IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 1, 2019, accepted December 14, 2019, date of publication December 25, 2019, date of current version January 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2962221

Can the Exchange Rate Be Used to Predict the Shanghai Composite Index?

JUN ZHANG[1], YUAN-HAI SHAO[2], LING-WEI HUANG[1], JIA-YING TENG[1], YU-TING ZHAO[1], ZHU-KAI YANG[1], AND XIN-YANG LI[1]

[1]School of Economics, Hainan University, Haikou, 570228, China
[2]Management School, Hainan University, Haikou, 570228, China

Corresponding author: Yuan-Hai Shao (shaoyuanhai21@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 11926349, Grant 61866010, Grant 11871183, Grant 61703370, and Grant 61603338, in part by the Natural Science Foundation of Hainan Province under Grant 118QN181, and in part by the Scientific Research Foundation of Hainan University under Grant kyqd(sk)1804.

ABSTRACT Stock index price forecasting is a consistent focus of business intelligence. Various factors influence stock index price forecasting, such as technical indicators, financial news, business status, and the macroeconomics situation. In addition, many studies have shown that the exchange rate is related to the stock index price; however, no study has examined whether the exchange rate can be used to forecast stock index prices. Therefore, this paper focuses on this topic and uses exchange rate to predict China stock index price for the first time. Firstly, we compare the association of China stock index price with different data sources to illustrate the feasibility of using the exchange rate to predict stock index prices. Then, we generate some additional technical features of the exchange rate and propose a strategy to predict the stock index price. Finally, we compare the forecast results of China's stock index price based on four data sources, i.e., technical indicators, exchange rate data, US market index data and finance news data from January 3, 2017 to March 20, 2019. Experimental results demonstrate that the performance of exchange rate data for stock index prediction is comparable to other popular data sources and that, in some prediction periods, the exchange rate outperforms such data sources. The results confirm that the exchange rate could be used for forecasting the Shanghai Composite Index prices.

INDEX TERMS Stock index prediction, exchange rate, finance news, technical indicators, SZ50, rolling window.

I. INTRODUCTION

Stock market prediction has attracted much attention from both academic and business [1]–[4], because prediction of financial time series is an interesting and difficult problem in business intelligence. Early studies on stock price prediction are based on the efficient market hypothesis (EMH) and random walk theory [5]. There have been many studies [6]–[9] that provide the evidence contrary to what is suggested by the EMH and random walk hypotheses. These studies show that the stock market can be predicted to some degree. In addition, [10]–[12] demonstrate that the stock price change is a highly nonlinear and dynamic process rather than a non-stationary random walk process. Since stock price movement is a complex nonlinear process, in recent years, data mining technology [13] [14] has been used to realize stock price prediction and shows great advantages than traditional technique analysis methods. Using data mining techniques for stock market prediction has become the mainstream approach.

Stock price prediction studies using data mining techniques fall into two broad perspectives. One perspective focuses on the choice of data sources for stock prediction [7], [13], and the other one is constructing different approaches of predicting the stock [14], [15]. The quality of the data determines the prediction performance, and more and more data sources are being used to predict stock prices.

Many factors affect the changes of stock prices [16]. Such factors include the macroeconomic environment, policies, the psychological behavior of the investor, and different factors may involve different data sources. Popular data sources are technical indicators [17], financial news data [18], and fundamental data [19]. Fundamental analysis is a superior method for long-term stability and growth, but it is difficult to predict the market accurately in the short-term using

The associate editor coordinating the review of this manuscript and approving it for publication was Alberto Cano.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

fundamental analysis [20]. In this paper, we focus on short-term forecasting; thus, fundamental data is not considered.

In general, technical indicators are the most common indicators in stock forecasting. 12 technical indicators were used by [17] to predict the direction of daily stock price changes in the Korea Composite Stock Price Index using a support vector machine (SVM) [21], [22]. [23] found that technical indicators had statistically and economically significant out-of-sample forecasting power and frequently outperformed macroeconomic variables in stock prediction. [24] used 10 indicators to predict the price movement of the CNX Nifty and the S&P Bombay Stock Exchange Sensex. The experimental results showed that the performance of all prediction models was improved when technical parameters were represented as trend deterministic data. In fact, technical indicators are commonly used data source for stock price forecasting. [25] used 6 popular indicators and trade volume as input indicators to predict the Shanghai Stock Exchange Composite Index and the Shenzhen Stock Exchange Component Index in the short. [26] used the weighted SVM to predict the turning point of stock price change with 14 technical indicators and combined the relative strength index (RSI). The experimental results showed that the trading income was relatively stable. Different technical indicators have different prediction effects; therefore, constructing and selecting appropriate technical indicators is crucial.

In addition to technical indicators, financial news is also an important data source for stock prediction. Many previous studies investigated textual financial predictions. Such studies primarily focused on classifying price direction [27], [28]. With the development of the Internet, finance related websites and applications constantly provide a large amount of textual data that increasingly affect market participants future price expectations [29]. [30] selected semantically relevant features to predict stock price movement and thus reduced the overfitting problem. [31] used the Financial Times' extensive daily financial news to quantify the relationship between financial market decisions and the development of financial news, and it found that there was a positive correlation between the number of daily quotes of the Financial Times Company and the daily trading volume of the company's stock. [8], [32], [33] used Twitter contents to predict the stock price or stock options pricing. Financial news is gradually becoming more and more important for predicting changes in stock prices.

Exchange rates also affect stock price change, and many scholars have studied the relationship between exchange rates and stock prices. [34] argued that fluctuations in the exchange rate of the local currency would affect a country's balance of payments, international competitiveness, and the country's actual output, and thereby affecting its company's cash flow and stock prices, which in turn would affect the country's stock market. [35] found a significant positive correlation between stock prices and exchange rates. [36] argued that there was also a systematic connection between the exchange rate and the Shanghai Stock Exchange 50 (SZ50). [37] found that there might be other common factors behind exchange rate and stock price changes, such as an increase in a country's real interest rates. Typically, an interest rate increase results in increased capital inflows and appreciation of the local currency. At the same time, the current value of the country's companies future cash flow was reduced, resulting in a decline in stock price. However, no studies have investigated the relationship between stock price movement and exchange rates. Such research is limited by the fact, that the original exchange rate data has only four features, i.e., daily open price, daily close price, daily highest price, and daily lowest price.

In this paper, we focus on predicting stock price movement based on the exchange rate and use the exchange rate to predict China stock price index for the first time. The exchange rate in our research is the benchmark exchange rate for RMB against us dollar. We found that the exchange rate changes have a co-movement relationship with the SZ50 close price through analysis the close price changes between exchange rate and SZ50. In addition, we compare this co-movement relationship to other popular data sources, including SZ50 technical indicators, financial news, and the Dow Jones Industrial Average (DJIA). The results demonstrate that there is a strong correlation between the exchange rate and the SZ50 on the close price. The comparison with the DJIA is provided for fairness. Note that the DJIA data is generated in the same way as the exchange rate data.

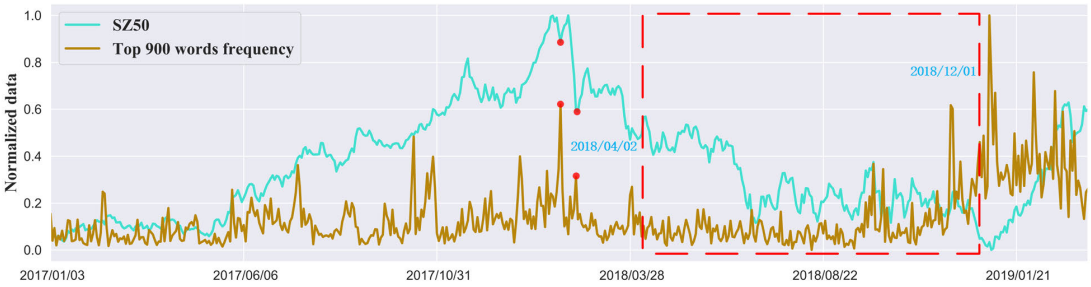
To predict the stock price index using exchange rate data, we generate some additional technical exchange rate indicators and use the generated data to predict SZ50's price movement in the next k-days, where k ranges from {1, 3, 5, 10, 15, 20}. The compared data sources include technical indicators on SZ50, financial news, and the DJIA. After data normalization and feature selection, we train the models and predict the ''up'' and ''down'' of the SZ50 price in the next k-days through a rolling window. The experimental results indicates that the performance of exchange rate data for stock index prediction is comparable to other popular data sources and in some prediction periods, outperforms other popular data sources, which confirms that the exchange rate can be used to predict the Shanghai Composite Index.

The remainder of this paper is organized as follows. In Section II, we explain why the exchange rate can be used for SZ50 prediction. In Section III, we extract the features and generate some additional exchange rate indicators for SZ50 prediction. In Section IV, we propose a data-driven approach using four data sources to predict SZ50 price movement. In Section V, we design experiments using exchange rate data to predict stock price movement and compare exchange rate data to other data sources relative to prediction performance. Conclusions and and recommendations for future work are provided in Section VI.

II. WHY EXCHANGE RATE CAN BE USED FOR SZ50'S PREDICTION?

In this section, the associations of SZ50 and other data sources (DJIA, financial news, and exchange rate) are

(a) SZ50 and DJIA



(b) SZ50 and Financial news

(c) SZ50 and Exchange rate

FIGURE 1. Description of different data sources.

compared to illustrate the feasibility and goodness of predicting stock index price movement using the exchange rate. The respective comparisons of SZ50 to commonly used data sources and to commonly used data sources and to the exchange rate are presented in the following subsections.

A. POPULAR DATA SOURCES AND SZ50

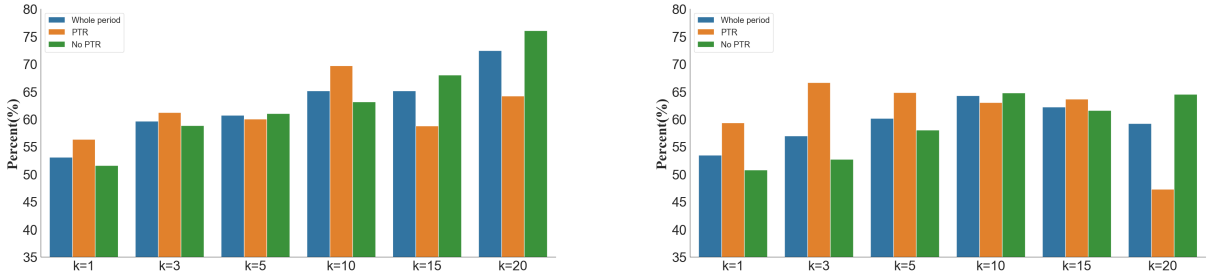
Many factors account for the stock index fluctuations. As mentioned in Section I, technical indicators, financial news, and the DJIA are the popular data sources.

To explain the rationality of using such popular data sources to predict the SZ50 in a simply way, the SZ50 close price is compared to the DJIA close price of and to the top words frequency (extracted from financial news on website[1]). The corresponding results are shown in Figure 1(a) and (b), respectively.

Figure 1(a) and (b) compare the co-movement between the SZ50 and other data sources (DJIA, Financial news) on the close price. The data was collected between January 3, 2017 and March 20, 2019. This period includes 538 trading days. In these figures, the horizontal axis is time (year/month/day) and the vertical axis is the normalized values of the closing price or words frequency. The time period in the red dotted box indicates that the period of the Sino-US trade war (PTR) began on April 1, 2018 and ceased on December 01, 2018.

Figure 1(a) illustrates the co-movement of the SZ50 and DJIA relative to the close price. The trend of DJIA close price is broadly in line with that of the SZ50 from January 2017 to approximately April 2018. We observe that the line of DJIA and SZ50 has the similarity on price changes direction except in the red dotted box which represents for the period of Sino-US trade war (PTR). In the period of red dotted box, the movement trend has been changed. With the intensification of Sino-US trade war, the lines of DJIA and SZ50 diverge dramatically and the previous correlation between two curves seem to be getting ambiguous. Not until

[1]http://www.eastmoney.com/ and http://www.hexun.com/



(a) SZ50 and Exchange rate (b) SZ50 and DJIA

FIGURE 2. The comparisons of different data sources on the direction of price changes in different periods.

the trade war suspended in December 2018, did the tendencies of DJIA and SZ50 become consistent again.

In Figure 1(b), the SZ50 close price is compared to the top 900 words [2] frequency. As can be seen, the top 900 words frequency often peaks when the SZ50 close price reaches a turning point, i.e., sharply declines or increases. Some instances are marked with red dots. Figure 1(a) and (b) indicate that some connections exist between these two data sources and the SZ50.

B. EXCHANGE RATE AND SZ50

Although several studies focus on the relationship between the exchange rate movement and stock index price [35], [37], the exchange rate has rarely been used to predict stock index prices. In this subsection, we show the co-movements of exchange rate and stock price changes. Here, the benchmark exchange rate is the RMB against the US dollar.

Figure 1(c) shows the normalized SZ50 close price along with that of the exchange rate. Note that the original line of the exchange rate has been rotated 180° along the middle yellow line. By and large, SZ50 and exchange rate fluctuations moved in roughly the same direction almost all periods, except during the PTR. During the PTR, the trend lines oscillated violently, given the uncertainty of a trade war.

To illustrate the co-movement of data sources and SZ50 numerically, the rise-and-fall of SZ50 and DJIA close prices and the exchange rate are compared in different periods. For the close price data at time i, its label represents price movements at time $(i+k)$. In other words, if the price at time $(i+k)$ is higher than the price at time i, the label of time i is "+1". Otherwise, the label is "−1". The label of the close price data at time i is defined as follows:

$$Label_k(i) = \text{sgn}\{cp(i+k) - cp(i)\}, \tag{1}$$

where cp is close price, $cp(i)$ is close price at time i, and sgn is the symbolic function.

After labelling the SZ50, DJIA, and exchange rate data, the consistency of the labels of SZ50 and DJIA, SZ50 and exchange rate are compared. Given (i, k), the frequency at which the labels from different data sources are the same is counted, the ratio of that frequency to the total number of the samples in a specific period is calculated. Then, we obtain the histograms in Figure 2, where the horizontal axis is different levels of k, i.e., the time span used to compute labels, and the vertical axis is the percentage (%) of the times that two labels from different data sources have the same (i, k).

From Figure 2(a), it can be seen that as the value of k rises, the proportions of both "Whole period" and "No PTR"(where "No PTR" means that the trade war period is excluded) increase accordingly, and reach their maxima at $k = 20$ (outnumbers 70 and 75, respectively), while the proportion of "PTR" changes irregularly. From Figure 2(b), one can draw an obvious conclusion that the consistency of the SZ50 and exchange rate is better than that of SZ50 and DJIA with regard to "Whole period" and "No PTR" at a relatively high k value.

In addition, after comparison between Figure 2(a) and 2(b), we find that the proportion of consistency on price change directions in Figure 2(a) is higher than that in Figure 2(b) for both "Whole period" and "No PTR". It is evident that the relationship of the close price co-movement between SZ50 and exchange rate is closer than the relationship between SZ50 and DJIA.

The above analysis shows that SZ50 close price has a strong connection with the exchange rate, and their co-movement is even stronger than that of SZ50 and data from some popular sources. Thus, the exchange rate is suitable for predicting the price movement of SZ50.

III. FEATURE GENERATION FOR FORECASTING SZ50

A. EXCHANGE RATE DATA ACQUISITION AND FEATURE GENERATION

In this subsection, we focus on constructing the features to predict SZ50 price movement based on 538 trading days, from January 3, 2017 to March 20, 2019. The original exchange rate data was collected from website [3], where daily open price, daily close price, daily highest price, and daily lowest price are obtained. These four indicators are used as original features.

Since technical analysis is the most common approach [17] to analyze future stock price by studying historical prices, and both stock prices and exchange rates are dynamic price change process, we generate some additional indicators from the original features according to technical analysis methods.

[2] which is ranking according to word frequency

[3] http://quote.eastmoney.com/forex/USDCNY.html.

In particularly, exchange rate price changes in the next k days are analyzed.

Therefore, the features of the exchange rate data include collected indicators and generated indicators. The collected indicators represent the 4 original features extracted from website,[3] and the generated indicators contain 32 features. These features are frequently used to measure whether the currency is in overbought or oversold territory. Detailed descriptions of the features are provided in the following.

1) COLLECTED INDICATORS

Collected indicators consist of daily open price, daily close price, daily highest price, and daily lowest price at time t. The generated indicators are calculated using these collected indicators.

2) GENERATED INDICATORS

Generated indicators include 32 features. First, we calculate the moving average of daily close price at time $t-5$, $t-10$, and $t-20$. Second, we use the indicators in Table 1 to construct the rest of the features, which are in different periods. The moving average method is a simple technology that smooths out price data and filters some noise.

The technical indicators are explained as follows.

- Stochastic %K and %D. Stochastic oscillator lines %K and %D are used together in a technical analysis to determine whether the currency is overbought or oversold. %K and %D values are limited between 0 and 100. %D line values that are greater than 80 indicate that the currency is overbought and values that are less than 20 indicate that it is oversold. %K values are the moving average of D%. Buy and sell signals can also be generated when lines %K and %D intersect: when %K crosses above %D, a buying signal is generated, and when %K crosses below %D a sell signal is generated.
- MACD. Moving Average Convergence Divergence (MACD) is a trend-following momentum indicator that shows the relationship between two moving averages of prices. It is designed to reveal changes in the strength, direction, momentum, and duration of a trend in exchange rate.
- Williams %R. Williams %R is a technical analysis oscillator showing the current closing price in relation to the high and low of the past N days (for a given N). It ranges from 0 and 100. When the value of Williams %R is greater than 80, the currency is in oversold territory. In contrast, the exchange rate price will be ''up'' in the following few days. When the value of Williams %R decreases to 20, the currency is in overbought territory, and the exchange rate price will be ''down'' in the following few days.
- CCI. Commodity Channel Index (CCI) measures variability from the normal range of prices. When the value is between -100 and 100, the CCI does not provide an obvious signal indicating whether to trade. However, when the value is greater than 100, the CCI indicates the end of a rising trend, i.e., exchange rate prices are likely to be ''down'' in a short time. In contrast, a CCI value less than -100 indicates that the currency is in oversold territory and thus, the price may be ''up'' in the future.
- RSI. Relative Strength Index (RSI). The RSI is a momentum indicator. It compares the magnitude of gains and losses over a specified time period. The RSI is used to measure the speed and change of price movements. It is primarily used to identify overbought or oversold conditions in the trading of an asset.
- A/D oscillator. Accumulation/Distribution oscillator (A/D oscillator). The A/D oscillator also follows the price trend, which means that, if the value at time t is greater than that at time $t-1$, the opinion on trend is ''up'' and if value at time t is less than that at time $t-1$, the opinion regarding price trends is ''down''.
- Momentum. Momentum is the rate of acceleration of exchange rate price or volume and refers to the force or speed of movement. Conventionally, momentum is defined as a rate.
- ROC. Rate of Change (ROC). The ROC is the rate at which the exchange rate price changes over a given period. It measures the difference between the price at time t and time $t-1$.
- Disparity-n. Disparity-n refers to the distance between the current price and the moving average over n days. It is often used to predict the turning point. If the value is too large, the end of a rising trend would be anticipated and the price would be expected to move downward.

In [17], [26], most of the above indicators are constructed using technique indicators of exchange rate price and are used as features to predict the movement of the exchange rate price. In this paper, we focus on the short-term prediction of the stock index. The periods of exchange rate data were 5-days, 10-days, and 20-days. The generated indicators we used are described as follows in Table 2, where the exchange rate features comprise 4 collected indicators and 32 generated indicators.

Our labels are determined by the SZ50 close price, and we use data from a previous day to predict the next day's stock price movement. If the price of day t is higher than that of day $t-1$, the label is ''+1''. Otherwise, the label is ''-1''. The label of the i^{th} sample is defined as follows:

$$Label(i) = \text{sgn}(cp(i+1) - cp(i)), \tag{2}$$

where $cp(i)$ is the stock close price at time i, and sgn is the symbolic function. When k in Eq.(1) is equal to 1, the Eq.(2) is the same with Eq.(1).

B. CONTRAST DATA SOURCES FOR FORECASTING SZ50

1) SZ50 TECHNICAL INDICATORS

In this paper, SZ50 technical indicators consist of collected indicators and some generated indicators. Collected indicators include daily open price, daily close price, daily high price, daily low price, and trade volume. The other generated indicators are generated according to Table 1. In addition,

TABLE 1. The constructed technical indicators of exchange rate for stock prediction.

Indicators' name	Formulas(n1<n2)	Description	Periods(n)
%K	$\frac{C_t - LL_{t-n}}{HH_{t-n} - LL_{t-n}} \times 100$	It compares where the exchange rate price closed relative to its price range over a given time period.	5
%D	$\frac{\sum_{i=0}^{n-1} \%K_{t-i}}{n}$	Stochastic %D. Moving average of %K	10
%J	$3 \times K_t - 2 \times D_t$	It is very sensitive to the direction of exchange rate price changes.	5/10
Slow %D	$\frac{\sum_{i=0}^{n-1} \%D_{t-i}}{n}$	Moving average of D%.	3/5
Momentum	$C_t - C_{t-n}$	It measures the amount that the exchange rate price has changed over a given time span.	5
Slow Momentum	$\frac{\sum_{i=0}^{n-1}(C_t - C_{t-n})}{n}$	Moving average of Momentum.	10
ROC	$\frac{C_t}{C_{t-n}} \times 100$	Price rate-of-change. It displays the difference between the current price and the price n days ago.	5/10
Williams' %R	$\frac{H_n - C_t}{H_n - C_n} \times 100$	It measures the amount that the exchange rate price has changed over a given time span.	5/10
A/D Oscillator	$\frac{H_t - C_{t-1}}{H_t - L_t} \times 100$	Accumulation/distribution oscillator. It is a momentum indicator that associates changes in price.	1
Disparity-n	$\frac{C_t}{MA_n} \times 100$	n-day disparity. It means the distance of current price and the moving average of n days.	5/10/20
OSCP	$\frac{MA_{n_1} - MA_{n_2}}{MA_{n_1}}$	Price oscillator. It displays the difference between two moving averages of the exchanege rate price.	(5,10)/ (10,20)
CCI	$\frac{M_t - SM_t}{0.015D_t} \times 100$	Commodity channel index. It measures the variation of the exchange rate price from its statistical mean.	5
RSI	$100\text{-}\frac{100}{1 + (\sum_{i=0}^{n-1} Up_{t-i}/n)/(\sum_{i=0}^{n-1} Dw_{t-i}/n)} \times 100$	Relative strength index. It is a price following an oscillator that ranges from 0 to 100.	5/10/20
DEA	$EMA(C_t, n1) - EMA(C_t, n2)$	It helps to determine if the market will reverse.	10/20
DIF	$EMA(DEA, n)$	Slow moving average of DEA.	5
MACD	$(DIF - DEA) * 2$	Difference between DIF and DEA.	5
BIAS	$(C_t - MA_n)/MA_n$	It measures the difference between the exchange rate price and its moving average during the fluctuation process.	5/10/20
CR	$\frac{\sum_{i=0}^{n-1}(H_t - YM_{t-n})}{\sum_{i=0}^{n-1}(YM_{t-n} - L_{t-n})} \times 100$	Comparison of bullish strength and short power.	5
Slow CR	$\frac{\sum_{i=0}^{n-1} CR_{t-i}}{n}$	Moving average of CR.	5

[a] C_t is the closing price at time t, L_t is the low price at time t, H_t is the high price at time t, MA_t is the moving average of t days, LL_t and HH_t means the lowest and the highest in the last t days; $M_t = (H_t + L_t + C_t)/3$; $SM_t = \sum_{i=1}^{n} M_{t-i+1}/n$; $D_t = \sum_{i=1}^{n} |M_{t-i+1} - SM_t|/n$; $YM_t = \frac{H_t + L_t}{2}$.
[b] EMA(indicators, period) the exponential moving average of the indicators in a period(n days) at time t, the formula is $EMA_t = \alpha * I_t$ $+(1-\alpha) * I_t - 1 + ... + (1-\alpha)^n * I_{t-n}$, where I is for the indicators in EMA(indicators, period), α is the smoothing coefficient.
[c] Up_t means upward-price-change and Dw_t means downward-price-change at time t.

the Chaikin Oscillator is added to clearly describe the volume change. Thus, the technical indicators have 39 features (the added indicators are trade volume, Chaikin Oscillator (5,10) and its moving average) on SZ50.

2) FINANCIAL NEWS DATA

In this paper, the news is extracted from websites.[4] We obtain 13417 news articles, including a title, the post time, and content from these websites. The articles were posted between January 2017 and March 2019. The articles are sorted in date order by the order of the days according to the post time. The China stock market is open for trading from 9:30 am to 11:00 am and 13:00 pm to 15:00 pm. Therefore, news released before 9:30 am is considered as information that was known prior to the opening of the stock market the next day. In addition, the financial news is available on business days and weekends; however, trading day labels and exchange rates are only available on trading days. We added

[4] http://www.eastmoney.com/ and http://www.hexun.com/

TABLE 2. Generated features of exchange rate data.

Data sources	All features				
Exchange rate	Open price	High price	Low price	Close price	MA(5)
	MA(10)	MA(20)	A/D Oscillator	OSCP(5,10)	OSCP(10,20)
	CR(5)	slow CR(5)	DIF(10)	DEA(20)	K%(5)
	MACD(5)	D%(10)	J%(10)	Slow D%(3)	Slow D%(5)
	CCI(5)	ROC(5)	ROC(10)	WR(5)	WR(10)
	BIAS(5)	BIAS(10)	BIAS(20)	Disparity(5)	Disparity(10)
	Disparity(20)	RSI(5)	RSI(10)	RSI(20)	Momentum(5)
	Slow Momentum(10)				

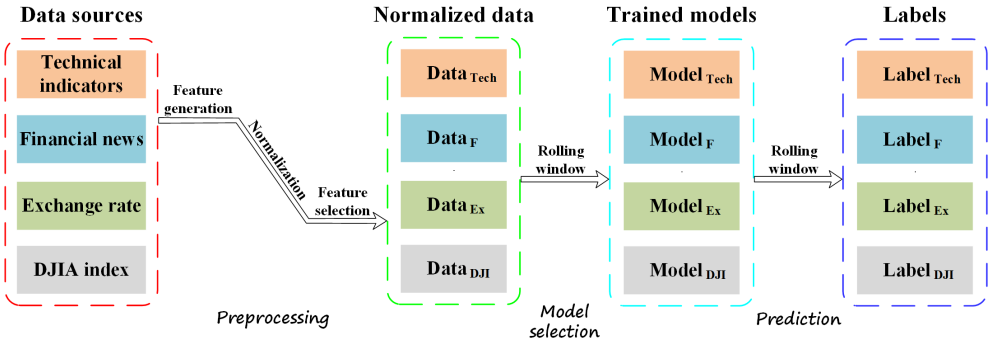


FIGURE 3. Proposed process on stock index prediction with different data sources.

the financial news published on non-trading day to the latest trading day.

We adopt the bag-of-words model [29] to quantize the text news. First, Jieba [5] is used to realize Chinese word segmentation, and 23818 words are obtained after filtering the stop words. Second, 4475 conjunctions, adverbs, attributives, and numbers are eliminate manually. The remaining 19343 words make up a feature dictionary. Finally, we calculate the term frequency-inverse document frequency (TF-IDF) of every word in the daily news as the values of characteristic vectors. In this way, we obtain 538 samples that correspond to 538 trading days.

3) DJIA

The US stock market is highly related to Chinese stock market [38], and usually, their indices move together. Four DJIA features are obtained, i.e., daily open price, daily close price, daily high price, and daily low price. In addition, we generate other technical indicators (in Table 1) using the same period of exchange rate data. After removing some technical indicators that are not available, we obtain 31 technical indicators. The labels are also the same as the exchange rate data which is dependent on SZ50 price changes on trading days.

IV. METHODOLOGY

In this section, we propose a data-driven approach to predict SZ50 price movement using four data sources. A flow chart of the approach is shown in Figure 3. Four data sources are represented by four data sources with four different colors. First, feature generation, normalization, and feature selection are performed to obtain the proper preprocessed data. Second, a rolling window technique is used to select the best model parameters on the normalized data. Third, the test data is predicted using the rolling window with the best parameters, and then the test data labels are obtained as output. The technical details are presented as follows.

A. FEATURE SELECTION

As mentioned above, feature selection is required in order to predict the SZ50 stock index. Here, we adopt two strategies to select features for four data sources.

- Feature selection for technical indicators, DJIA, and exchange rate
  A recursive strategy is used to select original features on these three data sources. The recursive strategy deletes one feature from the original features in a loop, and the deleted feature represents the highest accuracy (trained model without the deleted feature). Once the accuracy has decreased significantly (about 3%) from the best accuracy, the process is terminated. The selected features are shown in Table 3, where the numbers in parentheses indicate the period of the selected indicators. The numbers of features selected for (of technical indicators, exchange rate, and the DJIA) are 34, 34, and 26, respectively. From Table 3, it is evident that the features of the three data sources do not differ significantly relative to both quantity and category.
- Feature selection for financial news data
  First, the top 900 words (ranked by word frequency) are selected as the characteristic vectors of each text. Then, the TF-IDF of these words is calculated for the values of characteristic vectors.

[5] https://github.com/fxsjy/jieba

TABLE 3. Selected features of different data sources.

Data sources	Selected features				
Technical indicators	Open price	High price	Low price	Close price	MA(10)
	MA(20)	OSCP(5,10)	OSCP(10,20)	Volumn	CR(5)
	Slow CR(5)	DIF(10)	DEA(20)	MACD(5)	K%(5)
	D%(10)	J%(10)	Slow D%(3)	Slow D%(5)	CCI(5)
	ROC(5)	ROC(10)	WR(5)	WR(10)	BIAS(20)
	Disparity(5)	Disparity(10)	Disparity(20)	RSI(5)	RSI(10)
	RSI(20)	Momentum(5)	Slow Momentum(10)	Chaikin Oscillator(5,10)	
Exchange rate	Open price	High price	Low price	Close price	MA(5)
	MA(10)	MA(20)	A/D Oscillator	OSCP(5,10)	OSCP(10,20)
	CR(5)	Slow CR(5)	DIF(10)	DEA(20)	MACD(5)
	K%(5)	D%(10)	J%(10)	Slow D%(3)	CCI(5)
	ROC(5)	ROC(10)	WR(5)	WR(10)	BIAS(5)
	BIAS(10)	BIAS(20)	Disparity(5)	Disparity(10)	Disparity(20)
	RSI(5)	RSI(10)	Momentum(5)	Slow Momentum(10)	
DJIA	Open price	High price	Low price	Close price	A/D Oscillator
	MA(5)	MA(20)	OSCP(5,10)	OSCP(10,20)	Volumn
	DEA(20)	MACD(5)	K%(5)	D%(10)	J%(10)
	Slow D%(3)	Slow D%(5)	ROC(5)	ROC(10)	WR(5)
	WR(10)	BIAS(5)	BIAS(10)	BIAS(20)	Disparity(5)
	Disparity(10)	Disparity(20)			

$ChaikinOscillator = SMA(MID_t, 10) - SMA(MID_t, 5)$, where $MID_t = sum(volumn \cdot \frac{2 \cdot C_t - H_t - L_t}{H_t - L_t})$.

B. METHOD

As a successful classification tool, SVM can cope with nonlinear prediction well by introducing kernels. Stock price movement is a highly nonlinear process, so SVM is suitable to deal with stock prediction problem. In fact, SVM has been applied to time series data predictions, including stock price, such as in [7] [17]. Therefore, in this paper, we employ SVM to perform stock price movement prediction.

In the following, we briefly describe how SVM is used. Here, SVM is implemented to handle large margin linear classification by mapping the input data into the high-dimensional feature space. Given a training set with label pairs (x_i, y_i), $i = 1, ..., m$, where $x_i \in R^n$ and, $y_i \in \{+1, -1\}$, SVM first maps a sample x into $\phi(x)$ and then constructs a pair of support hyperplanes $w^\top \phi(x) + b = -1$ and $w^\top \phi(x) + b = 1$ in mapped space, and requires the different classes on both sides of the support hyperplanes as far as possible. At the same time, SVM maximizes the margin of these two support hyperplanes. In particular, it requires solution of the following optimization problem.

$$\begin{aligned} &\min_{w,b,\xi} \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\ &\text{s.t. } y_i(w^\top \phi(x_i) + b) \geq 1 - \xi_i, i = 1, 2 \ldots m \\ &\xi_i \geq 0, \quad i = 1, 2 \ldots m \end{aligned} \tag{3}$$

Here, C is a positive parameter. By introducing kernel $K(x_i, x_j) = (\phi(x_i) \cdot \phi(x_j))$ and solving the dual problem in Eq.(3), we could obtain its solutions. Then, the decision function is denoted as follows

$$y = \text{sgn}(f(x)) \tag{4}$$

where

$$f(x) = \sum_{i=1}^{m} y_i \alpha_i K(x_i, x) + b \tag{5}$$

and α_i is the dual variable.

We can see that only parameter C and the kernel parameter need to be regular. Therefore, SVM has fewer parameters compared to other prediction approaches. In addition, SVM can avoid the dimension disaster problem on stock prediction due to the existence of kernel function. Here, the radial basis function is used as the kernel function. Optimal parameter values are determined by grid search. We set the parameter C from $[5^{-8}, 5^{-7}, 5^{-6} \ldots 5^{6}]$, and the kernel parameter from $[2^{-8}, 2^{-7}, 2^{-6} \ldots 2^{6}]$.

C. ROLLING WINDOW

For time series problems, a common way for training a model is traditional k-fold cross-validation approach, but the approach will destroy the temporal information, because the k-fold cross-validation will break the order of stock price time series. Apart from k-fold cross-validation, rolling window [39] is also a common way for training classifiers in stock prediction, and it is designed for time series problems. Therefore, we combine rolling window and validation technique to select models. Specifically, the entire dataset is divided into three parts based on the time-line, which is divided into training data, validation data, and test data, e.g., three datasets in Table 4. Training data is used to train the models, validation data is used to select the parameters and prevent overfitting, and the test data is used to test the model. The rolling window technique is used both in training models and predicting the test data. Assume the window step is k, then the training data, validation data, and test data move forward k samples at a time. In this way, we can predict k samples with the latest

TABLE 4. Three datasets with different proportion of training data and validation data.

Dataset	Trainng data	Validation data	Test data
Dataset I	01/03/2017 - 09/14/2018	09/17/2018 - 12/17/2018	12/18/2018 - 03/20/2019
Dataset II	01/03/2017 - 08/03/2018	08/06/2018 - 12/17/2018	12/18/2018 - 03/20/2019
Dataset III	01/03/2017 - 06/22/2018	06/25/2018 - 12/17/2018	12/18/2018 - 03/20/2019

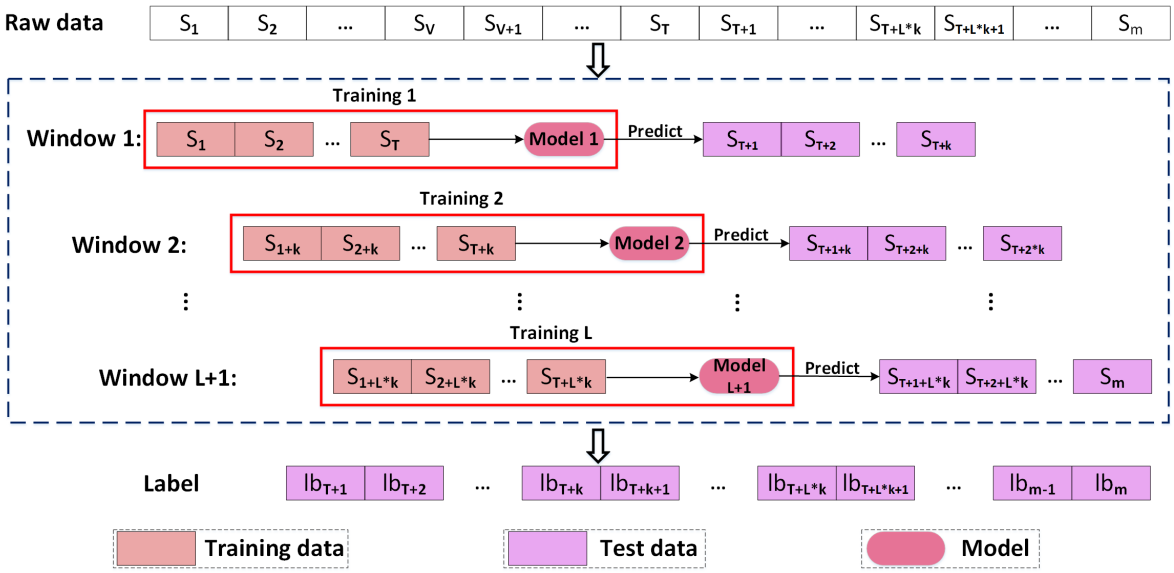


FIGURE 4. The process of rolling window with validation technique.

training data once at a time. We keep moving the training data, validation data, and test data until all test data have been predicted.

The rolling window process with a validation technique is shown in Figure 4. Given the raw data with m samples, $S_1, S_2, \ldots, S_m$, we use the rolling window technique to divide the samples into $[(n - T_2)/k + 1]$ ($[\cdot]$ denotes the rounding down operator) windows in chronological order, each window has k test samples. In each window, data is divided into two parts, that is, data for train (including training data and validation data) and data for test. For example, different color modules present different data sets or models, and the solid red box represents the training process. Models are trained by using both training data and validation data. Then, the test data in each window is predicted in each window until all test data in all windows has been predicted. In this way, we can obtain the labels of all test data. In order to explain the process of training a model in detail, we present the model training process separately, as shown in Figure 5.

Figure 5 illustrates the training process in window 1. As shown in Figure 5, the data for training the model is $S_1, S_2, \ldots, S_T$. We use $S_V, S_{V+1}, \ldots S_T$ as the validation data and divide the training process into $q + 1$ subprocesses in chronological order. For each subprocess, V samples are used to predict the next k samples in chronological order and these V samples move forward k steps in the next subprocess. The process terminates when all validation data have been predicted for one time. In this way, we can obtain the labels of validation data. Then we can obtain the best parameters using a grid search technique. Finally, the most recent V samples $S_{T-V+1}, \ldots, S_T$ are using as the training data to train the model 1 with the best parameters. Other rolling windows are similar with window 1. The remaining test data is predicted using a similar process, as shown in Figure 4 and Figure 5, respectively.

In addition, [39] implies that the length of windows step k is flexible. In fact, for stock price prediction, the length of windows step k differs. Researchers set different window step k, and the results are different [40]. Therefore, we set the values of k to $\{1, 3, 5, 10, 15, 20\}$ to observe the influence of k by setting different k values in the rolling window process.

V. EXPERIMENTS

A. FEATURE SELECTION RESULTS

In this subsection, we show the results of the recursive strategy for feature selection as described above.

The results of feature selection on different data sources for stock index prediction are shown in Figure 6, where the window step size is $k = 3$. The horizontal axis represents different data sources, and the vertical axis represents the prediction accuracy (%). The blue histogram represents prediction accuracy without feature selection, and the orange histogram represents prediction accuracy after feature selection. It is evident that prediction accuracy after feature selection is at least 10% higher than that of before feature selection. Thus, feature selection prior to predicting the price movement of the stock index is useful.

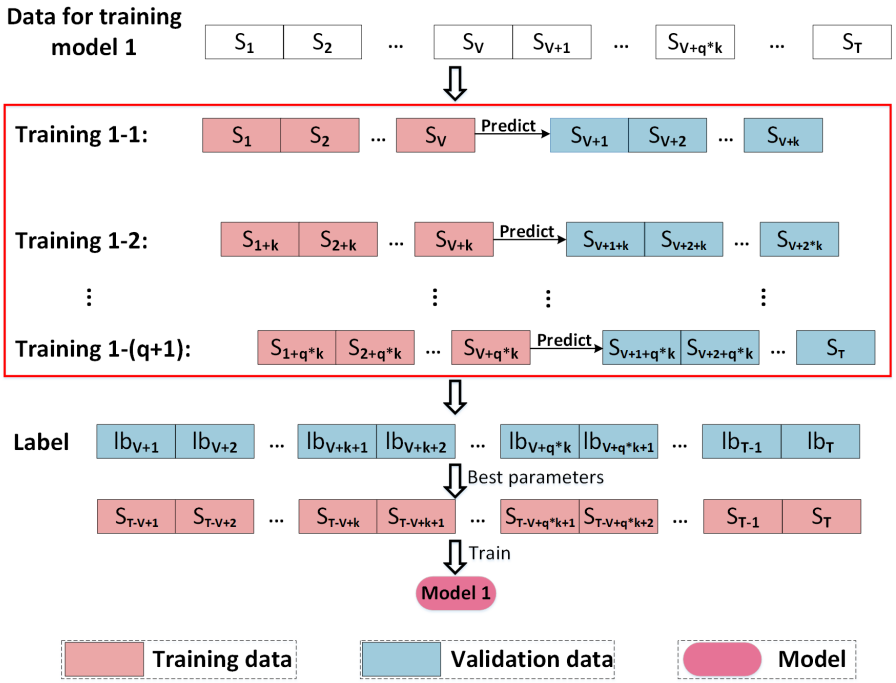


FIGURE 5. The details of training model 1.

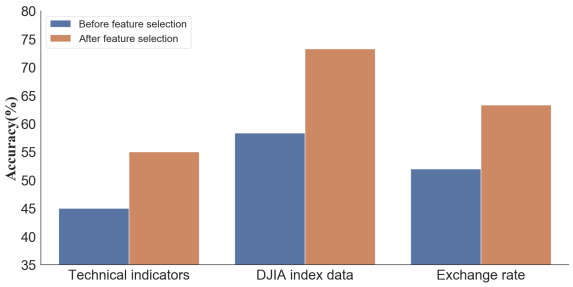


FIGURE 6. The results of feature selection on different data sources.

B. RESULTS OF COMPARING DIFFERENT DATA SOURCES

In this subsection, we show the prediction results on the price movement with different data sources at different window steps in different time periods. In the experiments, three datasets are constructed with period. In these three datasets, test data is the same, and the proportion of training data to validation data is set to {1:1, 1:1.5, 1:2}, respectively. Accuracy and Gmean are used to evaluate the performance, and the parameters are selected by the grid method.

The results are compared in Table 5. In Table 5, bold numbers represent the highest accuracy in horizontal contrast, and underlined numbers represent the highest accuracy in vertical contrast. For every dataset, we selected different window steps k to illustrate its influence. From Table 5, we can see that i) The window step k strongly influences prediction performance. The accuracy is approximately 7% floating up and down as changes. For technical indicators and the DJIA data, the most appropriate k is 3 or 5, and $k = 1$ is not the best choice for price prediction. This may be because the technical analysis method is more suitable for analyzing the trend over time. The exchange rate data gets better classification results at the most window step k, and it means the change of exchange rate has a great influence on the stock market. ii) The proper proportions of training data and validation data are different in different datasets, and the performance on Dataset I is obviously better than the performance on other datasets. iii) The same step window k and the proportion of training data and validation data with different data sources greatly influence prediction performance. iv) From the average accuracy of different data sources with different window steps, we can see that the average accuracy of exchange rate is higher than other data sources in most datasets.

To further compare the performance of different data sources, a two-sample t-test is done for the Accuracy and Gmean indicators (Table 5). The results are shown in Table 6. In Table 6, "Ex" indicates exchange rate data, and "Ex-Technical indicators" represents the comparison between exchange rate data and technical indicators for the two-sample t-test results. By comparing t-test values, we also compute a comprehensive metric Win-Tie-Loss (W-T-L) on accuracy to characterize the relative performance, which denotes the number of metrics where exchange rate data is significantly superior/equal/inferior compared to other data sources. The corresponding W-T-L values are listed at the bottom of Table 6. From Table 6, it can be seen that the performance of exchange rate data is statistically better than other data sources on most cases, as evidenced by the indicators of "Win" at 95% confidence level. Thus, it can be concluded that using exchange rate data to predict the price movement of SZ50 is at least comparable to using other popular data sources, and in some cases, exchange rate data outperforms other data sources.

TABLE 5. The experiment results of different data sources (Accuracy(%)/Gmean(%)).

	Data sources	k=1	k=3	k=5	k=10	k=15	k=20	Mean
Dataset I	Technical indicators	41.67/41.03	**55.00/54.64**	53.33/51.90	43.33/31.43	48.33/36.70	45.00/35.14	47.78/41.81
	Financial news data	56.67/56.95	53.33/49.69	**61.67/59.18**	55.00/53.18	61.67/51.03	56.67/51.90	57.51/53.49
	Exchange rate	61.67/60.12	**73.33/72.47**	63.33/61.41	61.67/60.12	61.67/60.12	56.67/56.85	63.06/61.85
	DJIA data	48.33/41.57	**63.33/51.90**	53.33/49.69	48.33/48.55	53.33/52.65	50.00/50.14	52.78/49.08
Dataset II	Technical indicators	45.00/45.20	**53.33/51.90**	53.33/49.69	51.67/49.69	45.00/37.61	48.33/47.85	49.44/46.99
	Financial news data	51.67/48.55	**63.33/59.18**	56.67/53.29	56.67/53.29	55.00/55.25	58.33/52.97	56.95/53.76
	Exchange rate	53.33/50.92	53.33/49.69	53.33/50.48	**53.33/51.85**	47.78/48.85	50.00/47.33	51.85/48.85
	DJIA data	50.00/50.14	51.67/51.85	48.33/47.85	55.00/45.20	**55.00/52.12**	43.33/41.44	50.56/48.10
Dataset III	Technical indicators	50.00/49.80	50.00/49.80	**55.00/54.64**	46.67/46.78	41.67/40.20	50.00/44.44	48.89/47.61
	Financial news data	50.00/44.44	48.33/48.55	50.00/47.38	**58.33/52.97**	53.33/53.60	53.33/53.60	52.22/50.09
	Exchange rate	53.33/52.65	55.00/55.25	**60.00/56.85**	58.33/54.43	56.67/54.43	53.33/49.69	56.11/53.88
	DJIA data	51.67/48.55	**53.33/52.65**	50.00/46.06	51.67/50.59	50.00/49.80	**53.33/52.65**	51.67/50.05

TABLE 6. Significant difference test for prediction effects based on different data sources ($\alpha = 0.05$).

	Ex-Technical indicators	Ex-Financial news	Ex-DJIA
Accuracy(t-value/p-value)			
Dataset I	4.8298/0.0007	2.0898/0.0632	3.1912/0.0096
Dataset II	1.2886/0.2266	0.2266/0.0206	0.6294/0.5431
Dataset III	3.4017/0.0068	2.1112/0.0609	3.5072/0.0057
W-T-L	2-1-0	0-2-1	2-1-0
Gmean(t-value/p-value)			
Dataset I	4.5104/0.0011	0.0119/0.0056	4.6525/0.0009
Dataset II	1.3090/0.2198	-2.5059/0.0311	0.9597/0.3598
Dataset III	2.7596/0.02014	2.0282/0.0700	2.6508/0.0242
W-T-L	2-1-0	1-1-1	2-0-1

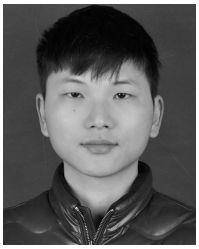
VI. CONCLUSION

In this paper, we use the exchange rate to forecast the price movement of the SZ50. For the exchange rate, the basic indicators are used to generate some additional indicators, and the empirical results prove the feasibility of generating features. Experiments show that the performance of exchange rate data is better than that of other data sources in most situations. Our datasets, code, and prediction tool have been uploaded to https://github.com/LeeRiking/CanExsp. All in all, the stock index price movement forecasting is a complicated problem that is influenced by many factors. Although we demonstrate that the exchange rate could be used to forecast the price movement of the SZ50, in future it would be interesting to introduce more factors and do additional feature construction and selection.

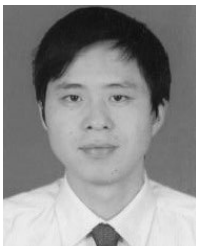
REFERENCES

[1] R. K. Nayak, D. Mishra, and A. K. Rath, "A naive SVM-KNN based stock market trend reversal analysis for Indian benchmark indices," *Appl. Soft Comput.*, vol. 35, pp. 670–680, Oct. 2015.
[2] L. Chen, Z. Qiao, M. Wang, C. Wang, R. Du, and H. E. Stanley, "Which artificial intelligence algorithm better predicts the Chinese stock market?" *IEEE Access*, vol. 6, pp. 48625–48633, Jul. 2018.
[3] N. Kanungsukkasem and T. Leelanupab, "Financial latent Dirichlet allocation (FinLDA): Feature extraction in text and data mining for financial time series prediction," *IEEE Access*, vol. 7, pp. 71645–71664, 2019.
[4] A. Picasso, S. Merello, Y. K. Ma, L. Oneto, and E. Cambria, "Technical analysis and sentiment embeddings for market trend prediction," *Expert Syst. Appl.*, vol. 135, pp. 60–70, Nov. 2019.
[5] E. F. Fama, "Efficient capital markets: A review of theory and empirical work," *J. Finance*, vol. 25, no. 2, pp. 383–417, May 1970.
[6] B. G. Malkiel, "The efficient market hypothesis and its critics," *J. Econ. Perspect.*, vol. 17, no. 1, pp. 59–82, 2003.
[7] B. Weng, M. A. Ahmed, and F. M. Megahed, "Stock market one-day ahead movement prediction using disparate data sources," *Expert Syst. Appl.*, vol. 79, pp. 153–163, 2017.
[8] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, Mar. 2011.
[9] J. R. Nofsinger, "Social mood and financial economics," *J. Behav. Finance*, vol. 6, no. 3, pp. 144–160, 2005.
[10] M. Frank and T. Stengos, "Measuring the strangeness of gold and silver rates of return," *Rev. Econ. Stud.*, vol. 56, no. 4, pp. 553–567, 1989.
[11] S. C. Blank, "'chaos' in futures markets? A nonlinear dynamical analysis," *J. Futures Markets*, vol. 11, no. 6, p. 711, 1991.
[12] G. P. Decoster, W. C. Labys, and D. W. Mitchell, "Evidence of chaos in commodity futures prices," *J. Futures Markets*, vol. 12, no. 3, p. 291, 1992.
[13] H. W. Wang, S. Lu, and J. C. Zhao, "Aggregating multiple types of complex data in stock market prediction: A model-independent framework," *Knowl.-Based Syst.*, vol. 164, pp. 193–204, Jan. 2019.
[14] J. L. Ticknor, "A Bayesian regularized artificial neural network for stock market forecasting," *Expert Syst. Appl.*, vol. 40, no. 14, pp. 5501–5506, 2013.
[15] X. Ding, Y. Zhang, T. Liu, and J. W. Duan, "Deep learning for event-driven stock prediction," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015.
[16] R. Dash and P. K. Dash, "A hybrid stock trading framework integrating technical analysis with machine learning techniques," *J. Finance Data Sci.*, vol. 2, no. 1, pp. 42–57, 2016.
[17] K. J. Kim, "Financial time series forecasting using support vector machines," *Neurocomputing*, vol. 55, nos. 1–2, pp. 307–319, Sep. 2003.
[18] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news: The AZFin text system," *ACM Trans. Inf. Syst.*, vol. 27, no. 2, p. 12, 2009.
[19] P. M. Dechow, A. P. Hutton, L. Meulbroek, and R. G. Sloan, "Short-sellers, fundamental analysis, and stock returns," *J. Financial Econ.*, vol. 61, no. 1, pp. 77–106, 2001.

[20] J. G. Agrawal, V. S. Chourasia, and A. K. Mittra, "State-of-the-art in stock prediction techniques," *Int. J. Adv. Res. Electr., Electron. Instrum. Eng.*, vol. 2, no. 4, pp. 1360–1366, 2013.
[21] V. Vapnik and V. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998, pp. 156–160.
[22] C. N. Li, Y. H. Shao, and H. Wang, "Single versus union: Non-parallel support vector machine frameworks," 2019, *arXiv:1910.09734*. [Online]. Available: https://arxiv.org/abs/1910.09734
[23] C. J. Neely, D. E. Rapach, J. Tu, and G. Zhou, "Forecasting the equity risk premium: The role of technical indicators," *Manage. Sci.*, vol. 60, no. 7, pp. 1772–1791, 2014.
[24] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques," *Expert Syst. Appl.*, vol. 42, no. 1, pp. 259–268, 2015.
[25] Y. Chen and Y. Hao, "A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction," *Expert Syst. Appl.*, vol. 80, pp. 340–355, Sep. 2017.
[26] H. M. Tang, P. W. Dong, and Y. Shi, "A new approach of integrating piecewise linear representation and weighted support vector machine for forecasting stock turning points," *Appl. Soft Comput.*, vol. 78, pp. 685–696, 2019.
[27] V. Cho, B. Wüthrich, and J Zhang, "Text processing for classification," *J. Comput. Intell. Finance*, vol. 7, no. 2, pp. 6–22, 1999.
[28] M. A. Mittermayer, "Forecasting intraday stock price trends with text mining techniques," in *Proc. 37th Annu. Hawaii Int. Conf. Syst. Sci.*, 2004, p. 10.
[29] Y. Shynkevich, T. M. McGinnity, S. A. Coleman, and A. Belatreche, "Forecasting movements of health-care stock prices based on different categories of news articles using multiple kernel learning," *Decis. Support Syst.*, vol. 85, pp. 74–83, May 2016.
[30] M. Hagenau, M. Liebmann, and D. Neumann, "Automated news reading: Stock price prediction based on financial news using context-capturing features," *Decis. Support Syst.*, vol. 55, no. 3, pp. 685–697, Jun. 2013.
[31] M. Alanyali, H. S. Moat, and T. Preis, "Quantifying the relationship between financial news and the stock market," *Sci. Rep.*, vol. 3, p. 3578, Dec. 2013.
[32] J. Si, A. Mukherjee, B. Liu, Q. Li, H. Li, and X. Deng, "Exploiting topic based Twitter sentiment for stock prediction," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2013, pp. 24–29.
[33] W. Wei, Y. X. Mao, and B. Wang, "Twitter volume spikes and stock options pricing," *Comput. Commun.*, vol. 73, pp. 271–281, Jan. 2016.
[34] R. Dornbusch and S. Fischer, "Exchange rates and the current account," *Amer. Econ. Rev.*, vol. 70, no. 5, pp. 960–971, 1980.
[35] R. Aggarwal, "Exchange rates and stock prices: A study of the us capital markets under floating exchange rates," *Akron Bus. Econ.*, vol. 12, pp. 7–12, 2003.
[36] A. Sensoy and B. M. Tabak, "Dynamic efficiency of stock markets and exchange rates," *Int. Rev. Financial Anal.*, vol. 47, pp. 353–371, Oct. 2016.
[37] Y. Wu, "Stock prices and exchange rates in VEC model—The case of Singapore in the 1990s," *J. Econ. Finance*, vol. 24, no. 3, pp. 260–274, 2000.
[38] B. Zhang and X. M. Li, "Has there been any change in the comovement between the Chinese and US stock markets?" *Int. Rev. Econ. Finance*, vol. 29, pp. 525–536, Jan. 2014.
[39] F. Ma, Y. Wei, and D. Huang, "Multifractal detrended cross-correlation analysis between the Chinese stock market and surrounding stock markets," *Phys. A, Stat. Mech. Appl.*, vol. 392, no. 7, pp. 1659–1670, 2013.
[40] R. Ren, D. D. Wu, and T. Liu, "Forecasting stock market movement direction using sentiment analysis and support vector machine," *IEEE Syst. J.*, vol. 13, no. 1, pp. 760–770, Mar. 2018.



JUN ZHANG received the bachelor's degree from the College of Civil Engineering and Architecture, Hainan University, China, in 2015. He is currently pursuing the degree with the School of Economics, Hainan University. He is also a member of the OPTIMAL Group, led by Y.-H. Shao. His research interests include machine learning, data mining, and stock prediction.



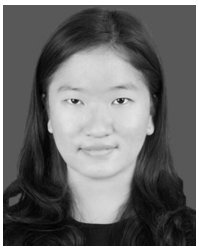
YUAN-HAI SHAO received the master's degree in information and computing science from the College of Mathematics, Jilin University, and the master's degree in applied mathematics and the Ph.D. degree in operations research and management from the College of Science, China Agricultural University, China, in 2006, 2008, and 2011, respectively. He is currently a Professor with the School of Management, Hainan University. His research interests include data mining, machine learning, and optimization methods. He has published more than 80 refereed articles on these areas.

LING-WEI HUANG received the bachelor's degree in mathematics and applied mathematics from the College of Mathematics and Econometrics, Hunan University, in 2015, where he is currently pursuing the degree with the School of Economics. His research interests include data mining, machine learning, and optimization methods.

JIA-YING TENG received the bachelor's degree in finance from Shandong University, China, in 2017. She is currently pursuing the degree with the School of Economics, Hainan University, led by Y.-H. Shao. Her research interests include machine learning, data mining, and quantitative investment.

YU-TING ZHAO received the bachelor's degree from the School of Mathematical Sciences, Xiamen University, China, in 2017. She is currently pursuing the degree with the School of Economics, Hainan University. Her research interests include data mining and machine learning.

ZHU-KAI YANG has been in economic statistics with the School of Economics, Hainan University since 2017. He is one of the undergraduate assistants of the I Do Lab, and focuses on data preprocessing and outlier detection. He devotes himself on finding patterns in the data and exploring them.



XIN-YANG LI has been in economics statistics with the Economics School, Hainan University, since 2017. She is one of the undergraduate assistants of the I DO Lab, and focuses on missing data filling and outlier detection in big data preprocessing. She researches on algorithms and programming in business data analysis.