Forecasting Stock Market Movement Direction Using Sentiment Analysis and Support Vector Machine

Rui Ren, *Member, IEEE*, Desheng Dash Wu, *Senior Member, IEEE*, and Tianxiang Liu

***Abstract*—Investor sentiment plays an important role on the stock market. User-generated textual content on the Internet provides a precious source to reflect investor psychology and predicts stock prices as a complement to stock market data. This paper integrates sentiment analysis into a machine learning method based on support vector machine. Furthermore, we take the day-of-week effect into consideration and construct more reliable and realistic sentiment indexes. Empirical results illustrate that the accuracy of forecasting the movement direction of the SSE 50 Index can be as high as 89.93% with a rise of 18.6% after introducing sentiment variables. And, meanwhile, our model helps investors make wiser decisions. These findings also imply that sentiment probably contains precious information about the asset fundamental values and can be regarded as one of the leading indicators of the stock market.**

***Index Terms*—Day-of-week effect, decision making, sentiment analysis, stock markets, text mining.**

I. Introduction

Forecasting stock market trends has been treated as one of the most challenging but important tasks. Stock market is a nonlinear and dynamic system, and investor sentiment constitutes a key factor of the financial market [1]. With the proliferation of news, blogs, forums, and social networking websites, textual content on the Internet provides a precious source to reflect investor sentiment and predicts stock prices as a complement to traditional stock market time series data. Hence an automated approach is required to distill knowledge from a large number of textual documents [2], [3]. Sentiment analysis is used to automatically extract views, attitudes, and emotions from the opinionated contents [4]. So, we employ sentiment analysis to construct sentiment indexes, and then aggregate them with stock market data to forecast movement direction.

In order to get an efficient and persuasive sentiment index, we take the day-of-week effect into consideration, which means that the average return on Mondays is much lower than that on the other days of the week [5]. It is one of the most well-known financial anomalies dating back to 1930 when Fred C. Kelly revealed the phenomenon on the U.S. markets where the returns had the tendency to decline on Mondays [6], [7]. Then, the effect is proved to exist in global stock markets [8]. The reasons probably include that a much larger amount of information is produced on weekends than weekdays. Most of the corporations tend to release news on Saturday, Sunday, or even on Friday nights just after the stock market is closed, as people have enough time to digest the bad news to prevent the dramatic fluctuations or remember the good news to boost companies' images. However, the day-of-week effect is seldom mentioned when it comes to calculating sentiment indexes.

Another difficulty in predicting stock movement direction is attributed to its nonlinear, dynamic, and evolutionary properties. Support vector machine (SVM) has been widely utilized since it can solve the nonlinear problem by converting it to a quadratic programming. Moreover, the solution of SVM is unique and globally optimal [9]. It can also reduce the overfitting problem by selecting the maximal margin hyperplane in the feature space [10]. To further address the problem, we implement five-fold cross validation. However, it leads to look-ahead bias, so we integrate SVM with a realistic rolling window approach to eliminate the bias. Empirical results illustrate that combining sentiment features with stock market data outperforms using only stock market data in forecasting movement direction. So, we can deduce that investor sentiment plays an important role on the stock market. Furthermore, sentiment probably contains precious information about the asset fundamental values and can be regarded as one of the leading indicators of the stock market.

Moreover, we have developed a practical trading strategy. The prediction results also imply trade order, 1 means buy order, whereas –1 means sell order. Thus, we can simulate how it behaves if people make investment decisions solely based on the results in a real market environment. We assume that short selling mechanism is allowed and there are no market frictions. The results present that by integrating sentiment indexes to the basic model, the investors can make more profit and at the same time bear fewer risks. In addition, a stop-loss order strategy is applied to limit the potential losses, and it accomplishes a much better performance.

Manuscript received June 9, 2017; revised September 20, 2017 and December 27, 2017; accepted January 4, 2018. This work was supported in part by the Ministry of Science and Technology of China under Grant 2016YFC0503606, in part by the National Natural Science Foundation of China under Grant 71471055 and Grant 91546102, in part by the Chinese Academy of Sciences (CAS) Frontier Scientific Research Key Project under Grant QYZDB-SSW-SYS021, and in part by the CAS Strategic Research and Decision Support System Development under Grant GHJ-ZLZX-2017-36. *(Corresponding author: Desheng Dash Wu.)*

R. Ren and T. Liu are with the School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: renrui115@mails.ucas.ac.cn; 201518009443015@mails.ucas.ac.cn).

D. Wu is with the School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China, and also with the Stockholm Business School, Stockholm University, Stockholm SE-106 91, Sweden (e-mail: dash@risklab.ca).

Digital Object Identifier 10.1109/JSYST.2018.2794462

The remainder of this paper is organized as follows. Section II highlights related literature. Section III puts forward a new sentiment analysis method and describes SVMs in detail. In Section IV, we describe data and present empirical results. Section V provides concluding remarks and future work.

II. Related Work

A. Sentiment Analysis in Finance Industry

Sentiment is an opinion or feeling you have about something according to the Longman Dictionary. Sentiment analysis is the method to transfer unstructured textual contents to structured data, and distill views, attitudes, and emotions by language processing, data mining, and computational linguistics [11], [12]. Investor sentiment constitutes a key factor of the financial market [1]. Baker and Wurgler [13] employ the equity share in new issues, the dividend premium and some other variables as sentiment proxies, and point out that investor sentiment affects the cross section of stock returns. Edmans *et al.* [14] use international soccer results as a mood variable and document a significant market decline after each loss. Afterwards, with the development of the sentiment analysis, researchers start to deal with written text that is a more direct way to express ideas and emotions. Tetlock [15] generates a pessimistic media factor in terms of the Wall Street Journal's "Abreast of the Market" column and finds that high pessimism has a negative effect on market prices followed by a subsequent reversion. Bollen *et al.* [16] present evidence that tweets posted on Twitter are a predictive factor of the Dow Jones Industrial Average (DJIA) values and find an accuracy of 86.7% in predicting the daily up and down changes in the closing values of the DJIA. Gillam *et al.* [17] concentrate on the volume of news to quantify the information incorporated in textual data and discover that it enhances earnings forecasting. Moreover, sentiment analysis is superior to the bag-of-words model at individual stock, sector, and index levels in predicting stock prices [12]. Oliveira *et al.* [18] propose an automated method to build a stock market sentiment lexicon to facilitate the research in the area. Nevertheless, the day-of-week effect is rarely mentioned in the study of investor sentiment.

B. SVM in Predicting Stock Market

SVM was proposed by Vapnik [19] and is a supervised learning method that can partially address the overfitting problem directly and formally [20]. With the help of kernel functions, such as radical basis function (RBF) kernel and polynomial kernel, it is able to solve the nonlinear problem by projecting it onto the high-dimensional feature space. Furthermore, it is a dynamic approach. Stock market is a nonlinear and dynamic system, so SVM has been widely applied to forecast stock prices, especially stock indexes. Huang *et al.* [9] employ SVM to predict the weekly movement direction of NIKKEI 225 Index and show that SVM outperforms the other classification methods, such as random walk model, quadratic discriminant analysis, and Elman backpropagation neural networks. An evolving least squares SVM is proposed by Yu *et al.* [10] to explore the trends of three important stock indexes, S&P 500 Index, DJIA Index, and New York Stock Exchange Index. Kara *et al.* [21] use SVM to forecast daily Istanbul Stock Exchange National 100 Index, and the average prediction performance is 71.52%. Besides, SVM is combined with other methods to achieve a better performance. A model by integrating SVM with several other classification methods, such as random walk model and Elman backpropagation neural networks, performs best in predicting NIKKEI 225 Index [9]. Pai and Lin [22] put forward a hybrid of an autoregressive integrated moving average model and an SVM model in forecasting stock prices, and the experimental results prove promising. Wu *et al.* [23] propose a method that integrates sentiment analysis into SVM and generalized autoregressive conditional heteroskedasticity to explore the relationship between stock price volatility and stock forum sentiment, and the method can effectively predict financial risk measured in volatility terms. This paper not only combines SVM with a rolling window approach to make our method more meaningful and practical in a financial domain, but also integrates sentiment analysis into a machine learning method based on SVM to forecast stock market movement direction with consideration of human emotions.

III. Methodology

This paper aims to forecast stock market movement direction by not only using financial market data, but also combining them with sentiment features that incorporate investor psychology. The features are extracted from unstructured news data automatically and then are expressed as sentiment indexes. In order to make the indexes more realistic and reliable, we take the day-of-week effect into consideration. Next, we employ SVM to forecast stock market trends, and make an adjustment to real market situations by use of a rolling window approach, and then compare the accuracy with the baseline method. Moreover, the prediction results are used to instruct investment decisions, and the performance of three different trading strategies are evaluated and compared. The overview of the stock market prediction architecture is illustrated in Fig. 1.

A. Investor Sentiment

This section is made up of three steps. We first build a web crawler to download news documents automatically from the Internet, and then construct daily sentiment indexes based on the corpus. At last, adjustments are made in consideration of the day-of-week effect.

Step 1 Web crawler: In this step, we aim to build a web crawler to automatically download the targeted textual documents from the Internet and store them to a database for further processing. The framework is clarified in Fig. 2. The web crawler begins with the seeds in the form of a list of URLs. The scheduler manages the queue of URLs, deciding the priority and eliminating duplicate parts. Next, the downloader is responsible for acquiring the web pages from the Internet and providing them to the spider, which is used to parse the pages and extract the targeted contents. What we need to obtain is comprised of two sections: one is the textual news with the date from the websites, and

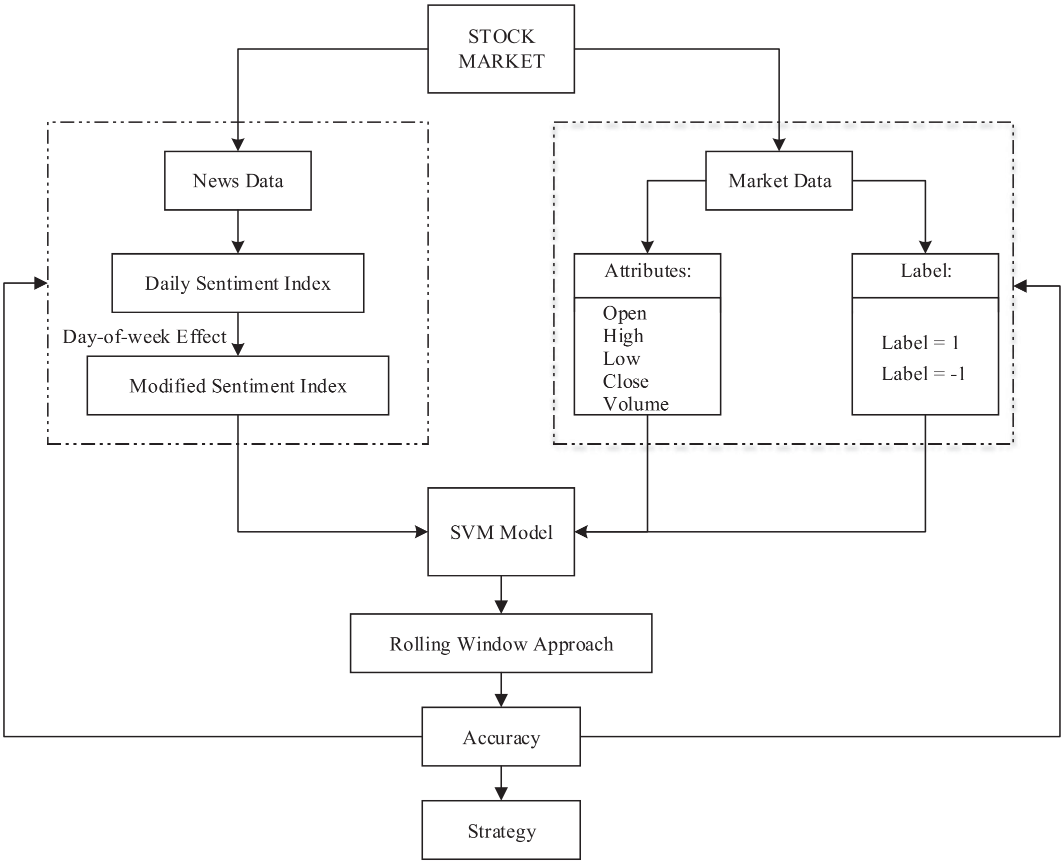


Fig. 1. Overview of stock market prediction architecture.

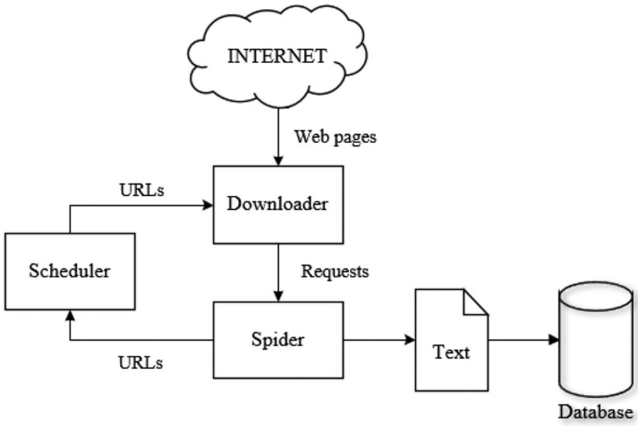


Fig. 2. Framework of the web crawler.

then we store the precious data into the database; the other is the URLs contained in the pages, and then the URLs are transported to the scheduler. The procedures are repeated until we get hold of all the targeted textual documents. Each of the documents is displayed as time, headline, and contents in the database.

Step 2 Daily sentiment: A sentence-based sentiment analysis approach is used to process the textual data during a specific period. We regard a sentence as a unit to interpret the meaning of the whole document instead of a single word because a sentence can express a relatively complete meaning and help address the ambiguity problem. As a result, a document is divided into sentences first. Next, we segment the sentences into separate words, then project the words onto the sentiment space, count the number of positive and negative words, assign a specific sentiment value, and decide the polarity of each sentence based on HowNet and Chinese Sentiment Analysis Ontology Base. HowNet is an online common-sense knowledge base unveiling interconceptual relationships and interattribute relationships of concepts as connoted in lexicons of the Chinese and their English equivalents [24]. Chinese Sentiment Analysis Ontology Base is constructed by Dalian University of Technology and depicts words and phrases from various aspects containing part of speech, polarity, and sentiment intensity. After that, we categorize each document. As there may be a large number of posts or articles in a day, a daily sentiment index S_t is calculated as

$$S_t = \begin{cases} 2{M_t}^{\text{bull}}/({M_t}^{\text{bull}} + {M_t}^{\text{bear}}) - 1, & {M_t}^{\text{bull}} > {M_t}^{\text{bear}} \\ 0, & {M_t}^{\text{bull}} = {M_t}^{\text{bear}} \\ 1 - 2{M_t}^{\text{bear}}/({M_t}^{\text{bull}} + {M_t}^{\text{bear}}), & {M_t}^{\text{bull}} < {M_t}^{\text{bear}} \end{cases} \tag{1}$$

where ${M_t}^{\text{bull}}$ denotes the number of positive comments, whereas ${M_t}^{\text{bear}}$ denotes the number of negative comments in day t. The value of S_t ranges from –1 to 1, where 0 means people hold a neutral position. And, if the value is larger than 0, it means most people take a positive view; if the value is less than 0, it means most people take a negative view.

Step 3 Modified sentiment: The day-of-week effect is one of the most well-known financial anomalies [5], which means that the average return on a Monday is much lower than that on the other days of the week. The reason includes that large amount of news is reported on the weekend or on Friday just after the market is closed. With such considerable and valuable information to deal with, investors are very likely to change their mind and take actions on Mondays. Furthermore, corporations also tend to release important news on the weekend to ensure the stability of the stock and boost the public image. If it is bad news, investors will have enough time to digest and accept it, whereas if it is good news, companies can continuously spread out news to make it known by more and more people and expand their coverage.

In this procedure, we aim to gauge the effect from Saturday to Monday by using an exponential time function as news has greater impact when it is more recent. Barberis *et al.* [25] introduced a measure of "sentiment" by using an exponential function on past price changes on the stock market, accordingly, we define sentiment on Monday as a weighted average of past sentiment where the weights decrease exponentially. The expression is clarified as

$$S_m = e^{-\lambda t_1} S_1 + e^{-\lambda t_2} S_2 + e^{-\lambda t_3} S_3 \tag{2}$$

where S_1, S_2, and S_3, respectively, stands for Saturday sentiment, Sunday sentiment, and Monday sentiment; S_m is the modified Monday sentiment; $\lambda(\lambda > 0)$ is prescribed; $t_1 = 2$, $t_2 = 1$, and $t_3 = 0$.

Similarly, the stock market is also closed on national holidays or on some special days, so we generalize (2) to more common occasions. Assume, there are n holiday days on the stock market, then the sentiment on $n + 1$th day is represented as

$$S_{n+1} = e^{-n\lambda} S_1 + e^{-(n-1)\lambda} S_2 + \cdots + e^{-\lambda} S_n + S_{n+1}. \tag{3}$$

B. Support Vector Machine

SVM is a supervised machine learning model for classification, which was proposed by Vapnik in the 1990s [19]. Assume that there is an input space X, an output space Y, and a training dataset T

$$T = \{(x_i, y_i), i = 1, \ldots, l\} \in (X \times Y)^l \tag{4}$$

where $x_i \in R^n$, $y_i \in Y = \{-1, 1\}$, and then introduce a transformation $\mathrm{x} = f(x)$ such that $R^n \rightarrow H$, where H is the Hilbert space, so the training set is then denoted as

$$T_f = \{(x_i, y_i), i = 1, \ldots, l\} \in (H \times Y)^l \tag{5}$$

where $\mathrm{x}_i = f(x_i) \in H$ and $y_i \in Y = \{-1, 1\}$. Thus, we can find a linear separating hyperplane $(w^* \cdot \mathrm{x}) + b^* = 0$ in the Hilbert space, and then we can obtain a separating hyperplane $w^* \cdot f(x) + b^* = 0$ and a decision function $D(x) = \mathrm{sgn}((w^* \cdot \mathrm{x}) + b^*) = \mathrm{sgn}(w^* \cdot f(x) + b^*)$ in the original space R^n.

SVM is an optimization problem that aims to maximize the margin. The margin between two hyperplanes in the Hilbert space is $2/\|w\|$. The two hyperplanes are classified as

$$(w \cdot \mathrm{x}) + b = 1 \text{ and } (w \cdot \mathrm{x}) + b = -1. \tag{6}$$

An SVM model can be represented as

$$\min_{\omega, b, \xi} \ \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{l} \xi_i \tag{7}$$

$$\text{s.t. } y_i((w \cdot f(x_i)) + b) \geq 1 - \xi_i, i = 1 \tag{8}$$

$$\xi_i \geq 0, i = 1, \ldots l \tag{9}$$

where ξ_i is a tolerable training error, and C is a positive constant parameter to evaluate the tradeoff between training errors and margin maximization. In order to solve the problem, we can transform it to its dual problem, and the solution set of the dual problem is the same as the QP problem as shown in the following equation:

$$\min_{\alpha} \ \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j (f(x_i) \cdot f(x_j)) - \sum_{j=1}^{l} \alpha_j \tag{10}$$

$$= \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{j=1}^{l} \alpha_j \tag{11}$$

$$\text{s.t. } \sum_{i=1}^{l} y_i \alpha_i = 0 \tag{12}$$

$$0 \leq \alpha_i \leq C, i = 1, \ldots l \tag{13}$$

where $\alpha = (\alpha_1, \ldots, \alpha_l)^T$ is a Lagrange multiplier, $K(x_i, x_j)$ is defined as a kernel function with $K(x_i, x_j) = (f(x_i) \cdot f(x_j))$, and $(\cdot)$ denotes the inner product in the Hilbert space. There are many kinds of kernel functions, such as RBF kernel $K_{\mathrm{rbf}}(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2)$ and polynomial kernel $K_{\mathrm{poly}}(x_i, x_j) = ((x_i \cdot x_j) + 1)^d$, where γ and d are kernel parameters. If we choose a prescribed parameter C and a proper kernel function $K(x_i, x_j)$, we can compute the solution $\alpha^* = ({\alpha_1}^*, \ldots, {\alpha_l}^*)^T$ of QP problem (11)–(13), and then b^* is calculated as

$$b^* = y_j - \sum_{i=1}^{l} y_i {\alpha_i}^* K(x_i, x_j). \tag{14}$$

Finally, we construct the decision function (15) that can be used to classify

$$D(x) = \mathrm{sgn}\left(\sum_{i=1}^{l} y_i {\alpha_i}^* K(x_i, x_j) + b^*\right). \tag{15}$$

IV. Experiment

A. Data Description

We intend to explore the trend of a very important index in China, the SSE 50 Index, not only by using stock market data but also exploiting news documents related to it and its constituents. The SSE 50 Index is a primarily blue-chip stock index on the Shanghai stock market, and it is made up of the 50 largest stocks of good liquidity and representativeness. Conventional time series data include opening price, closing price, high for the day, low for the day, trading volume in number of shares, trading volume in RMB, change in RMB, and change in percentage. We download such data of the SSE 50 Index and its 50 constituents from the Wind Economic Database, which is the market leader in China's financial information service industry.

Accordingly, we applied the web crawler we built to download all the posts and documents of the 51 shares from the Sina stock forum and Eastmoney stock forum over the period between June 17th, 2014 and June 7th, 2016, including 485 trading days. The two forums are widely regarded as active and mainstream communities in China. The number of reviews of each stock is 37 855 on average, peaking at 23 236 and reaching the lowest point at 7797. The details are illustrated in the second column of Table I. The total number of the reviews on the Sina stock forum and Eastmoney stock forum is 1 930 592 after filtering and denoising during the given period.

B. Sentiment Calculation

Under Steps 2–3 in Section III, we can compute 51 sentiment indexes for 51 stocks. In Step 2, we first segment each document into several sentences by identifying punctuations, such as "," "." and "!" Then sentences are divided into separate words, and if there appears a negative word, it is treated as a whole with the word next to it. For example, if people say "我不满意这股票 (I'm not satisfied with the stock)," after word segmentation, the sentence becomes four words "我(I'm)" "不(not)" "满意(satisfied with)" "这(the)" "股票 (stock)." If we directly project the words to the sentiment space, the program will tell us the sentence is optimistic because of the positive word "满意(satisfied with)." So, we need to treat "不(not)" "满意(satisfied with)" as a whole "不满意 (not satisfied with)" so that we can find the true meaning. Then, we need to categorize each document, assume there are p_i positive sentences and n_i negative sentences in document *i*; if $p_i > n_i$, the document is positive; if $p_i = n_i$, the document is neutral; if $p_i < n_i$, the document is negative. And, then we find on the day t, the number of positive comments is ${M_t}^{\text{bull}}$ and the number of negative comments is ${M_t}^{\text{bear}}$, so a daily sentiment index is calculated by using formula (1), with the value ranging from –1 to 1, where 0 means people hold a neutral position. And, if the value is between 0 and 1, it means people hold a positive view; if the value is between –1 and 0, it means people take a negative view. Then, by considering the day-of-week effect, the modified sentiment indexes are calculated according to (2) and (3). The third to seventh columns of Table I show the major statistics of the modified sentiment indexes, including the mean, median, standard deviation, skewness, and kurtosis, of each stock sentiment index. It demonstrates that the majority of investors have a positive view on the 51 stocks since the mean and median of most stocks are positive rather than negative. From Fig. 8 or 9, we can find that the price of the SSE 50 Index increases at first and then drops suddenly, but overall people are optimistic. It implies that people are not ready for the decrease. In fact, many individuals and companies lost a great deal during the period that is also called the stock market disaster in China.

TABLE I
STATISTICS OF SENTIMENT INDEXES

	Number	Mean	Median	Std. Dev.	Skewness	Kurtosis
1	23236	0.35637	0.40967	0.26127	2.01860	8.80423
2	81248	0.53358	0.58165	0.20246	3.20545	20.72166
3	33196	0.21777	0.25837	0.21603	1.60730	9.98526
4	34591	0.32120	0.37197	0.21405	1.82987	8.50063
5	27811	0.47152	0.51065	0.17151	3.64575	25.45141
6	18754	0.29471	0.34524	0.21371	1.81329	9.32399
7	33537	0.36117	0.39134	0.17573	2.23436	15.70559
8	34968	0.52412	0.57139	0.17972	3.05854	19.38373
9	9853	0.28766	0.31879	0.19666	1.21639	9.65186
10	40644	0.42828	0.48220	0.22630	2.27251	10.46903
11	31386	0.25575	0.30260	0.20527	2.23891	11.81875
12	36994	0.45809	0.50773	0.20597	2.51271	14.05555
13	17032	0.18246	0.22299	0.25101	1.05804	5.89966
14	10766	0.19424	0.23456	0.20181	1.91364	10.09988
15	28200	0.39511	0.42887	0.16154	2.17121	16.48476
16	26608	0.13916	0.19250	0.21458	1.61710	7.60458
17	17052	0.17892	0.24357	0.24875	1.50026	5.94911
18	46914	0.18769	0.24859	0.23176	1.80984	8.71453
19	27024	0.64060	0.67804	0.16707	4.05740	31.82885
20	28158	0.19699	0.25102	0.20822	2.06887	10.11175
21	33764	0.24181	0.28347	0.20219	1.65773	8.80704
22	38000	0.31557	0.36938	0.20474	2.16160	11.06345
23	37925	0.40178	0.46159	0.21796	2.42526	11.64860
24	57848	0.63125	0.69425	0.25169	2.64671	12.65547
25	43594	0.04643	0.10727	0.26298	1.06220	5.51963
26	19409	0.37464	0.41017	0.20768	1.63800	10.20172
27	14538	0.40000	0.44882	0.19462	1.72010	10.35468
28	17470	0.39013	0.42699	0.17367	3.03270	20.76046
29	22766	0.41687	0.38462	0.14956	-0.78036	22.28815
30	26006	0.52197	0.57188	0.19350	2.46205	13.48589
31	41348	0.58017	0.63323	0.20431	2.50395	13.37901
32	30269	0.26305	0.30691	0.17272	1.92448	12.25382
33	37313	0.46347	0.52139	0.23625	1.91042	8.52865
34	24289	0.54739	0.62792	0.20351	3.30866	19.27568
35	16982	0.43495	0.48230	0.20400	2.15579	10.87906
36	79823	0.15132	0.20275	0.25479	0.94820	6.24546
37	21645	0.00751	0.01479	0.21188	-0.26603	7.62239
38	71082	0.35197	0.40319	0.20811	2.16227	11.59339
39	54386	0.50157	0.56257	0.21386	2.50068	13.03801
40	42746	0.30985	0.35117	0.19425	1.78007	10.41846
41	18339	0.59117	0.64865	0.20903	2.44041	13.36085
42	25493	0.53206	0.58415	0.20860	2.55991	14.14896
43	57196	0.53761	0.58005	0.17451	2.78345	18.64874
44	55932	0.44264	0.47679	0.14944	2.96612	22.85836
45	12858	0.11312	0.15774	0.20706	2.08796	11.04161
46	152533	0.17153	0.22287	0.23559	1.41102	6.72343
47	104189	0.58461	0.65802	0.23296	2.65091	13.82870
48	7797	0.55417	0.60022	0.16712	2.91624	20.59029
49	16293	0.31672	0.36630	0.24080	1.39273	7.65575
50	17807	0.48988	0.53705	0.19318	2.10251	11.62268
51	122980	0.42754	0.46735	0.19745	2.69997	16.05438

This table illustrates the major statistics of the 51 sentiment indexes. The second column shows the number of comments on each stock. The third to seventh column, respectively, display the mean, median, standard deviation, skewness, and kurtosis of each stock sentiment index.

Nevertheless, we do not utilize them directly to explore the stock market trend, but select 8 sentiment indicators based on 51 sentiment indexes. The reasons include that the number of the features of market data and sentiment indexes needs to be balanced. Although we have already computed 51 sentiment indexes, the number of market data is around 10, so it is unfair for the market attributes to some extent. Next, too many variables are inclined to cause the problem of overfitting. Furthermore, we find that using 8 sentiment features achieves a better result in forecasting the SSE 50 Index than using all 51 indexes. Inspired by the market attributes, sentiment features consist of the highest of modified sentiment indexes, the lowest of modified

TABLE II
DESCRIPTION OF EIGHT FEATURES

Features	Description
s1	The average of modified sentiment indexes
s2	The highest of modified sentiment indexes
s3	The lowest of modified sentiment indexes
s4	The median of modified sentiment indexes
s5	The value between the highest and the lowest
s6	The change of the average
s7	The percentage of the average change
s8	The standard deviation of modified sentiment indexes

sentiment indexes, the median of modified sentiment indexes, the average of modified sentiment indexes, the difference between the highest and the lowest, the change of the average (a certain day's average minus the last day's average), the percentage of the average change (a certain day's average minus the last day's average and divided by the last day's average), and the standard deviation of 51 modified sentiment indexes. Table II describes eight selected features. Fig. 3 sheds light on the trend of the 8 variables after standardization in 484 trading days.

C. Prediction

First, we need to label the data according to the following equation:

$$\text{Label} = \begin{cases} 1, & \text{Close}_{t-1} < \text{Close}_t \\ -1, & \text{Close}_{t-1} > \text{Close}_t \end{cases} \tag{16}$$

where Close_t denotes the close price of the SSE 50 Index, and Close_{t-1} stands for the close price on the previous day. Besides, 1 also means buy order as it indicates the increase, whereas -1 means sell order as it implies the decline.

Next, we implement two experiments to predict the index movement direction. Experiment 1 is to use market data, which include opening price, closing price, high for the day, low for the day, trading volume in number of shares, trading volume in RMB, change in RMB, and change in percentage. And then, we combine them with sentiment features for Experiment 2. We employ classification accuracy Acc to assess the performance, as shown in the following equation:

$$\text{Acc} = \frac{T_{++} + T_{--}}{T_{++} + T_{--} + F_{-+} + F_{+-}} \tag{17}$$

where T_{++} denotes that the true value is $+1$ and the prediction value is also $+1$; T_{--} denotes that the true value is -1 and the prediction value is also -1; F_{+-} denotes that the true value is $+1$, whereas the prediction value is –1; F_{-+} denotes that the true value is -1, whereas the prediction value is $+1$.

A fivefold cross-validation approach is adapted to train an SVM model. Eventually, we find the proper parameters and the kernel functions to achieve the best performance. Figs. 4 and 5 document the processes of parameter selection. Panel A of Table III sheds light on the prediction results. For Experiment 1, the accuracy can be 79.96%, and we use RBF kernel function, $C = 256, \gamma = 0.9942$; for Experiment 2, the accuracy can be as high as 97.73%, and we employ RBF kernel function, $C = 181.0193, \gamma = 0.005524$.

However, the two kinds of methods cannot be applied in forecasting stock market movement direction for the reason that they lead to look-ahead bias, which is created by the use of information or data that would not have been known or available during the period being analyzed. For example, we have some data from January to May, and implement the fivefold cross validation approach. When we use the data from February to May as training set and the data in January as testing set, it is impossible because in January, we will never know what will happen from February to May. On the other hand, that does not mean the method is useless, and it is an important procedure to select the proper kernel functions and parameters as well as address the overfitting problem. In other words, the purpose of the procedure is not to forecast but select the proper kernel functions and parameters.

We utilize a realistic rolling window approach to overcome the challenge, and accordingly, we need to single out a best window for both experiments. The principle of choosing the rolling windows is that we use n previous days to forecast the next day's movement direction, repeat the procedure, and change the value of n until an SVM model achieves the highest accuracy with the parameters and the kernel function we have already selected. Figs. 6 and 7 display the rolling window choosing processes. For Experiment 1, the optimal rolling window is 68, and the highest accuracy is 71.33%; for Experiment 2, the optimal rolling window is 76, and the highest accuracy is 89.93%. It is clear from each figure that the accuracy is relatively stable at around the optimal rolling window. And it remains lower than the highest accuracy after 80 days in Fig. 6 or 90 days in Fig. 7, which are not shown in the figures due to the large size.

Panel B of Table III sheds light on the prediction accuracy of SVM with rolling windows. We can see from the table that adding sentiment features to the baseline model helps boost the prediction performance significantly. The reasons why the empirical result of forecasting the market index movement direction can be as high as 89.93% probably include that investor sentiment plays a very important role on the stock market. Furthermore, sentiment contains valuable knowledge about the asset values and can be considered as one of the leading indicators of the stock market.

In addition, LR is used to reexamine the conclusions. LR is a very important method in prediction, and it is good at modeling the probability of a response based on a set of predictor variables. In order to compare with SVM, fivefold cross validation is also applied to forecast the movement direction of the SSE 50 Index. For Experiment 1, the accuracy is 70.96%, GD with the maximum number of iterations set to 600 is implemented to converge the result; for Experiment 2, the accuracy is 86.59%, stochastic GD with the maximum number of iterations set to 1000 is implemented to converge the result. Panel C of Table III confirms that investor sentiment is vital to the stock prices; and illustrates that the accuracy of LR with fivefold cross validation is acceptable, but it is not only less than SVM with fivefold cross validation but also less than SVM with a rolling window approach, suggesting that our method is realistic and efficient.

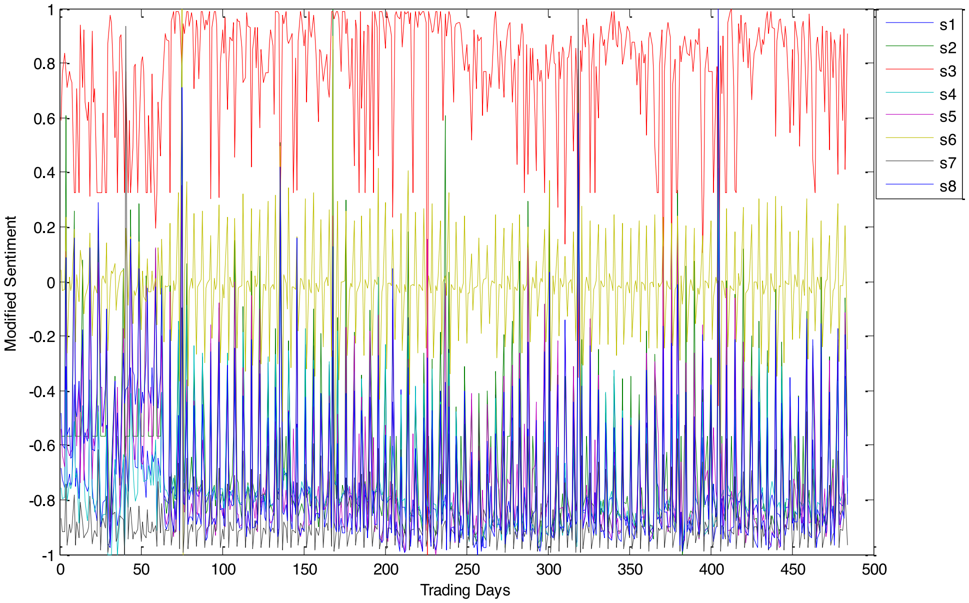


Fig. 3. Sentiment features in trading days.

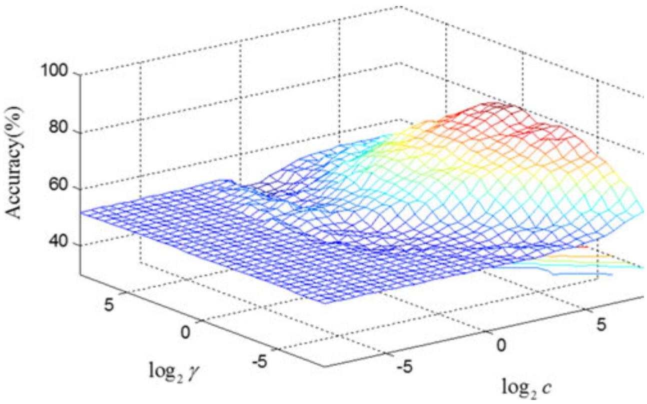


Fig. 4. Parameter selection process for Experiment 1.

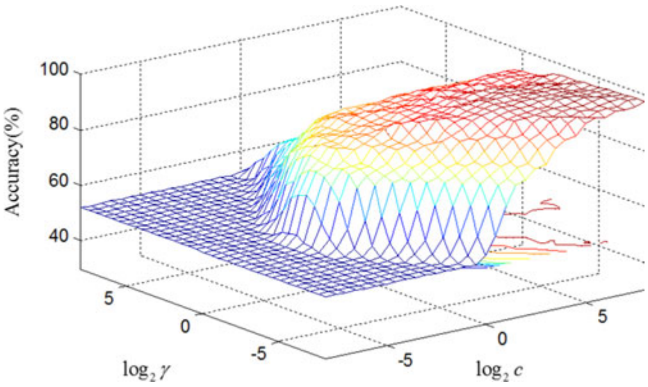


Fig. 5. Parameter selection process for Experiment 2.

D. *Investment Performance*

This section tries to discover if the prediction results are of benefit to the investment. Some measures are employed to evaluate and compare the performance of the methods. AI is computed based on the stock points. For example, if we buy a stock at the price of 100 and sell it at 150, then we earn 50 stock points and AI is 50 stock points; after that, we short the equity at 150 and liquidate the position at 120, then we make 30 stock points and AI becomes 80 stock points. Maximum drawdown (MDD) is the maximum decline of a series from a peak to a trough over a specified time period [26]. MDD at time T is expressed as

$$\mathrm{MDD} = \sup_{t\in[0,T]} \left[\sup_{s\in[0,t]} X(s) - X(t) \right] \tag{18}$$

where $X(t)$ is a random process on $[0, T]$. MDD time illustrates when MDD occurs. The expected maximum drawdown (EMD) is an estimate of the maximum losses average, based on a geometric Brownian motion assumption. MDD and EMD are regarded as indicators of downside risk. Sharpe ratio (SR) is a way to gauge the performance of an investment by calculating the adjusted-risk return [27], which is defined as

$$\mathrm{SR} = \frac{r_a - r_f}{\sigma_a} \tag{19}$$

where r_a is the mean of the asset returns, σ_a is the standard deviation of the asset returns, and r_f denotes the risk-free rate and set to be 0 in this paper.

We have previously mentioned that the label 1 means buy order, whereas –1 means sell order. Accordingly, we follow the prediction results to buy or sell, then find if it is beneficial to support investment decisions and reduce financial risk. We postulate that short selling mechanism is allowed and there are no market frictions. The prediction results of Experiment 1 and

TABLE III
PREDICTION RESULTS

Panel A: Prediction accuracy of support vector machine (SVM) with fivefold cross validation				
	Accuracy		C	γ
Experiment 1	0.7996		256	0.9942
Experiment 2	0.9773		181.0193	0.0055
Panel B: Prediction accuracy of SVM with rolling windows				
	Accuracy	C	γ	Rolling window
Experiment 1	0.7133	256	0.9942	68
Experiment 2	0.8993	181.0193	0.0055	76
Panel C: Prediction accuracy of logistic regression (LR) with fivefold cross validation				
	Accuracy		Max	Optimization
Experiment 1	0.7096		600	Gradient descent (GD)
Experiment 2	0.8659		1000	Stochastic GD

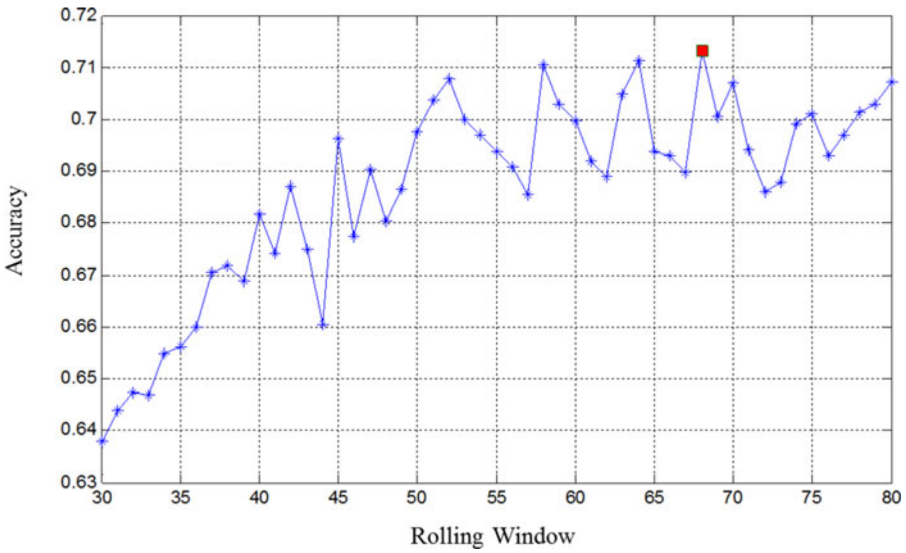


Fig. 6. Rolling window choosing process for Experiment 1.

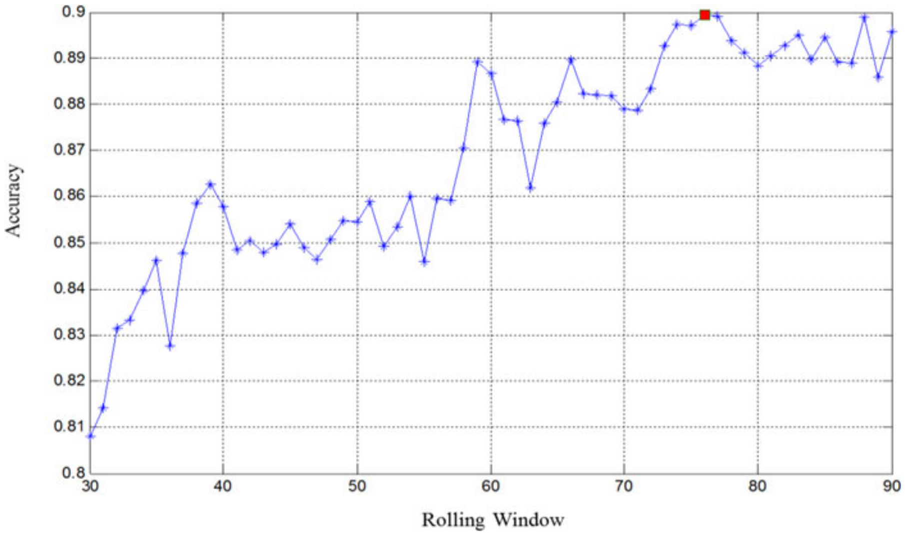


Fig. 7. Rolling window choosing process for Experiment 2.

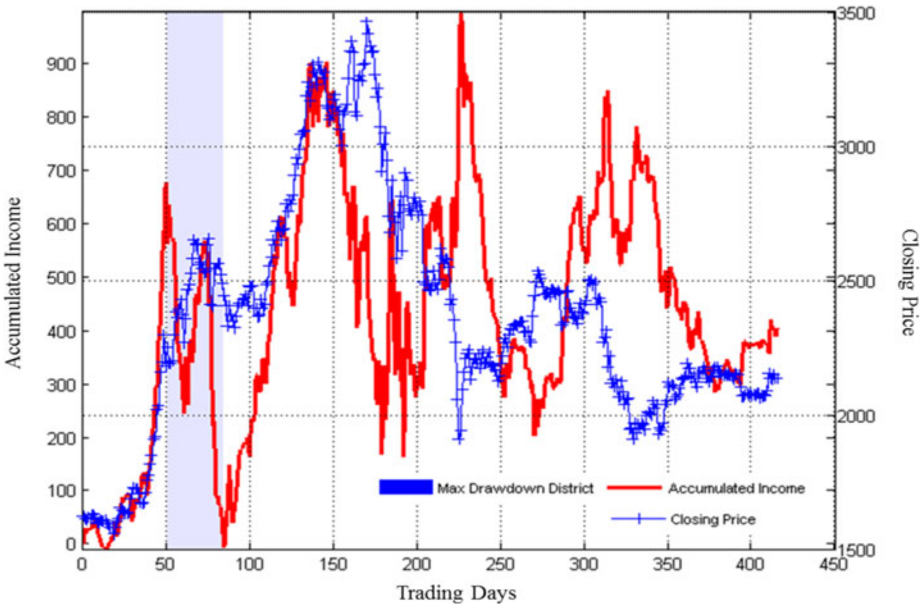


Fig. 8. Accumulated income (AI) for Experiment 1.

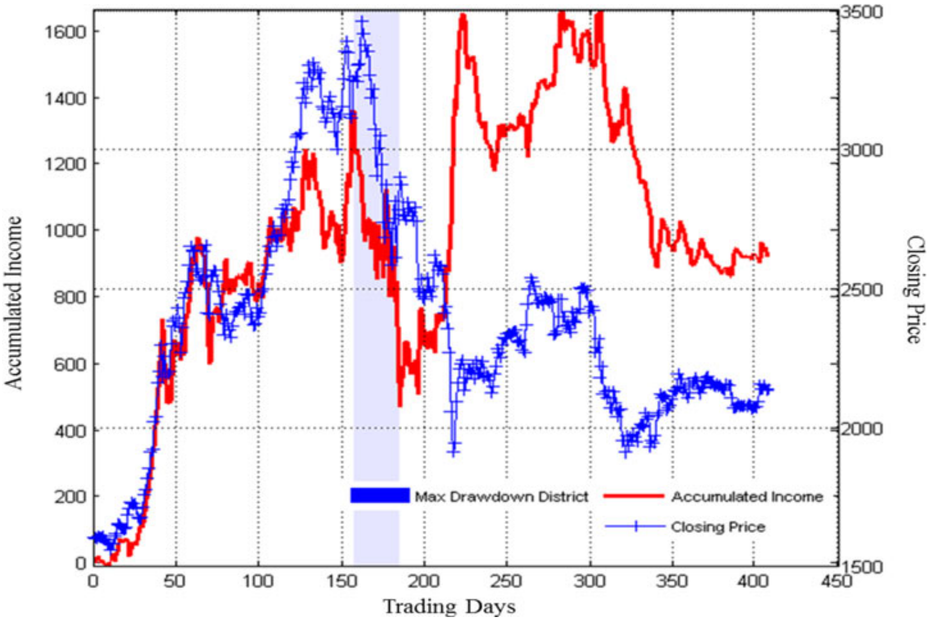


Fig. 9. AI for Experiment 2.

Experiment 2 are, respectively, utilized to compute AI and MDD. Figs. 8 and 9 demonstrate AI compared with the trend of the closing price of the SSE 50 Index, and highlight the MDD district of AI simultaneously. It can be seen from the line graphs that the AI of Experiment 2 (916.6264 stock points) is more than two times than that of Experiment 1 (404.8598 stock points). Moreover, although both the methods fail to detect the dramatic decline at first, Experiment 2 predicts the trend afterward, and is able to uncover the following rise. The sharp decrease is known as the Chinese stock market crash in 2015. Besides, the MDD of Experiment 2 is 0.3770, whereas the MDD of Experiment 1 is 0.4073. Similarly, the EMD of the next 30 days for Experiment 2 (0.1882) is also lower than that of Experiment 1 (0.2546). This implies that sentiment features help to reduce risks for investors and institutions. In addition, SR significantly went up by adding sentiment variables, which indicates that the results of

TABLE IV
INVESTMENT PERFORMANCE

	AI/stock points	SR	MDD	MDD time/ trading days	EMD
Experiment 1	404.8598	0.3263	0.4073	50–85 (35 days)	0.2546
Experiment 2	916.6264	0.8263	0.3770	157–185 (28 days)	0.1882
Experiment 2 (strategy)	1300.4639	1.2248	0.3034	306–384 (78 days)	0.1572

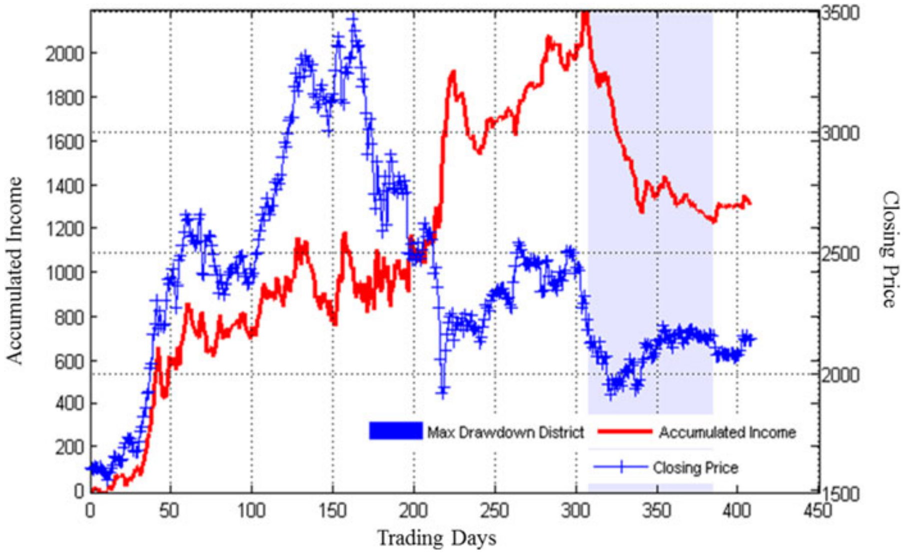


Fig. 10. Stop-loss order strategy for Experiment 2.

Experiment 2 can help investors make higher profits with the same risk.

Finally, we try to discover whether we can achieve a better performance based on the prediction results of Experiment 2. Hence a stop-loss order strategy is applied to limit the potential losses. We set the stop order to be 95 stock points, which means that we would stop to trade if 95 stock points had already been lost in a trading day. The strategy accomplishes a much better performance, and all of the measures that are displayed in the third row of Table IV improve significantly. From Figs. 9 and 10, we can point out that a stop-loss order can put an end to a losing period, but it cannot turn it into a win. However, the reduced loss also means an increase in a final AI. As a result, we can deduce that our approach can be of great benefit to investors if combined with a proper strategy, for example a stop-loss order strategy. In other words, the method is useful to decision-making processes that are pervasive phenomena of nature [28].

V. Conclusion and Future Work

In this paper, we aim to exploit investor sentiment to forecast stock market movement direction by emphasizing the role of investors. Investor psychology drives the stock market [1] and it matters for our research. Accordingly, user-generated content on the Internet provides a precious source to reflect investor psychology. Sentiment analysis is used to convert unstructured textual documents into daily sentiment indexes. Furthermore, the financial anomaly day-of-week effect that means the average return on Mondays is much lower than that on the other days of the week probably influences the precision of the sentiment indexes, so we adjust the indexes by introducing an exponential function on past sentiment changes on weekends and then generalize to holidays. Correspondingly, Sina Finance and Eastmoney, two typical financial websites, were selected as experimental platforms to obtain a corpus of financial review data. Then, the machine learning model SVM is employed to predict a very important index in China, the SSE 50 Index, by implementing fivefold cross validation and a realistic rolling window approach. Empirical results illustrate that combining sentiment features with stock market data can achieve a much better performance than just using stock market data in forecasting movement direction. The accuracy can be as high as 89.93% with a rise of 18.6% after introducing sentiment variables. Furthermore, if combined with a stop-loss order strategy, our approach can help investors reduce risks and make wiser decisions. In addition, we find that sentiment probably contains precious information about the asset fundamental values and can be regarded as one of the leading indicators of the stock market. For future work, we consider expanding the time interval, which also means crawling more textual documents from the Internet. And, it is imperative to boost the efficiency of the method to process voluminous data in real time.

References

[1] R. J. Shiller, *Irrational Exuberance*. Princeton, NJ, USA: Princeton Univ. Press, 2000.

[2] I. Perikos and I. Hatzilygeroudis, “Recognizing emotions in text using ensemble of classifiers,” *Eng. Appl. Artif. Intell.*, vol. 51, pp. 191–201, 2016.

[3] B. Wu, X. Zhou, Q. Jin, F. Lin, and H. Leung, “Analyzing social roles based on a hierarchical model and data mining for collective decision-making support,” *IEEE Syst. J.*, vol. 11, no. 1, pp. 356–365, Mar. 2017.

[4] B. Liu and L. Zhang, “A survey of opinion mining and sentiment analysis,” *Mining Text Data*. New York, NY, USA: Springer, 2012.

[5] R. J. Shiller, “From efficient markets theory to behavioral finance,” *J. Econ. Perspectives*, vol. 17, no. 1, pp. 83–104, 2003.

[6] F. C. Kelly, *Why You Win or Lose: The Psychology of Speculation*. North Chelmsford, Massachusetts, USA: Courier Corp., 2003.

[7] E. D. Maberly, “Eureka! Eureka! Discovery of the monday effect belongs to the ancient scribes,” *Financial Anal. J.*, vol. 51, pp. 10–11, 1995.

[8] J. Zhang, Y. Lai, and J. Lin, “The day-of-the-week effects of stock markets in different countries,” *Finance Res. Lett.*, vol. 20, pp. 47–62, 2017.

[9] W. Huang, Y. Nakamori, and S.-Y. Wang, “Forecasting stock market movement direction with support vector machine,” *Comput. Oper. Res.*, vol. 32, no. 10, pp. 2513–2522, 2005.

[10] L. Yu, H. Chen, S. Wang, and K. K. Lai, “Evolving least squares support vector machines for stock market trend mining,” *IEEE Trans. Evol. Comput.*, vol. 13, no. 1, pp. 87–102, 2009.

[11] C. C. Aggarwal and C. Zhai, *Mining Text Data*. New York, NY, USA: Springer, 2012.

[12] X. Li, H. Xie, L. Chen, J. Wang, and X. Deng, "News impact on stock price return via sentiment analysis," *Knowl.-Based Syst.*, vol. 69, pp. 14–23, 2014.

[13] M. Baker and J. Wurgler, "Investor sentiment and the cross-section of stock returns," *J. Finance*, vol. 61, no. 4, pp. 1645–1680, 2006.

[14] A. Edmans, D. Garcia, and Ø. Norli, "Sports sentiment and stock returns," *J. Finance*, vol. 62, no. 4, pp. 1967–1998, 2007.

[15] P. C. Tetlock, "Giving content to investor sentiment: The role of media in the stock market," *J. Finance*, vol. 62, no. 3, pp. 1139–1168, 2007.

[16] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, 2011.

[17] R. A. Gillam, J. B. Guerard, and R. Cahan, "News volume information: Beyond earnings forecasting in a global stock selection model," *Int. J. Forecast.*, vol. 31, no. 2, pp. 575–581, 2015.

[18] N. Oliveira, P. Cortez, and N. Areal, "Stock market sentiment lexicon acquisition using microblogging data and statistical measures," *Decis. Support Syst.*, vol. 85, pp. 62–73, 2016.

[19] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 2013.

[20] M. Cecchini, H. Aytug, G. J. Koehler, and P. Pathak, "Detecting management fraud in public companies," *Manage. Sci.*, vol. 56, no. 7, pp. 1146–1160, 2010.

[21] Y. Kara, M. A. Boyacioglu, and Ö. K. Baykan, "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul stock exchange," *Expert Syst. Appl.*, vol. 38, no. 5, pp. 5311–5319, 2011.

[22] P.-F. Pai and C.-S. Lin, "A hybrid ARIMA and support vector machines model in stock price forecasting," *Omega*, vol. 33, no. 6, pp. 497–505, 2005.

[23] D. D. Wu, L. Zheng, and D. L. Olson, "A decision support approach for online stock forum sentiment analysis," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 8, pp. 1077–1087, Aug. 2014.

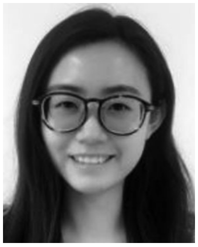
[24] Z. Dong, Q. Dong, and C. Hao, "HowNet and its computation of meaning," in *Proc. 23rd Int. Conf. Comput. Linguistics, Demonstrations*, 2010, pp. 53–56.

[25] N. Barberis, R. Greenwood, L. Jin, and A. Shleifer, "X-CAPM: An extrapolative capital asset pricing model," *J. Financial Econ.*, vol. 115, no. 1, pp. 1–24, 2015.

[26] M. Magdon-Ismail, A. F. Atiya, A. Pratap, and Y. S. Abu-Mostafa, "On the maximum drawdown of a Brownian motion," *J. Appl. Probab.*, vol. 41, no. 1, pp. 147–161, 2004.

[27] W. F. Sharpe, "The sharpe ratio," *J. Portfolio Manage.*, vol. 21, no. 1, pp. 49–58, 1994.

[28] V. Shukla, G. Auriol, and K. W. Hipel, "Multicriteria decision-making methodology for systems engineering," *IEEE Syst. J.*, vol. 10, no. 1, pp. 4–14, Mar. 2016.



Rui Ren (M'17) is currently working toward the Ph.D. degree at the School of Economics and Management, University of Chinese Academy of Sciences, Beijing, China.

Her research interests include sentiment analysis, text mining, and behavioral finance.

Desheng Dash Wu (M'09–SM'14) is with the School of Economics and Management, University of Chinese Academy of Sciences, Beijing, China, and also with the Stockholm Business School, Stockholm University, Stockholm, Sweden. He has authored or coauthored more than 100 papers in refereed journals such as *Production and Operations Management, Decision Support Systems, Decision Sciences, Risk Analysis*, IEEE TRANSACTIONS ON SYSTEMS MAN AND CYBERNETICS, etc. He is the Editor of the Springer book series entitled "Computational Risk Management." His research interests include enterprise risk management in operations, performance evaluation in financial industry, and decision sciences.

Dr. Wu has been an Associate Editor/Guest Editor for the IEEE TRANSACTIONS ON SYSTEMS MAN AND CYBERNETICS, *Annals of Operations Research, Computers and Operations Research, International Journal of Production Economics, Omega*, etc. He is elected member of the European Academy of Sciences and Arts.

Tianxiang Liu is currently working toward the Ph.D. degree at the School of Economics and Management, University of Chinese Academy of Sciences, Beijing, China.

His research interests include investment and asset pricing.