. This section also introduces the opinion phrase extraction rules for identifying opinion words and degree words from the reviews posted on social media to quantify reviews into sentiment scores. Section 4 first presents the screened chip indicators that are influential with respect to stock price change based on the results of logistic regression analysis. Then, by considering the sentiment scores, an improved prediction model is proposed. Finally, Section 5 presents conclusions and future work.

2 | RELATED WORK

Behavioral finance scholars argue that stock market prices do not fully follow random walk and efficient market assumptions, which means stock prices can be predicted to some extent. Behavioral finance has proven that financial decisions are clearly driven by emotions (Rubbaniy, Asmerom, Rizvi, & Naqvi, 2014). The sentiments of reviews posted on social media reportedly affect trading, investment decisions, and activities in the stock exchanges, and thus become a valuable resource (Zhu et al., 2014). In recent years, significant progress has been made in sentiment tracking techniques that directly extract public sentiments from social media content. In addition to using chip indicators and machine learning to predict stock price, public sentiments have also been integrated into stock price prediction models (Bollen et al., 2011). This section first introduces the model of using machine learning technology to predict stock prices and identifies the chip indicators relevant to stock change. Next, the literature using social media mining techniques to predict stock price change is reviewed.

2.1 | Chip-indicator-based prediction model

Chips are kinetic energy of the stock market. Investors with more funds generally have the ability to influence the stock market. It is beneficial to make a winning investment decision by analyzing the chip indicators to grasp the flow of funds. Information related to the chips of Taiwanese stocks, such as changes in the trading volume of domestic and foreign investors, are transparent and open quantitative data.

Using machine learning prediction models and the chip indicators to predict stock market fluctuations is not uncommon in the literature. Researchers of machine learning prediction models use different indicators depending on their research objectives and purposes.

For example, Liao and Wang (2010) proposed a stochastic time-effective neural network for financial market prediction. C. M. Hsu (2011) offered 16 technical indicators to forecast the stock market. Kanas and Yannopoulos (2001) proposed several financial indicators to forecast the stock market. Ballings, Van den Poel, Hespeels, and Gryp (2015) argued that predicting the direction of stock prices is very beneficial for investment and the prediction accuracy of using logistic regression is good.

In order to screen for appropriate indicators to construct a forecasting model, this study collates variables of chips from relevant literature in the financial field. In sum, we collected 26 variables, as shown in Table 1. Since too many variables are unfavorable for analysis and decision making, this study uses binary logistic regression to find the key variables of chips and establish a predictive model in Section 4.

2.2 | Predict stock prices using social media mining

In recent years, there have been several studies exploring the possibility of using financial news and classification techniques to improve the traditional machine learning prediction model. Wuthrich et al. (1998) collected news articles from five popular financial websites and used text mining techniques and various neural networks to predict trends in the Hong Kong market. Their results proved to be better than the accuracy of random predictors. Based on a self-developed system for classifying online financial news, Wu (2007) claimed that their trading strategy of investing in the Taiwan Stock Weighted Index (TWSI) can yield an average return of 5.4% per month. Using text mining technology, Micu, Mast, Milea, Frasincar, and Kaymak (2009) developed a web ontology language-based system to track the news of the Nasdaq index and predict the potential polarity (positive, neutral or negative) of its effect on the company. S. Wang, Zhe, Kang, Wang, and Chen (2008) designed an ontology-based expert reasoning system to integrate the domain knowledge by building data mining models consisting of multiple news-related variables with certain financial trading activities. Based on financial news, Schumaker et al. (2012) used comprehensive language, finance, statistics, and technical sentiment analysis to develop a stock price forecasting engine, which performed better than the market average.

Zhao and Wang (2015) proposed an outlier data mining algorithm that used anomalies on distributions of trading volume to predict upward trends of stock prices. Although they claimed that this method could effectively predict stock trend and make profits on the Chinese stock market in long-term usage, the characteristics used to predict stock prices are not sufficient to make accurate predictions (Y. Wang & Wang, 2016).

Due to the development of the Internet and the popularity of various social media, it is easier to get information by crawling data from websites, including documents, news, blogs, forums, emails, etc. Among them, Sina Weibo, Twitter and Facebook attract increasing numbers of users to receive and share information or comments because of their wide content and fast dissemination.

With the rise of social media and the promotion of relevant platforms for user-generated content, user opinions have begun to play a greater role in the stock market. Zhu et al. (2014) reported in 2008 that about 25% of adults are indirectly dependent on investment advice spread through social media. In response to the increase in online news and social media, web-mining-based forecasting methods have been extensively studied to improve the performance of financial market forecasts (Xu, Li, Jiang, & Cheng, 2012). Social media mining stems from the relevant areas of data mining, which explores the pattern of structured data rather than unstructured data. Text sentiment analysis, also known as opinion mining, has been used to discover and classify the sentiments of stock comments from social media into a set of predefined emotion categories.

TABLE 1 Chip indicators related to stock price fluctuations

Stock price change (SPC)	Trading volume (TV)	Turnover in value (TIV)	Change of turnover in value (CTIV)
Margin loan buy (MLB)	Margin loan sell (MLS)	Margin loan cash pay (MLCP)	Margin balance long (MBL)
Stock loan buy (SLB)	Stock loan sell (SLS)	Stock loan cash pay (SLCP)	Margin balance short (MBS)
Foreign investors buy (FIB)	Foreign investors sell (FIS)	Juridical person buy (JPB)	Juridical person sell (JPS)
Investment trust buy (ITB)	Investment trust sell (ITS)	Proprietary trader buy (PTB)	Proprietary trader sell (PTS)
Foreign investors shareholding (FISH)	Foreign investors shareholding ratio (FISHR)	Day offset of margin purchasing and short selling (DOMPSL)	Ratio of margin loan to stock load (RMLSL)
Naked day trade (NDT)	Margin day trade (MDT)		

Market sentiment is an important qualitative factor lacking in most machine learning prediction models in the literature. Since market sentiment follows a random distribution, just as stock prices follow random walks, it is necessary to determine whether using the sentiment evaluated from social media to predict stock price is effective or not. Das and Chen (2007) extracted sentiments from online messages and examined the relationship between sentiment and stock values. Gilbert and Karahalios (2010) confirmed that public mood has an impact on the stock market by extracting an indicator of public anxiety from LiveJournal posts and verifying that its changes could predict the S&P 500 index. Bollen et al. (2011) explored the association between Twitter users' sentiments and the Dow Jones Industrial Average (DJIA), and found that public sentiment can be tracked from social media and correlated with the DJIA. Xu et al. (2012) used text mining techniques to mine sentiments hidden in Twitter tweets, and the results showed that the method can quickly learn the sentiments of online users. Kim, Jeong, and Ghani (2014) designed a method that included natural language processing classification and the extraction of sentiments and opinions expressed by authors, but the effects were not satisfactory.

Zhu et al. (2014) proposed a Microblogging Influence Detection Scheme (RMIDS) to detect the impact of microblogging posts on the stock market. Their results showed that the sentiments in microblogging is significantly related to Chinese stock prices. Bogle and Potter (2015) proposed a hybrid prediction model that included sentiment analysis and machine learning algorithms including decision trees, neural networks, and support vector machines to predict the Jamaica Stock Exchange. In addition to using four major factors affecting stock price trends, including closing price, trading volume, market index, and daily turnover rate, Y. Wang and Wang (2016) also focused on the impact of investor sentiment and proposed a sentiment discrepancy index (SDI). Although they argued that stock price prediction models could forecast more accurately by using social media mining combined with other information, their reason for defining SDI was not clearly explained.

The aforementioned literature generally agrees on the close correlation between market sentiment and stock change. However, many studies applied classification techniques to predict the potential polar effect of market sentiments on stock change, which is not suitable for combining with a machine learning predictive model. Although a few studies proposed hybrid prediction models by using machine learning algorithms that included market sentiments and the factors affecting stock price, the results were not satisfactory. This study argues that the decision-making process of investors should be influenced by rationality and emotions, and both factors should be included in the prediction model of stock price. Therefore, this study hopes to analyze social media comments by text mining technology and quantify these comments into sentiment scores. Thus this study proposes a regression model combining both indicators of rationality and indicators of emotions to improve the accuracy of prediction.

3 | METHODOLOGY

3.1 | Research framework

This study first collects chips data for Hon Hai Precision Industry Co. (HHPIC) to establish a logistic prediction model and identify important indicators. HHPIC is currently ranked second in terms of the Taiwan stock market value. We collect the reviews and replies about HHPIC posted on PTT, which is the largest and most famous terminal-based bulletin board system (BBS) site in Taiwan. After conducting word segmentation and part-of-speech (POS) tagging on the collected HHPIC reviews, we construct opinion phrase extraction rules to identify the opinion words associated with the attribute words. Next, we use the number of replies as weights to calculate the positive and negative scores of each review to obtain quantified sentiment scores. Finally, we explore the relationship between the sentiment scores and prediction accuracy of the stock price change extent, and propose an improved stock price prediction model based on the sentiment scores. Figure 1 illustrates the framework of this study.

3.2 | Data sources

The collected HHPIC-related data consist of two types: (1) chip indicators and (2) social media reviews posted on PTT responding to news and replies. The data of HHPIC chip indicators shown in Table 1 were collected from the Taiwan Stock Market Observation Post System. In total, data were collected on 245 days during 2017.

This study collected HHPIC-related social media reviews and replies posted on PTT in 2017. PTT has more than 1.5 million registered users, with more than 20,000 reviews (articles) and 500,000 replies (comments) posted every day. The founding purpose of PTT was to provide an online platform for academic purposes; therefore, many studies use PTT social media reviews as a source of research data. For example, C. M. Huang, Chan, and Hyder (2010) described how to apply user reports on PTT to construct a disaster emergency response and disaster public health management system when typhoon Morakot hit Taiwan in 2009. S. C. Hsu, Chiu, Hung, and Chen (2013)

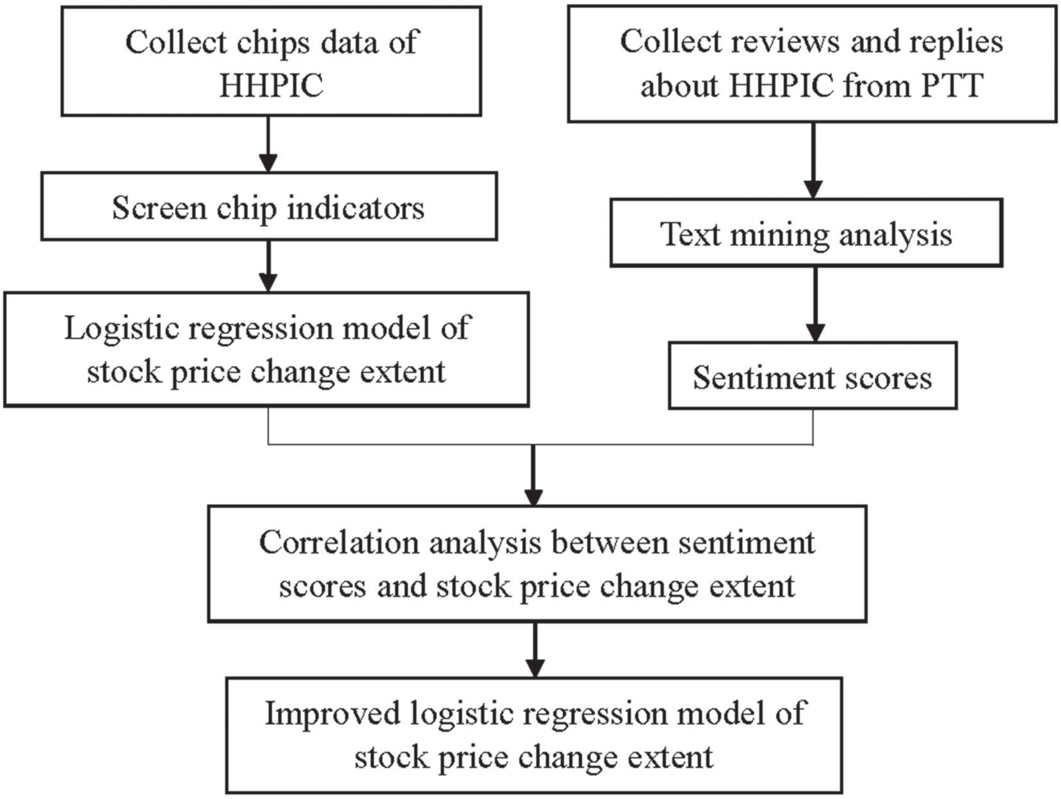

Collect chips data of HHPIC
Screen chip indicators
Logistic regression model of stock price change extent
Collect reviews and replies about HHPIC from PTT
Text mining analysis
Sentiment scores
Correlation analysis between sentiment scores and stock price change extent
Improved logistic regression model of stock price change extent

FIGURE 1 The research framework

analyzed 150 valid samples obtained from PTT and found that an increasing number of patients and their families were learning about diseases and health through the Internet, and that many cancer patients supported each other on the Internet. Chen (2014) used PTT discussions of Taiwanese netizens regarding the Diaoyutai/Senkaku Islands to investigate how Taiwanese people responded to territorial disputes, and he found evidence of a national identity issue in sovereignty disputes.

This study recorded the title, URL, author, time, text, and replies information for each review including the number of replies, the author ID of the reply, and the tags and content of the replies. In total, we recorded 68 reviews and 4,708 replies. All forecasts in this study are out of sample. Among the collected data, 70% of the data (171 days from January 3 to September 14, 2017) were used as training data to stabilize the prediction model, and 30% of the data (74 days from September 15 to December 29, 2017) were used as testing data to evaluate the prediction accuracy.

3.3 | Screening of chip indicators and construction of prediction model

Since too many chip indicators are collected to support decision making by investors, the indicators must be screened based on consideration of their statistical significance. The machine learning methods commonly used in the literature to establish prediction models are mostly support vector machine (SVM), neural networks and logistic regression. Many studies use logistic regression to screen variables. For example, Austin and Tu (2004) applied logistic regression to select variables and determine the probability of death after being in hospital. In view of the effectiveness of the binary logistic regression model, this study uses this method to establish a prediction model and identify the important factors affecting the extent of stock price change.

This study conducts backward elimination logistic regression by starting with a saturated model and then iteratively removing the least useful independent variables (predictors). The SPSS v. 22 software system is used to perform the experiments. The stock price change extent is represented by P, which contains three levels: 0%, 0.5%, and 1%. The outcome $Y = 1$ means that tomorrow's change extent of HHPIC will exceed P, otherwise $Y = 0$.

3.4 | Opinion mining

After collecting data from PTT, we conduct word segmentation on the corpus and tag the POS of each word using the Chinese knowledge information processing group

(CKIP), which is a Chinese parser developed by Academia Sinica.

To transfer the reviews into scores, PTT users' opinions associated with attribute words must be identified. The methods of extracting opinions are generally divided into two types. The first is to extract the required opinions from an established database, such as the National Taiwan University Sentiment Dictionary (NTUSD), HowNet and E-HowNet. The second is to use a combination of words or grammatical structure to extract opinions. For example, Turney (2002) proposed five patterns of tags for English reviews, and the matching adjectives are identified as opinions. Hu and Liu (2004) proposed using adjectives near the attribute words as opinion words. Popescu and Etzioni (2007) proposed 10 extraction rules for extracting opinion words and noted that opinion phrases could be noun, verb, or adverb phrases. J. Y. Huang (2017) proposed six patterns of opinion phrase extraction rules after studying the grammatical structure of attribute words and opinion words in many reviews. Since this study deals with Chinese reviews, we borrow the opinion phrase extraction rules from Huang.

After identifying the opinion words associated with the attribute words, this study compares the extracted opinion words with the NTUSD and determines their polarity. It is worth noting that, when a sentence has a negative word such as “no” or “does not have,” the polarity of the sentence is likely to be reversed. That is, if there is a negative word before or after the positive opinion, it is a comment expressing negative emotions.

There may be adverbs with enhanced semantics before and after the opinion words, called degree words. Because degree words have a strengthening effect on the opinion words, they also affect the opinion scores of public sentiment. This study uses HowNet to quantify the degree of adverbs into six levels according to their intensities, including “extreme,” “most,” “very,” “more,” “over,” and “insufficiently,” which are rated in sequence from 6 to 1, respectively.

3.5 | Sentiment scores

Each review on PTT has different effects on the change extent of stock prices for tomorrow. For a news review, the more people that reply to it, the higher its importance. Therefore, the number of replies can be regarded as the weight of the reviews. This study uses degree words and opinion words to calculate the basic score of the review, and then uses the number of replies as the weight to obtain the final sentiment score.

Assuming there are j reviews posted on PTT for daily news, we define the index t_{ij} as the attribute word of sentence i in review j. The polarity of the opinion words is set as $P\,(t_{ij}) = 1$ for a positive opinion and $P\,(t_{ij}) = -1$ for a negative opinion. We assign values to degree words, $D\,(t_{ij})$, which range from 1 to 6 according to their intensities. The negation modifier $N\,(t_{ij}) = -1$ when a negation word is present, otherwise it is equal to 1.

To compute the sentiment score for each review j, this study transforms each sentence in each review via its corresponding quadruple $<t_{ij}, P\,(t_{ij}), D\,(t_{ij}), N\,(t_{ij})>$ by the following formula:

$$\mathrm{S}_j = \sum\nolimits_i P\left(t_{ij}\right) \times D\left(t_{ij}\right) \times N\left(t_{ij}\right) \tag{1}$$

The total sentiment score for each day is computed as follows:

$$S_T = \sum\nolimits_j S_j \times W_j \tag{2}$$

where W_j is the weighting of review j and is equal to the total number of replies of review j. We present a review posted on PTT on January 8, 2017 as an example, shown in Table 2, to explain the steps of sentiment score calculation. After employing text mining technology, there are nine sentences matching the opinion phrase extraction rules. Among these, seven sentences express positive opinions and two sentences express negative opinions. The degree levels corresponding to each sentence are listed in Table 2. No negation modifiers are found. According to Equation 1, we have the scores +31 for positive expressions and − 8 for negative expressions. Therefore, the basic score of the review is 23. Since the number of replies for this review is 65, according to Equation 2, the sentiment score of this review is 23 × 65 = 1,495.

4 | DATA ANALYSIS AND DISCUSSION

This section presents the results of logistic regression analysis and social media mining, and discusses the

TABLE 2 Example of calculating sentiment score of a review

		Number of sentences matching opinion phrase extraction rules	
Level	**Rating**	**Positive opinions**	**Negative opinions**
Insufficiently	1	0	0
Over	2	1	0
More	3	1	1
Very	4	0	0
Most	5	4	1
Extreme	6	1	0

performance of using the results of social media mining to improve the prediction model.

4.1 | Screening of chip indicators

Based on the literature, this study collected a total of 26 chip indicators related to stock price fluctuations, as shown in Table 1. We used these indicators as the independent variables for constructing the binary logistic regression model. Although three different stock price change extent values (P)—0%, 0.5%, and 1%—were investigated, only the calculation steps of $P = 0\%$ are presented here, owing to space limitations.

This study uses the Hosmer–Lemeshow test to estimate the goodness-of-fit of the logistic regression model and the Wald statistic to test the statistical significance of each coefficient in the model. During the process of backward elimination, the indicators with small values in the Wald test are considered insignificant and are therefore screened out. From the above analysis, the appropriate model for $P = 0\%$ with which to fit the data is the following:

$$\begin{aligned} Y = &-0.393 - 4.240x_1 + 3.55x_2 - 9.072x_3 + 3.984x_4 \\ &+ 8.021x_5 - 4.356x_6 - 3.418x_7 \\ &+ 68.204x_8 - 67.816x_9 + 5.274x_{10} \end{aligned} \tag{3}$$

where x_1 is SPC, x_2 is CTIV, x_3 is MLB, x_4 is MLS, x_5 is SLB, x_6 is DOMPSL, x_7 is NDT, x_8 is FISH, x_9 is FISHR and x_{10} is PTB. A positive regression coefficient means that the explanatory variable increases the probability of the outcome whereas a negative regression coefficient means that the variable decreases the probability of that outcome (J. Y. Huang, 2014).

This study uses a confusion matrix to summarize the performance of logistic regression prediction for binary classification tasks. This square matrix consists of columns and rows that list the number of instances as absolute or relative *actual class* versus *predicted class* ratios, as shown in Table 3. Taking $P = 0\%$ as an example, TP (true positive) is the number of correct predictions of the stock price that increase more than 0%, TN (true negative) is the number of correct predictions of the stock price that increase less than 0%, FP (false positive) is the number of incorrect predictions of the stock price that increase more than 0%, and FN (false negative) is the number of incorrect predictions of the stock price that increase less than 0%.

If the values of TP and TN are high, investors will increase profit or reduce loss according to the confusion matrix. During the process of backward elimination for 0%, the values in the confusion matrix change for each step, as shown in Table 4. The TP value in step 1 is 69 and it increases to 74 in step 25, meaning the prediction accuracy increases as we conduct the logistic regression with the process of backward elimination. Moreover, although the values of FP increase from 25 to 27 in Table 4, the overall prediction accuracy increases from 71.9% to 73.7%. By conducting forward selection using the same data, we get TP = 73, TN = 47, FN = 17, and FP = 34. The prediction accuracy is 70.2%, which is somewhat lower than the results using backward elimination. Therefore, this study uses backward stepwise logistic regression to construct the prediction model and screen the key chip indicators.

To compare the prediction accuracy between rolling and nonrolling estimation, we divide all collected data (245 days) into 10 parts, and conduct rolling estimations steps proposed by Tashman (2000) as follows:

Step 1.. Using 70% of the total data, from day 1 to 171, as the training data to estimate the prediction accuracy of part 8 (day 172–195), part 9 (day 196 to 221), and part 10 (day 222–245), respectively.

Step 2.. Using 70% of the total data, from day 25 to 195, as the training data to estimate the prediction accuracy of part 9 (day 196–221), and part 10 (day 222–245), respectively.

Step 3.. Using 70% of the total data, from day 51 to 221, as the training data to estimate the prediction accuracy of part 10 (day 222–245).

Step 4.. Calculate the averaged prediction accuracy.

As shown in Table 5, the analysis results show that the averaged prediction of rolling estimations is no better than that of nonrolling estimation. Therefore, we adopt a nonrolling estimation method for accuracy prediction in this study.

According to the analysis results of the backward selection method, the 26 indicators listed in Table 1 are screened. The optimal chip indicators of the logistic prediction model and prediction accuracy are slightly

TABLE 3 Confusion matrix

	Prediction	
Actual	TP	FN
	FP	TN

TABLE 4 Confusion matrix by backward elimination for P=0%

Step	1	T69	FN = 21	Overall prediction accuracy = 71.9%
		F27	54	
	25	T74	FN = 16	Overall prediction accuracy = 73.7%
		F29	52	

TABLE 5 Prediction accuracy comparison between rolling and nonrolling estimation

				P-value		
Training data	**Part**	**Testing day**		**0%**	**0.5%**	**1%**
Day1–171	8	Day172–195	Total testing days	24	24	24
			Correct days	15	14	18
			Prediction accuracy	0.625	0.583	0.750
	9	Day 196–221	Total testing days	25	25	25
			Correct days	15	18	19
			Prediction accuracy	0.620	0.72	0.76
	10	Day 222–245	Total testing days	25	25	25
			Correct days	15	18	19
			Prediction accuracy	0.60	0.72	0.76
Day 25–195	9	Day 196–221	Total testing days	25	25	25
			Correct days	15	19	19
			Prediction accuracy	0.60	0.76	0.76
	10	Day 222–245	Total testing days	25	25	25
			Correct days	12	17	19
			Prediction accuracy	0.48	0.68	0.76
Day196–221	10	Day 222–245	Total testing days	25	25	25
			Correct days	12	17	19
			Prediction accuracy	0.48	0.68	0.76
Averaged prediction accuracy of rolling estimations				0.564	0.690	0.758
Prediction accuracy of non-rolling estimations				0.608	0.743	0.757

different for different degrees of stock price change extent, as shown in Table 6.

Some of the indicators appearing in Table 6 exist in more than one degree of stock price change extent, including TV, CTIV, MLB, NDT, FISH, and FIB, showing their important influence on stock prices. In general, the larger the value of NDT—that is, the naked day trade—the more active the stock is. As can be seen in the data, NDT exists in all three *P* conditions, meaning the investors of HHPIC tend to conduct short-swing trading. It is worth noting that the prediction for larger stock price change extent is closely related to foreign capital and trading volume; undoubtedly, the investment decisions of foreign investors make a large impact on HHPIC stock price fluctuations.

TABLE 6 Screened indicators for different *P*-values

	P-value		
	0%	**0.5%**	**1%**
Screened indicators	NDT SPC FISHR MLB DOMPSL FISH PTB CTIV SLB MLS	NDT TV FIB MLB RMLSL	NDT TV FIB MBS RMLSL FISH PTB CTIV TIV MDT ITS
Number of days of data collection	110	74	40
Prediction accuracy	60.81%	74.32%	75.67%

4.2 | Improved prediction model using sentiment scores

This study uses the reviews and replies of one previous day to predict the stock price change extent. This study conducted word segmentation and POS tagging for the reviews through the CKIP system. Based on the phrase extraction rules combined with the opinion words database, this study identifies attribute words, opinion words and degree words related to the stock price change extent of HHPIC. After using the number of replies as a weighting, the daily reviews are quantified into a sentiment score.

In order to confirm the feasibility of using the sentiment score to improve the logistic regression prediction model, this study first investigates the correlation between the sentiment score, number of replies, trading volume, stock price change extent, and prediction accuracy.

4.2.1 | Correlation analysis of trading volume with sentiment score and number of replies

As shown in Table 7, the trading volume is statistically significantly positive correlated with the number of

TABLE 7 Correlation analysis of trading volume with sentiment score and number of replies

Number of replies	Correlation coefficient	0.200
	Significance	0.000
Sentiment score	Correlation coefficient	0.031
	Significance	0.583

replies, meaning that the higher the number of replies, the higher the trading volume. Large numbers of replies mean that there are more responses to specific news reviews, which may incur more frequent transactions, resulting in a higher trading volume. On the other hand, the trading volume is not significantly correlated with sentiment score. This is probably because when the sentiment score is a positive high value, everyone is optimistic that the stock price is set for a boost, which generates a mentality of reluctance to sell, resulting in a lower trading volume. Conversely, when the sentiment score is a negative large value, everyone is pessimistic about the stock and few people are willing to buy stocks, which also results in a low trading volume.

TABLE 8 Correlation analysis of stock price change with sentiment score and number of replies

Number of replies	Correlation coefficient	−0.074
	Significance	0.135
Sentiment score	Correlation coefficient	−0.186
	Significance	0.000

4.2.2 | Correlation analysis of stock price change with sentiment score and number of replies

As shown in Table 8, the stock price change is statistically significantly negative correlated with the sentiment score. A high sentiment score indicates that many of the investors' moods are high and of the same polarity, and the synchronized investment decisions result in a large stock price change. On the other hand, although a large number of replies may result in a higher trading volume, the varied investors may have different opinions on the reviews, which makes the stock price change insignificantly related to the number of replies.

In Figure 2, the solid lines represent the sentiment scores and the dotted lines represent the daily stock price change of HHPIC. The relationship between the two is worth noting at three time periods within 2017. First, HHPIC announced its 2016 financial report at the beginning of April. The net profit after tax for the whole year surged to a record high of 148.6 billion, which is inconsistent with the expectation of foreign investors. Since the haze of poor revenue suddenly swept away, the total trading volume reached a record high of 170,000 on the April 5, 2017. However, the sentiment score is inconsistent with the stock price change. One of the reasons is that the content of the news is quite different from what was expected. The other reason is that the stock market closed due to four consecutive holidays right after the news was published. In other words, holidays dilute the intensity of discussion on this issue. Although there are not many voices on PTT,

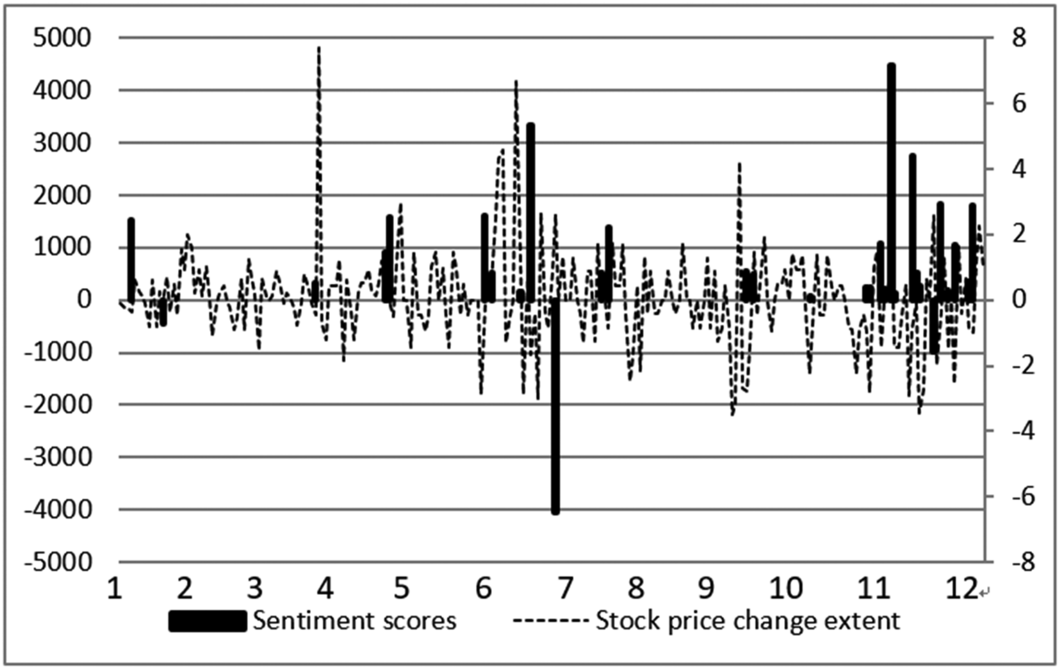


FIGURE 2 Comparison of sentiment score with stock price change

investors rushed to buy HHPIC stocks on April 5 after 4 days of waiting. The second period ranged from June to September, during which the sentiment scores were larger (positive or negative) when the stock price fluctuated violently. During the third period, from November to the end of 2017, situations of high sentiment score occurred frequently, meaning HHPIC was hotly debated by investors and the variation of the stock price change was large.

This study conducts correlation analysis between the sentiment score and stock price change, since their relationship is not easy to confirm from Figure 2. According to the test of the three correlation analysis methods, as can be seen from Table 9, the stock price change and sentiment scores have significant relationships and all of them are negatively correlated.

4.2.3 | Correlation analysis of sentiment scores with prediction accuracy

If the accuracy of the logistic regression prediction is lower in the periods with high sentiment scores, then it is insufficient simply to use the chip indicators for prediction when the influence of the news is high. This study compares the averaged forecasting accuracy of 39 days with higher sentiment scores to the averaged forecasting accuracy of the whole year. As shown in Figure 3, the average accuracy of the prediction model does decrease during the 39 days when the sentiment score is high, regardless of the stock price change. Because sentiment score has a certain correlation with prediction accuracy, this study considers sentiment score as a variable affecting the accuracy of prediction and accordingly proposes an improved prediction model.

TABLE 9 Correlation analysis of the stock price change and sentiment scores

Correlation method	Correlation coefficient and significance	Results
Pearson	Correlation coefficient Significance	−0.134 0.036
Kendall's tau	Correlation coefficient Significance	−0.183 0.000
Spearman's rho	Correlation coefficient Significance	−0.228 0.000

4.2.4 | Improved logistic regression prediction model

Since sentiment score is significantly correlated with stock price change, this study considers sentiment score as a new variable in the logistic regression model and reevaluates the change in prediction accuracy. The new prediction model equation for $P = 0\%$ is as follows:

$$\begin{aligned} Y = {} & 6.089 - 33.486x_1 + 50.812x_2 - 46.560x_3 + 9.224x_4 \\ & + 24.808x_5 - 12.029x_6 - 22.371x_7 - 4.562x_8 \\ & - 3.534x_9 + 11.608x_{10} + 3.346x_{11} + 6.051x_{12} \end{aligned} \tag{4}$$

where x_1 is sentiment score (SS), x_2 is TV, x_3 is TIV, x_4 is CTIV, x_5 is MBS, x_6 is MDT, x_7 is RMLSL, x_8 is NDT, x_9 is FIB, x_{10} is FISH, x_{11} is ITS, and x_{12} is PTB.

As shown in Figure 4, after adding sentiment scores to the proposed model, the prediction accuracy of the stock

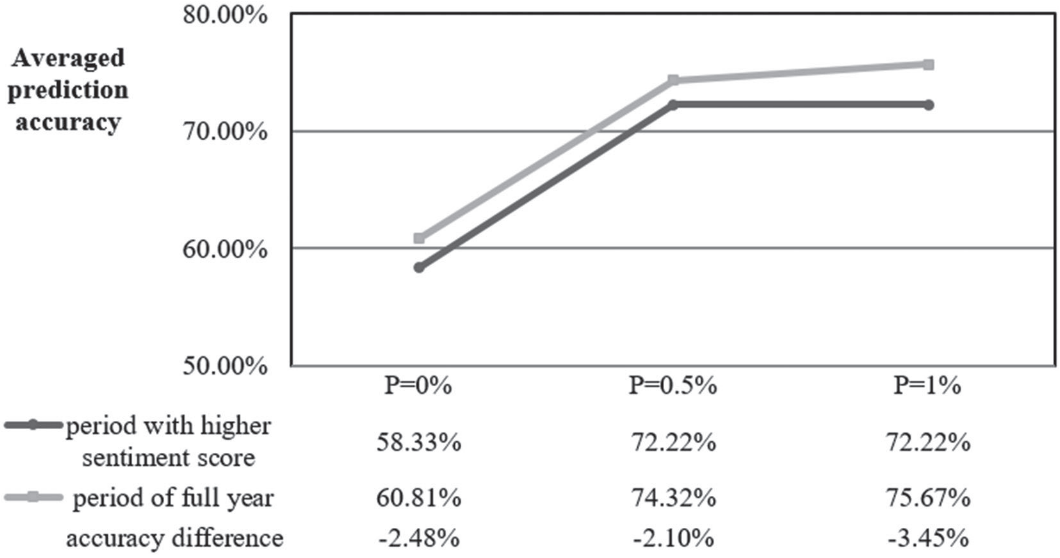
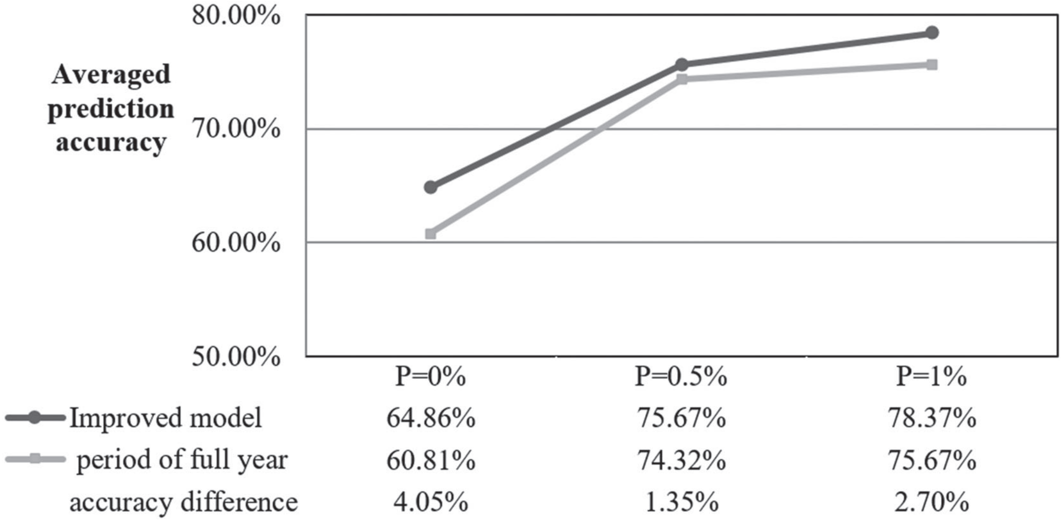


FIGURE 3 Comparison of average prediction accuracy for a high sentiment score period and for the whole year



	P=0%	P=0.5%	P=1%
Improved model	64.86%	75.67%	78.37%
period of full year	60.81%	74.32%	75.67%
accuracy difference	4.05%	1.35%	2.70%

FIGURE 4 The improved prediction accuracy of the proposed prediction model

price change extents of 0%, 0.5%, and 1% can be increased by 4.05%, 1.35%, and 2.70%, respectively.

5 | CONCLUSIONS AND FUTURE WORK

5.1 | Research contributions

The major academic contributions of this study are fourfold. First, this study employed text mining technology to quantify the investment opinions of social media into sentiment scores. Second, this study demonstrated a correlation between the sentiment score and the stock price change. Third, this study demonstrated that during periods with high sentiment scores the prediction accuracy of the logistic regression model decreased. Fourth, this paper proposed an approach to improve prediction accuracy by incorporating the sentiment scores into a logistic regression prediction model.

The practical contributions of this study are the disclosure of the key chip indicators that are influential to stock price change. Moreover, the key indicators are different for prediction models of different stock price change extent. Based on these results, investors can develop better investment strategies with fewer variables.

Notably, Table 8 shows a negative correlation between the stock price change and sentiment score. We note similar results in other studies. X. Zhang, Fuehres, and Gloor (2011) found a significantly negative correlation between the percentage of emotional tweets and the Dow Jones, Nasdaq, and S&P 500, which suggested that public sentiment and emotions in user-generated content on Weibo could drive stock market changes. Schumaker et al. (2012) conducted experiments using the stock price forecasting engine AZFinText and found that investors reacted more strongly to negative news reports. The authors argued that this result may be attributed to market traders operating in a contrarian manner, such as selling stocks in response to good news releases and buying stocks in response to bearish news releases (Schumaker et al., 2012). Thus whether the news is good or bad should be determined by the stock price trend after the news is published because the market reaction is not always consistent with the tone of the news. Rumors or estimations released by market analysts are often widely circulated before the official news release and the stock price has already responded to the news. Therefore, it is not uncommon for the market to retreat, defying good news, or to rally, despite bad news.

5.2 | Future work

This study has some limitations that require further consideration. First, as this study analyzed only three values of stock price change extent, future studies could consider a larger range of change extents. Second, this study considered only the number of replies for the weighting of the sentiment score. However, the polarity of the replies on the review were not all consistent. In future, we may consider incorporating positive and negative opinions of the replies to develop a more sophisticated method for calculating sentiment scores. Third, future studies could collect reviews from different social media platforms, although their phrase extraction rules may need to be redefined. Fourth, prediction methods other than logistic regression may be considered for comparison. Fifth, to confirm whether the proposed model of sentiment score is applicable to other stocks, more stocks should be selected for future analyses.

DATA AVAILABILITY STATEMENT

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

ORCID

Jia-Yen Huang https://orcid.org/0000-0001-6984-0057

REFERENCES

Adebiyi, A. A., Ayo, C. K., Adebiyi, M. O., & Otokiti, S. O. (2012). Stock price prediction using neural network with hybridized market indicators. *Journal of Emerging Trends in Computing and Information Sciences*, *3*(1), 1–9.

Austin, P. C., & Tu, J. V. (2004). Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *Journal of Clinical Epidemiology*, *57*(11), 1138–1146.

Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. *Expert Systems with Applications*, *42*(20), 7046–7056.

Bogle, S. A., & Potter, W. D. (2015). SentAMaL: A sentiment analysis machine learning stock predictive model. In *Proceedings of the International Conference on Artificial Intelligence (ICAI)* (pp. 610–615). Berlin, Germany: Springer.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8.

Chen, B. (2014). Sovereignty or identity? The significance of the Diaoyutai/Senkaku Islands dispute for Taiwan. *Perception*, *19*(1), 107–119.

Das, S. R., & Chen, M. Y. (2007). Yahoo! For Amazon: Sentiment extraction from small talk on the web. *Management Science*, *53*(9), 1375–1388.

Gallagher, L. A., & Taylor, M. P. (2002). Permanent and temporary components of stock prices: Evidence from assessing macroeconomic shocks. *Southern Economic Journal*, *69*(2), 345–362.

Gilbert, E., & Karahalios, K. (2010). Widespread worry and the stock market. In *4th International AAAI Conference on Weblogs and Social Media (ICWSM)* (pp. 59–65). Menlo Park, CA: AAAI.

Hsu, C. M. (2011). A hybrid procedure for stock price prediction by integrating self-organizing map and genetic programming. *Expert Systems with Applications*, *38*(11), 14026–14036.

Hsu, S. C., Chiu, C. M., Hung, Y. W., & Chen, L. H. (2013). Exploring driving factors of providing online social support: An example of PTT anti-cancer board. *Sun Yat-Sen Management Review*, *21*(3), 511–544.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 168–177). New York, NY: ACM.

Huang, C. M., Chan, E., & Hyder, A. A. (2010). Web 2.0 and internet social networking: A new tool for disaster management? Lessons from Taiwan. *BMC Medical Informatics and Decision Making*, *10*(1), 57.

Huang, J. Y. (2014). Innovative replenishment management for perishable items using logistic regression and grey analysis. *International Journal of Business Performance Management*, *15*(2), 138–157.

Huang, J. Y. (2017). Web mining for the mayoral election prediction in Taiwan. *Aslib Journal of Information Management*, *69*(6), 688–701.

Kanas, A., & Yannopoulos, A. (2001). Comparing linear and nonlinear forecasts for stock returns. *International Review of Economics and Finance*, *10*(4), 383–398.

Kim, Y., Jeong, S. R., & Ghani, I. (2014). Text opinion mining to analyze news for stock market prediction. *International Journal of Advances in Soft Computing and its Applications*, *6*(1), 1–13.

Liao, Z., & Wang, J. (2010). Forecasting model of global stock index by stochastic time effective neural network. *Expert Systems with Applications*, *37*(1), 834–841.

Micu, A., Mast, L., Milea, V., Frasincar, F., & Kaymak, U. (2009). Financial news analysis using a semantic web approach. In *Semantic knowledge management: An ontology-based framework* (pp. 311–328). Hershey, PA: IGI Global.

Mittermayer, M. A. (2004). Forecasting intraday stock price trends with text mining techniques. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*. Washington, DC: IEEE Computer Society.

Popescu, A. M., & Etzioni, O. (2007). Extracting product features and opinions from reviews. In *Natural language processing and text mining* (pp. 9–28). London, UK: Springer.

Qian, B., & Rasheed, K. (2007). Stock market prediction with multiple classifiers. *Applied Intelligence*, *26*(1), 25–33.

Rechenthin, M., Street, W. N., & Srinivasan, P. (2013). Stock chatter: Using stock sentiment to predict price direction. *Algorithmic Finance*, *2*(3–4), 169–196.

Rubbaniy, G., Asmerom, R., Rizvi, S. K. A., & Naqvi, B. (2014). Do fear indices help predict stock returns? *Quantitative Finance*, *14*(5), 831–847.

Schumaker, R. P., Zhang, Y., Huang, C. N., & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems*, *53*(3), 458–464.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, *16*(4), 437–450.

Tsaih, R., Hsu, Y., & Lai, C. C. (1998). Forecasting S&P 500 stock index futures with a hybrid AI system. *Decision Support Systems*, *23*(2), 161–174.

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 417–424). Stroudsburg, PA: Association for Computational Linguistics.

Wang, S., Zhe, Z., Kang, Y., Wang, H., & Chen, X. (2008). An ontology for causal relationships between news and financial instruments. *Expert Systems with Applications*, *35*(3), 569–580.

Wang, Y., & Wang, Y. (2016). Using social media mining technology to assist in price prediction of stock market. In *IEEE International Conference on Big Data Analysis (ICBDA)* (pp. 1–4). New York, NY: IEEE.

Wu, Y. (2007). *Predicting the trend of Taiwan Weighted Stock Index with text mining techniques* (Master's thesis). Department of Information Management, National Central University, Taiwan.

Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K., & Zhang, J. (1998). Daily stock market forecast from textual web data. In *IEEE International Conference on Systems, Man, and Cybernetics* (Vol. 3) (pp. 2720–2725). New York, NY: IEEE.

Xu, W., Li, T., Jiang, B., & Cheng, C. (2012). Web mining for financial market prediction based on online sentiments. In *Proceedings of the Pacific Asia Conference on Information Systems (PACIS)* (p. 43). Atlanta, GA: Association for Information Systems.

Yang, S. Y., Mo, S. Y. K., & Liu, A. (2015). Twitter financial community sentiment and its predictive relationship to stock market movement. *Quantitative Finance*, *15*(10), 1637–1656.

Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, *14*(1), 35–62.

Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting stock market indicators through twitter "I hope it is not as bad as I fear". *Procedia: Social and Behavioral Sciences*, *26*, 55–62.

Zhang, X., Shi, J., Wang, D., & Fang, B. (2017). Exploiting investors social network for stock prediction in China's market. *Journal of Computational Science*, *28*, 294–303.

Zhao, L., & Wang, L. (2015). Price trend prediction of stock market using outlier data mining algorithm. In *IEEE Fifth International Conference on Big Data and Cloud Computing (BDCloud)* (pp. 93–98). New York, NT: IEEE.

Zhu, N., Wang, Y., Cheng, C., Xu, W., Zhang, Y., Zou, P., & Awan, M. S. K. (2014). Regression-based microblogging influence detection framework for stock market. *Journal of Networks*, *9*(8), 2129–2136.

AUTHOR BIOGRAPHIES

Jia-Yen Huang is currently a professor in the Department of Information Management, National Chin-Yi University of Technology, Taiwan. He received the Ph.D. degree in Mechanical Engineering from National Taiwan University in 1987. His present research interests include Data mining, Innovation Management, and Patent Analysis.

Jia-Hao Liu received his Master's degree in the Department of Information Management from National Chin-Yi University of Technology in 2018. His research interests include data mining and computer programming. He is currently a data engineer of Essences Information Co., Ltd.

How to cite this article: Huang J-Y, Liu J-H. Using social media mining technology to improve stock price forecast accuracy. *Journal of Forecasting*. 2020;39:104–116. https://doi.org/10.1002/for.2616