WILEY | Hindawi

Research Article

The Effect of Online Investor Sentiment on Stock Movements: An LSTM Approach

Gaoshan Wang [ID],[1,2] **Guangjin Yu** [ID],[1] **and Xiaohong Shen** [ID][3]

[1]*School of Management Science and Engineering, Shandong University of Finance and Economics, Jinan 250014, China*
[2]*Institute of Digital Economy, Shandong University of Finance and Economics, Jinan 250014, China*
[3]*School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan 250014, China*

Correspondence should be addressed to Gaoshan Wang; gaoshanwang@126.com

With more and more investors exerting their voices through network forums or social media platforms, the relationships between online investor sentiment and stock movements have drawn more and more attention. In this paper, we crawl stock comments from China's most popular online stock forum, East Money (www.eastmoney.com), and then develop a sentiment classifier using the LSTM method. Using the online investor sentiment of the stock forum, we explore the effect of online investor sentiment on the stock movements of CSI300. The results show that online investor sentiment has a significant positive impact on both stock return and trading volume and remains significant after controlling book-to-market ratio, BETA, and market value. Moreover, investor sentiment has a significant positive impact on order imbalance of big trade, which represents the main flow of money in the market. As a result, investor sentiment has a positive impact on the major fund flows in the market. In other words, an increase in investor sentiment can boost the major money flows in the market to some extent. From a practical point of view, investor sentiment can assist investors to make investment decisions and help the government to regulate the stock market.

1. Introduction

Increasingly, individual investors move to online stock forums or social media platforms not only to find information on stocks but also to express opinions and comments related to stocks. The individual investor generated contents (IIGC) on these platforms have relations to stock movements in two ways. Firstly, online stock comments are emotional reflections of individual investors whose decision-making is influenced by their emotions. Secondly, some investors may use online stock comments as input for their investment decisions. Hence, the role of online stock forums or social media platforms in the stock market has drawn more and more attention from researchers and practitioners. And IIGC has become an important new source to derive investor sentiment, which in many researchers' points of view, is important information that can affect stock movements [1].

We call investor sentiment that is derived from IIGC online investor sentiment. Although the relationships between online investor sentiment and stock movements have been widely analyzed recently [2–8], there is no consensus regarding the causality between investor sentiment and stock movements [9]. For example, some empirical results show that investor sentiment affects stock return [5], while others show that investor sentiment is positively affected by prior stock price performance but has no predictive power for stock movements [10]. In addition, most research on this topic uses data in stock markets of developed countries more than emerging markets of developing countries. However, emerging stock markets in developing countries may have features that differ from those in developed countries.

Our objective is to investigate whether online investor sentiment has a significant effect on stock movements in Shanghai Stock Exchange (SHSE) and Shenzhen Stock Exchange (SZSE) in China. The Chinese stock market has its own characteristics that differ from those in developed

markets. One of the most prominent features of the Chinese stock market is that most of the investors are individual investors but not institutional investors. According to Shanghai Stock Exchange Statistics Annual (2019), as of December 31, 2018, there were 38.60 million shareholder's accounts in SHSE, among which 38.51 million accounts belonged to individual investors. The proportion of individual accounts in total accounts was more than 99%. Moreover, the holding value of individual investors was 4550.6 billion RMB while that of institutional investors was only 3227.9 billion RMB in SHSE by 2018. Zhu et al. [11] point out that in China stock market, the main investors are individual investors who have little experience in value investing and are always influenced by other investors, so it is normal for the stock prices to deviate from their intrinsic value and to fluctuate sharply.

Nowadays, with the development of Big Data and machine learning techniques, it is available to derive investor sentiment from a large number of online stock comments. In this paper, we adopt long short-term memory (LSTM) networks to achieve this goal. Firstly, we collect stock comments from an online stock forum in China, which is East Money (hereafter, EM; http://www.eastmoney.com). Secondly, we train a sentiment classifier based on LSTM and divide stock comments into three groups, that is, positive, negative, and neutral groups. Thirdly, we develop an online investor sentiment index with the statistics of the three groups. And finally, we investigate the relationships between online investor sentiment and stock movements in SHSE and SZSE.

The main contributions of this study are summarized as follows. First of all, the accuracy of the emotion classifier developed by the LSTM method is higher than that of dictionaries and Bayesian methods used in the past literature. Second, we focus on big trades in the stock market. Big trade is different from block trade in China. The former is conducted in the same way as ordinary trade, but a large amount of money is likely to cause fluctuations in the stock market. The latter is conducted on a separate special trading platform. As a result, institutional investors or individuals with deep pockets often opt for big trade rather than block trade to hide their information. To the best of our knowledge, our study is the first to explore the impact of investor sentiment on stock movements using big trade measures rather than block trade measures. Thirdly, the data of the stock forum is updated in real-time. The text data of investor comments can be periodically acquired through crawlers to obtain investor sentiment, so as to effectively assist investors in making decisions.

This paper is structured as follows. Section 2 describes the related works on investor sentiment measurement and the relationships of investor sentiment and stock movements. Section 3 describes the data and methodology used in this study. Section 4 provides empirical results and discussions. Finally, conclusions and further research are given in Section 5.

2. Related Literature

2.1. Measurement of Investor Sentiment. The key problem to study investor sentiment is how to measure investor

sentiment accurately. Several different investor sentiment indexes have been established in the existing literature. Generally, there are two main methods to measure investor sentiment. The first is to use market indicators, which include trading volume and turnover, mutual fund flows, closed-end fund discounts, dividend premium, volume of initial public offerings, and first-day returns on IPOs [12, 13]. The second is to use direct indicators, which are often obtained by questionnaire. Its greatest advantage is its simplicity and directness. The mostly used direct indicators include the University of Michigan Consumer Sentiment Index, American Association of Individual Investors Index (AAII), the Investors Intelligence Sentiment Index, and the UBS/GALLUP Index for Investor Sentiments [12, 14, 15]. Both methods have defects. Firstly, the market-based indicators are an indirect reflection of investor sentiment. Secondly, although questionnaire-based indicators are direct reflection of investor sentiment, it is often time-consuming and expensive to obtain them.

With the advent of Internet Big Data Era, the investor sentiment measurement method has made new progress. Capturing online investor sentiment from IIGC has intrigued many researchers. The most popular social media platforms are Twitter, Weibo, stock forums, and so on. A lot of researchers get IIGC from these social media platforms. Bollen et al. [16] analyze the text content of daily Twitter feeds and derive investor sentiment by two-mode tracking tools, that is, Opinion Finder and Google-Profile of Mood States (GPOMS). Cheng and Lin [17] derive investor sentiment from Weibo with ICTCLAS3.0. Chen et al. [18] extract investor sentiment from Seeking Alpha (http://seekingalpha.com). Li et al. [19] derive investor sentiment from Sina (http://www.sina.com) and East Money (http://www.eastmoney.com).

In addition, researchers have developed new ways to generate online investor sentiment: one based on dictionaries and the other based on machine learning. The accuracy of the first method depends on the sentiment dictionary. Researchers often use public dictionaries, such as GI (General Inquirer), LM (Loughran–McDonald), SentiWordNet, and HowNet. Loughran and McDonald [20] develop an alternative negative word list to derive online investor sentiment. Eickhoff and Muntermann [21] calculate online investor sentiment with the General Inquirer (GI) and Harvard IV-4 dictionary. Garcia [22] derives investor sentiment from Twitter with Harvard-4 dictionary. Sul, Dennis, and Yuan [5] also calculate investor sentiment with Harvard-4 dictionary. However, these dictionaries have a small size vocabulary. What is more, in the field of finance, some general emotional words in public dictionaries have no clear emotions any longer, and there are also a lot of verbal languages in financial online reviews. All these make the public general emotion dictionaries not fit for the financial area well [19]. Some researchers then develop specific dictionaries for the financial area, but there are few that could deal with the Chinese text.

Due to the flaw of the first method, researchers are inclined to adopt a machine learning based approach, which can be further divided into three categories: supervised

machine learning based method, unsupervised machine learning based method, and semisupervised machine learning based method. And the most popular of the three is the supervised machine learning based method, whose accuracy depends on the feature extraction. With a large enough training set and a good feature set, the supervised machine learning based method will achieve a high classification accuracy [23]. Li et al. [3] apply a naive Bayesian learning algorithm to obtain investor sentiment with data from Twitter. Borovkova and Tsiamas [24] adopt the LSTM neural network to gain investor sentiment and compare it with lasso and ridge logistic classifiers.

2.2. The Relationship between Investor Sentiment and Stock Movements. Different from the efficient market hypothesis (EMH), behavioral finance theory argues that investors are often irrational, and their decision-making is limited by their cognitive abilities [25–27]. For example, the investors may be overoptimistic or overpessimistic about stock related news. This means that the basic assumption of EMH may be unreliable in some circumstances and, in turn, suggests that investors' emotions may play a significant role in the determination of stock prices [28]. Hence, the effect of investor sentiment on stock movements has gained more and more attention. In this paper, we mainly focus on three indicators of stock movements, which are stock return, trading volume, and order imbalance of big trade. We, especially, want to know whether investor sentiment has a relation to main fund flows which can be reflected by big trade, so we focus on the order imbalance of big trade.

The relationship between investor sentiment and stock return is complex. According to behavioral finance theory, investors' decision-making is affected by their cognition and emotions, and the presence of a large number of emotionally driven investors can lead to price deviations from fundamental value. Therefore, stock return may be affected by investor sentiment, which is supported by most research. Lee et al. [29] investigate the role of investor sentiment in stock market volatility and excess earnings and find that the change of mood is negatively related to market volatility and is positively related to excess returns. Schmeling [30] finds that investor emotion has a negative impact on the average returns of stock market. Huang et al. [31] also show that investor sentiment has a significant negative effect on stock return. However, by mining investor sentiment from an online stock market forum, Hu and Tripathi [32] find that investor sentiment has a positive effect on stock return. Sul et al. [5] find that both positive sentiment and negative sentiment have a significant effect on stock return with investor sentiment indices mined from Twitter. Dimpfl and Kleiman [15] find investor pessimism has a negative effect on market return and has a positive effect on volatility and trading volume. Bouteska [33] finds that there is a positive correlation between investor sentiment standard deviation and cumulative abnormal return, and the main factor leading to this relationship is investor conservatism.

Using the number of net added accounts as a proxy for investor sentiment, Chu et al. [9] find that investor sentiment has no significant effect on stock return on weekly basis but find a significant effect with low-frequency timescale. Also, on weekly basis and adopting Baidu Searching Index as the agent variable of investor sentiment, Xie and Wang [4] give the opposite conclusion that high-frequency components of investor sentiment have a significant effect on stock return. Corredor et al. [34] find that investor sentiment has a significant negative effect on stock return when using global sentiment indices whereas finding no significant effect when using local sentiment indices. Adopting Baker and Wurgler (2006) sentiment index and on monthly basis, Jiang et al. [35] find that, in the longer run, investor sentiment has a significant effect on stock return but has no significant effect with shorter-term investor horizons. Ni et al. [36] find investor sentiment has a positive impact on stock return in 3 months whereas having a negative effect from 6 months to 12 months. He et al. [37] apply the American Association of Individual Investor (AAII) index as the proxy of investor sentiment and find that investor sentiment in different periods has a different effect on stock return; specifically, the present investor sentiment has a marked and positive impact on stock return; however, the past investor sentiment has played the opposite role.

As to trading volume, Baker and Stein [38] point out that it increases with investor sentiment. Baker and Wurgler [39] further argue that trading volume can be regarded as an investor sentiment index. Ryu et al. [40] also use trading volume as one of the investor sentiment proxies. Li et al. [41] find that there is significant comovement between the amount of investor attention and trading volume contemporaneously. By mining investor sentiment from daily content from a popular Wall Street Journal column, Tetlock [42] points out that when noise traders experience a negative belief shock, they sell stocks to arbitrageurs, increasing volume, and finds that unusually high or low pessimism predicts high market trading volume. With Google Search, Joseph et al. [43] find that, over a weekly horizon, online search intensity reliably predicts abnormal trading volume. By exploiting Facebook's Gross National Happiness Index, Siganos et al. [44] find that negative sentiments are related to increases in trading volume and return volatility. However, Oliveira et al. [7] show that the investor sentiment obtained through microblogging has no obvious effect on the prediction of trading volume.

Big trade, or large trade, according to CSMAR (China Stock Market and Accounting Research Database) research data service, refers to a transaction, the trading volume of which is larger than 100,000 shares. Big trade is different from block trade in China, the minimum threshold of which is 300,000 shares according to the rules of SHSE. In addition, the former is completed in normal trading hours, while the latter is completed on an off-exchange platform, mostly after the close of the trading day and information on trades is reported on the SHSE website in a timely fashion. Big trade can be seen as an indicator of information asymmetry. Early studies suggest that informed traders prefer using large trades to minimize transaction costs and to maximize the profit gained from their informed trading activities. This is because they face competition from other informed traders

and their private information could be short-lived [45]. On the other hand, to hide information, informed traders like to trade on the medium size transactions and prefer to big trade more than block trade. Hence, big trade could reflect the direction of major funds better than block trade, and studying big trade and capturing its patterns are significant to the operation of the stock market. For exploring whether investor sentiment is a predictive indicator of private information, we use order imbalance of big trade to reflect the overall private information and the main fund flow of the market. Order imbalance, proxying for unobservable traders' intention, measures the excess amount of buy orders over sell orders and often signals either private information or the arrival of public information [46]. Hung [47] finds that investor sentiment significantly affects the order submission behavior in the market and that investors become more active in optimistic times. Chiarella et al. [48] find that investor sentiment has an effect on order submission.

From the above analysis, we can see that the relationships between investor sentiment and stock movements are very complicated. There is much literature studying the influence of online investor sentiment on stock movements, but there is no unified conclusion so far. To the best of our knowledge, we think that the reason for mixed results might lie in different sentiment proxies, or different time horizons, or different markets, and so forth, especially some researchers use aggregate-level index while others use individual-level index. In this study, we use individual-level indices and daily data to examine the impact of investor sentiment on the performance of emerging market equities. In particular, we use large trading indicators to explore whether investor sentiment is related to major fund flows in the market, which complements and deepens existing achievements.

3. Methodology

Firstly, we construct an online investor sentiment index, and then we analyze the relationships of investor sentiment and stock return, trading volume, and order imbalance of big trade.

3.1. Development of Online Investor Sentiment Index. In this paper, the development of online investor sentiment includes 7 steps, as shown in Figure 1.

3.1.1. Data Collection. We develop a web crawler program with Python language, which can collect online stock comments according to stock code. In addition, our crawler is in compliance with the robots exclusion protocol and it spares bandwidth with interval crawling as well as avoiding a large number of simultaneous requests. Using the crawler, we collect online stock comments from EM (http://www.eastmoney.com).

3.1.2. Data Cleaning. Our goal is to obtain individual investor sentiment; however, there is a lot of online reviews or news posted by various organizations. Hence, we need to remove those institutional comments or news.
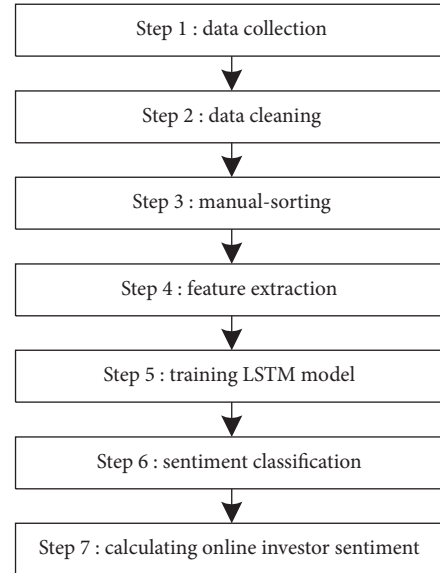

FIGURE 1: The steps of development of online investor sentiment.

3.1.3. Manual Sorting. The accuracy of a machine learning based classifier depends on a high quality manually annotated corpus. To get such a corpus, firstly, we randomly select 30,000 online stock comments. Because each comment is a short and simple sentence which can be easily read and understood by undergraduate students or people with at least a college degree, then we hire three financial graduate students who have investment experience to divide the comments into three groups according to their emotional polarities, that is, positive, negative, and neutral groups. Finally, the corpus can be used to train the machine learning model.

3.1.4. Feature Extraction. We use Jieba (https://pypi.org/project/jieba/) segmentation tool to cut sentences of the corpus into terms. And Jieba is a very popular Chinese segmentation tool used by researchers [49]. Then, we adopt the Word2Vec (https://pypi.org/project/word2vec/) model, which is a neural network-based model and quantifies the word into a vector considering the context, to extract the features of the terms [50].

3.1.5. Training LSTM Model. LSTM is a special type of Recurrent Neural Network (RNN), which is a kind of artificial neural network that is connected by nodes to form a cycle. RNN can simulate human reading sequence to read serialized data and transmit information through coding and memory of hidden layer neurons. However, because the vanilla RNN can only remember the short-term sequence data, it cannot solve the problem of long-term dependencies, and the phenomenon of "vanishing gradient" will appear.

LSTM extends RNN with three control units (i.e., input gate, output gate, and forget gate) and a memory cell. As the information enters the model, the cell in LSTM will judge the information, and the information that conforms to the rules will be kept, the information that does not conform to the

rules will be forgotten, and the memory function in time will be realized through the door switch to prevent the gradient from disappearing. The LSTM formulas are as follows:

$$\begin{aligned} f_t &= \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right), \\ i_t &= \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right), \\ C_t' &= \tanh\left(W_C x_t + U_c h_{t-1}\right) + b_c, \\ C_t &= f_t * C_{t-1} + i_t * C_t', \\ o_t &= \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right), \\ h_t &= o_t^* \tanh\left(C_t\right), \end{aligned} \tag{1}$$

where f_t, i_t, and o_t are forget gate, input gate, and output gate; x_t is the input vector; h_t is the state of the output units; C_t is the cell state vector; and U, W, and b are parameter matrices and vector (during training, each gate will get its own U, W, and b). Furthermore, σ is a sigmoid function and tanh is the hyperbolic tangent function:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \tag{2}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}}.$$

We build a training set and a testing set with the above corpus. Specifically, we take 80% of the corpus as the training set and the rest as the testing set, and we train the LSTM classifier with Python language, the learning curves shown in Figure 2.

We use the confusion matrix to calculate performance metrics, including accuracy, precision, recall, and F1, as follows:

$$\begin{aligned} \text{accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \\ \text{precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ F_1 &= \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \end{aligned} \tag{3}$$

where TP is the true positive, TN is the true negative, FP is the false positive, and FN is the false negative.

In this paper, we need to divide online stock comments into three classes, that is, positive class, neutral class, and negative class, so we calculate the performance metrics of each class, as shown in Table 1.

The accuracy of the positive group is 80.53% and the accuracy of the negative group is 72.66%, which are higher than the accuracy of Cheng and Lin [17], who analyze the investor sentiment and stock return of the Chinese stock market using sentiment dictionary.

3.1.6. Sentiment Classification.
We target the constituents of CSI300, which is composed of 300 stocks with the largest
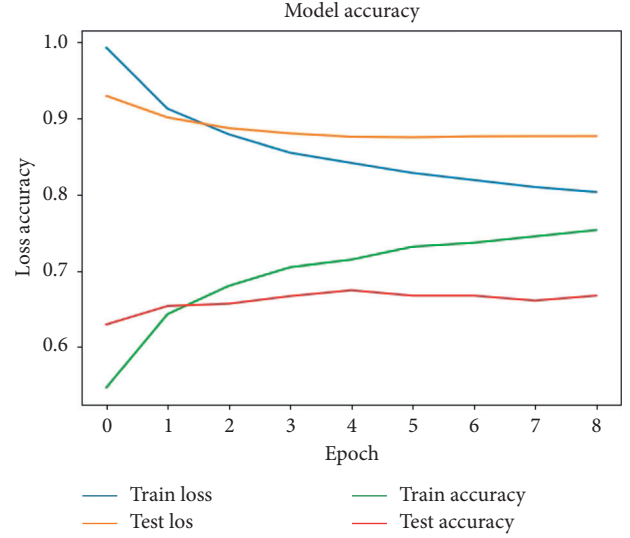

FIGURE 2: The learning curves of the training process.

TABLE 1: Performance metrics of the LSTM classifier.

Class	Accuracy	Precision	Recall	F_1
Positive	0.8053	0.7182	0.6890	0.7033
Neutral	0.8030	0.7153	0.6889	0.7019
Negative	0.7266	0.5776	0.6234	0.5997

market capitalization and liquidity from the entire universe of listed A-share companies in China. And their comments are classified into three groups with the trained LSTM classifier.

3.1.7. Calculating Online Investor Sentiment.
Much research adopts the following equation to calculate online investor sentiment for it is very robust [5, 51]:

$$S_{k,t} = \ln \frac{1 + M_{k,t}^{\text{pos}}}{1 + M_{k,t}^{\text{neg}}}, \tag{4}$$

where $S_{k,t}$ denotes the investor sentiment of stock k on day t. $M_{k,t}^{\text{pos}}$ denotes the number of positive comments of stock k on day t. $M_{k,t}^{\text{neg}}$ denotes the number of negative comments of stock k on day t.

However, there exists a silent majority effect on Internet; that is, many users do not post comments but only browse comments. Therefore, the users who post comments cannot represent the total users and equation (4) may have a large bias. In this paper, we consider the reading volume of comments for it represents attentions or opinions of silent majority, and we adopt an improved form of equation (4), as follows:

$$S_{k,t} = \ln \frac{1 + \sum_i R_{i,k,t}^{\text{pos}}}{1 + \sum_j R_{j,k,t}^{\text{neg}}}, \tag{5}$$

where $R_{i,k,t}^{\text{pos}}$ denotes the reading volume of the positive comment i of stock k on day t, $R_{j,k,t}^{\text{neg}}$ denotes the reading volume of the negative comment j of stock k on day t.

3.2. Analysis Approach. We denote investor sentiment, stock return, trading volume, and order imbalance of big trade with S, R, V, and B, respectively. Schmeling [30] assumes that investor sentiment in period $t - 1$ would affect the stock return in period t. In addition, Gross-Klussmann et al. [52] and Xie and Wang [4] also propose similar assumption on daily frequency. So, according to these documents, we analyze the effect of investor sentiment at day $t - 1$ on the stock return, trading volume, and order imbalance of big trade at day t. The equations are as follows:

$$
\begin{aligned}
R_t &= \alpha + \beta \cdot S_{t-1} + \varepsilon, \\
V_t &= \alpha + \beta \cdot S_{t-1} + \varepsilon, \\
B_t &= \alpha + \beta \cdot S_{t-1} + \varepsilon,
\end{aligned} \tag{6}
$$

where S_{t-1} denotes the investor sentiment on day $t - 1$, R_t denotes the stock return on day t, V_t denotes the trading volume on day t, and B_t denotes the order imbalance index of big trade on day t, which is calculated as follows:

$$
B_t = \ln \frac{1 + \text{Buy}_t}{1 + \text{Sell}_t}, \tag{7}
$$

where Buy_t denotes the trading volume of big trade whose direction is to buy on day t, Sell_t denotes the trading volume of big trade whose direction is to sell on day t.

In addition, prior research also argues that the stock market movements would also be influenced by their past [41]; then, the effects of one-day lagged stock return, trading volume, and order imbalance of big trade are also analyzed in this paper. Furthermore, we also include three control variables: book-to-market ratio, BETA, and market value according to past research [5, 10]. The book-to-market ratio is used to measure a company's growth, and irrational investors often pursue short-term speculative gains, while high-growth companies do not conform to their investment logic. BETA is a measure of a stock's risk, that is, how the price of the stock moves relative to the overall market. The market value is the product of the closing price of the stock on the trading day and the total number of shares issued. These three variables would also have effects on stock market movements according to prior research. Thus, the regression equations are as follows:

$$
\begin{aligned}
R_t &= \alpha + \beta \cdot S_{t-1} + \gamma \cdot R_{t-1} + \eta \cdot \text{Controls}_{t-1} + \varepsilon_t, \\
V_t &= \alpha + \beta \cdot S_{t-1} + \gamma \cdot V_{t-1} + \eta \cdot \text{Controls}_{t-1} + \varepsilon_t, \\
B_t &= \alpha + \beta \cdot S_{t-1} + \gamma \cdot B_{t-1} + \eta \cdot \text{Controls}_{t-1} + \varepsilon_t.
\end{aligned} \tag{8}
$$

3.3. Data Source. The online reviews are collected from EM (http://www.eastmony.com), which is one of the most popular financial online forums in China. Using the developed crawler, we get 12,280,974 online comments related to the constituents of CSI300. We classify the comments into three classes, that is, positive, negative, and neutral groups. Then, we calculate the investor sentiment of each stock at every trading day according to equation (5). We also get transaction data of these stocks, including trading volume,

close price, stock return, big trade, market value, BETA, and book-to-market ratio from CSMAR. Finally, after removing records with missing values, we get 105,937 records of 125 stocks, ranging from August 1, 2014, to June 15, 2018. And the descriptive statistics of the key variables are shown in Table 2.

The relationships of the variables could be displayed visually; for example, Figure 3 shows the historical trends of investor sentiment, stock return, trading volume, and order imbalance of large trade of SH600036.

4. Analysis Results

4.1. Stationarity Test. In order to avoid spurious regression, we use the ADF test to investigate the stationarity of the variables. The test results in Table 3 show that all variables are stationary time series and there is no unit root, so regression test can be carried out further.

Moreover, we provide a correlation analysis of variables, which is shown in Table 4. We can see a significant correlation between online investor sentiment and stock return, trading volume, and order imbalance of large trade. In addition, in order to avoid multicollinearity problem between variables, we use the variance inflation factor (VIF) to determine whether there is- a multicollinearity problem in our OLS regression model. According to the collinearity diagnostic criteria of Hair (1995), when $0 < \text{VIF} < 10$, there is no multiple collinearity. Our results show that the maximum VIF is 1.12, which is far less than 10, indicating that our model has no serious multicollinearity problem.

4.2. The Effect of Investor Sentiment on Stock Return. There is evidence that shows investor sentiment has an effect on stock return, shown in Table 5, from which we can see that the investor sentiment on day $t - 1$ has a significant effect on the stock return on day t, with the coefficient equalling 0.0011 ($p = 0.000 < 0.01$). After controlling the stock return, market value, book-to-market ratio, and BETA, the coefficient of investor sentiment is also significant, equalling 0.0009 ($p = 0.000 < 0.01$). Our finding that investor sentiment has a positive effect on stock return is similar to the findings of Hu and Tripathi [32] and Bouteska [33]. The regression results also show that market value, book-to-market ratio, and stock return on day $t - 1$ have significant effects on stock return on day t.

4.3. The Effect of Investor Sentiment on Trading Volume. The analysis results also show that investor sentiment has a significant relation to trading volume, as shown in Table 6. We can see that investor sentiment on day $t - 1$ has a significant effect on the trading volume on day t, with a coefficient equalling 0.0154 ($p = 0.000 < 0.01$). After adding lagged market value, lagged book-to-market ratio, lagged BETA, and lagged trading volume into the regression equation, the coefficient of investor sentiment on day $t - 1$ is also significant, equalling 0.0012 ($p = 0.022 < 0.05$). Impressively, when we add one-day lagged trading volume to

Table 2: Descriptive statistics.

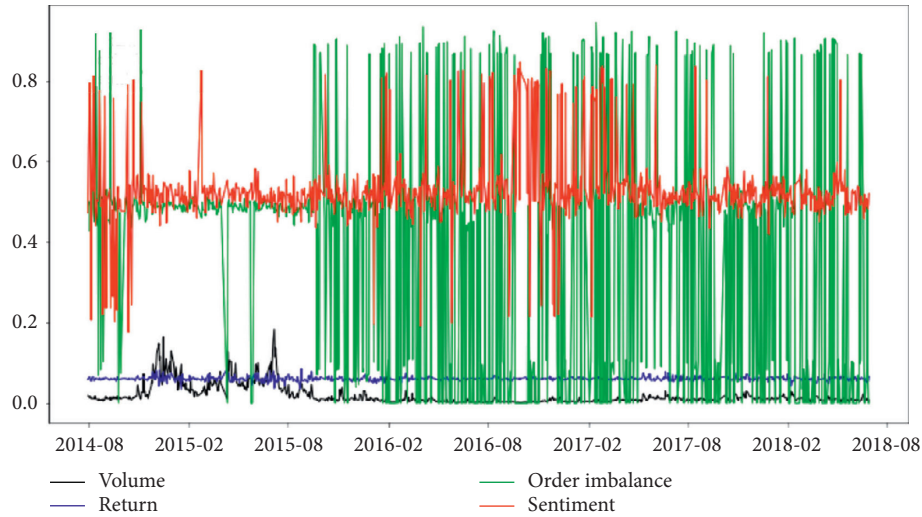| Variables	Mean	Std	Min	Max
Sentiment (S)	0.1611	2.3826	−15.0960	15.0822
Volume (V)	114464	173806.1	1554.3	2089495
Stock return (R)	0.0009	0.0274	−0.1014	0.1020
Order imbalance (B)	−0.3481	5.5534	−16.5944	16.7762
Market value (MV)	114464	173806.1	1554.3	2089495
Book-to-market ratio (BMR)	0.5039	0.3618	0	2.2120
BETA	1.0280	0.5042	−2.3719	5.4902

Figure 3: Historical trend of SH600036 (normalized).

Table 3: Results of ADF test for stationarity.

Variables	ADF test statistic	Test critical values			p value
		1% level	5% level	10% level	
Sentiment	−47.42181	−3.430243	−2.861377	−2.566723	0.0001
Return	−42.34230	−3.430243	−2.861377	−2.566723	0.0000
Volume	−16.62150	−3.958120	−3.409844	−3.126628	0.0000
Order-imbalance	−79.42494	−3.430243	−2.861377	−2.566723	0.0001
Book-to-market ratio	−11.85718	−3.430243	−2.861377	−2.566723	0.0000
Market value	−8.817231	−3.430243	−2.861377	−2.566723	0.0000
Beta	−20.33678	−3.430243	−2.861377	−2.566723	0.0000

the regression equation, the R-squared significantly rises to 82.41%, much higher than before. In addition, market value and BETA on day $t − 1$ also have significant effects on trading volume on day t while book-to-market ratio does not.

4.4. The Effect of Investor Sentiment on Order Imbalance of Big Trade. There is also evidence that shows investor sentiment has a significant effect on order imbalance of big trade, as shown in Table 7. We can see that investor sentiment on day $t − 1$ has no significant effect on order imbalance of big trade on day t, with the coefficient equalling 0.1083 ($p < 0.1$). After adding the lagged order imbalance, market value, book-to-market ratio, and BETA into the regression equation, the

coefficient of investor sentiment is still significant, equalling 0.0310 ($p < 0.01$). In addition, the market value, book-to-market ratio, BETA, and order imbalance of big trade on day $t − 1$ also have significant effects on order imbalance of big trade on day t. In order to further confirm whether investor sentiment has an impact on the imbalance of big trading orders, the regression of investor sentiment is eliminated in Case (3), and it is found that R^2 decreases from 20.39 to 20.38, indicating that the inclusion of investor sentiment in the model can improve the goodness of fit of regression. Combined with the above analysis, investor sentiment has a positive impact on the order imbalance of big trade. In other words, the higher the investor sentiment, the more optimistic the investor's expectation to the market and the more obvious the order imbalance of large trade.

TABLE 4: Correlation matrix.

	S	R	V	B	BMR	MV	Beta
S	1.0000						
R	0.0912***	1.0000					
V	0.0425***	0.1031***	1.0000				
B	0.0458***	0.1342***	0.2648***	1.0000			
BMR	0.0171***	−0.0234***	0.0791***	0.1793***	1.0000		
MV	0.0472***	0.0045	0.3019***	0.1211***	0.2649***	1.0000	
Beta	−0.0106***	−0.0063**	0.0668***	0.0814***	−0.1092***	−0.1902***	1.0000

Note. Data period is from August 1, 2014, to June 15, 2018. T-statistics are in parentheses and significance at the 0.01, 0.05, or 0.1 level is indicated by an ***, an **, or an *, respectively.

TABLE 5: Regression results of stock return and investor sentiment.

	Case (1)	Case (2)	Case (3)	Case (4)
Const	0.0599*** (345.06)	0.0607***(235.66)	0.0587***(125.28)	0.0583*** (111.37)
MV_{t-1}		−0.0013*** (−2.48)	−0.0013** (−2.50)	−0.0014** (−2.54)
BMR_{t-1}		0.0015*** (5.00)	0.0016*** (5.32)	0.0016*** (5.29)
$BETA_{t-1}$		−0.0005 (−1.08)	−0.0005 (−1.02)	−0.0005 (−1.02)
R_{t-1}			0.0327*** (5.94)	0.0318*** (5.88)
S_{t-1}	0.0011** (3.15)			0.0009** (2.57)
$R^2(\%)$	0.02	0.03	0.15	0.16
N	83,324	83,324	83,324	83,324

Note. Data period is from August 1, 2014, to June 15, 2018. T-statistics are in parentheses and significance at the 0.01, 0.05, or 0.1 level is indicated by an ∗∗∗, an ∗∗, or an ∗, respectively.

TABLE 6: Regression results of investor sentiment and trading volume.

	Case (1)	Case (2)	Case (3)	Case (4)
Const	0.0072*** (4.87)	−0.0320*** (−3.60)	−0.0030** (−3.24)	−0.0037*** (−3.27)
MV_{t-1}		0.1265*** (3.52)	0.0104** (3.11)	0.0103** (3.11)
BMR_{t-1}		0.0049 (0.29)	0.0021 (1.08)	0.0021 (1.08)
$BETA_{t-1}$		0.0783*** (4.72)	0.0077*** (4.34)	0.0077*** (4.34)
V_{t-1}			0.8722*** (169.15)	0.8721*** (168.52)
S_{t-1}	0.0154*** (4.58)			0.0012* (1.75)
$R^2(\%)$	0.17	10.41	82.41	82.41
N	83,324	83,324	83,324	83,324

Note. Data period is from August 1, 2014, to June 15, 2018. T-statistics are in parentheses and significance at the 0.01, 0.05, or 0.1 level is indicated by an ***, an **, or an *, respectively.

TABLE 7: Regression results of investor sentiment and order imbalance of big trade.

	Case (1)	Case (2)	Case (3)	Case (4)
Const	0.2467*** (15.82)	−0.1271 (−1.61)	−0.0713 (−1.52)	−0.0871* (−1.90)
MV_{t-1}		0.3606** (2.77)	0.2097** (2.65)	0.2083** (2.64)
BMR_{t-1}		0.7253*** (4.10)	0.4506*** (4.27)	0.4506*** (4.27)
$BETA_{t-1}$		0.6650*** (5.02)	0.3851*** (4.96)	0.3856*** (4.97)
B_{t-1}			0.3995*** (22.68)	0.3992*** (22.61)
S_{t-1}	0.1083*** (5.03)			0.0310** (2.25)
$R^2(\%)$	0.09	5.26	20.38	20.39
N	83,324	83,324	83,324	83,324

Note. Data period is from August 1, 2014, to June 15, 2018. T-statistics are in parentheses and significance at the 0.01, 0.05, or 0.1 level is indicated by an ***, an **, or an *, respectively.

4.5. Endogenous Problem. Endogeneity will bias the empirical results. We discuss the possible endogeneity in this paper from the causes of endogeneity problems.

(1) Reverse causality: since this paper uses the online investor sentiment on day $t − 1$ to predict the dependent variable on day t (including stock return,

trading volume, and order imbalance of big trade), the possibility of reverse causality between the dependent variable and independent variable can be excluded to avoid the possible bias caused by this problem.

(2) Measurement error: all the variables in this paper are from the actual measurement data, no proxy variables are used, and the possibility of endogeneity caused by measurement error is also small.

(3) Omission of variables: in this paper, factors of dependent variables contained in the residual term are mainly considered.Because the online investor sentiment on day $t - 1$, although not affected by the dependent variable on day t, is likely to be affected by the dependent variable on day $t - 1$., so dependent variable on day $t - 1$and control variables are gradually added in equation (6) to eliminate the possibility of endogeneity to the greatest extent.

(4) Although variables such as lagged stock return are controlled, there may still be variables that influence both emotions and dependent variables. Therefore, the liquidity of individual stocks on the daily frequency is selected as the instrumental variable of online investor sentiment, and the 2SLS method is used to more accurately measure the impact of online investor sentiment on the dependent variable. The results show that all the instrumental variables have passed the significance level test at level of 5%, which proves that the selected instrumental variables are more appropriate. By comparing 2SLS regression with OLS results in the abovementioned regression, it can be found that the coefficient size and direction of all variables are basically the same, and the significance and coefficient symbol direction of instrumental variables are highly consistent with the aforementioned regression results, which further demonstrates the reliability of empirical results.

4.6. Robustness Check. Considering that different time spans may have an impact on the empirical results, we further reorganize the data of all variables within the limit of one week to analyze whether the impact of online investor sentiment on stock return, trading volume, and order imbalance of big trade is still robust after the expansion of time span. The regression results show that after adjusting the time span to one week, online investor sentiment still has a significant positive impact on the stock return, trading volume, and order imbalance of big trade. This indicates that our research conclusion is robust.

4.7. Discussion. This paper investigates the impact of online investor sentiment on stock movements. The empirical analysis results show that the lagged investor sentiment has a significant positive impact on stock return, trading volume, and order imbalance of big trade. According to the theory of emotional consistency, the influence of positive emotions on people's decision-making behavior should be positive.

Investors in a positive emotion usually make more optimistic decisions, which tend to overestimate the expected return of a company, while underestimating the corresponding risks, and therefore their willingness to invest in the stock market will increase. If the overall market sentiment is positive, OTC investors will also be eager to buy stocks. And if investor sentiment tends to be negative, investors will be more cautious or even avoid investment decisions in a state of negative sentiment, and they are more worried about the risks of stocks and tend to underestimate the value of stocks. In conclusion, when market investor sentiment tends to be positive, it will boost stock prices in the short term, leading to higher stock returns, and vice versa. Therefore, investor sentiment could be seen as a predictor of stock movement to some extent. Our findings imply that online investor forums or social networks can be regarded as information systems that carry investor attitudes and opinions. With more and more investors using online stock forums or social networks, market managers or policymakers should use big data analytics to strengthen the supervision and guidance of online forums or social networks, track the public opinion dynamics, and prevent potential stock market risks.

Meanwhile, the analysis results also show that investor sentiment has a positive association with large trades. Because informed institutional investors or deep-pocketed individual investors often adopt big trade, so the imbalance of big trade could reflect the distribution of information they have and can reflect the flow of main funds. Therefore, the significant effect of investor sentiment on order imbalance of big trade maybe have twofold meanings. Firstly, the informed investors may reveal their private information on the online stock forum. Secondly, the institutional investors may make their decisions by referring to online investor sentiment.

In addition, stock return, trading volume, and order imbalance of big trade are also significantly influenced by their own lagged values. Most notably, the influence of one-day lagged trading volume on trading volume is significantly larger than that of stock return and order imbalance of big trade, which shows that the trading volume in China stock market exhibits high time serial correlation.

5. Conclusions

This paper develops an online investor sentiment index with a machine learning method and analyzes the relations between investor sentiment and stock movements. We can get three points. Firstly, the analysis results show that investor sentiment has a significant effect on stock return. Specifically, with investor sentiment increasing, the stock return on the next day will rise. Secondly, the analysis results also show that investor sentiment has a significant positive impact on trading volume, and lagged trading volume is a good predictor of trading volume itself in the short term. Finally, investor sentiment has some partial effect on the order imbalance of big trade of the next day. Our results confirm the conclusions of previous scholars' researches, indicating

that deriving investor sentiment index from the online stock forum is feasible and useful.

The main contribution of this paper is to explore the impact of online investor sentiment on stock movements through big trade indicator for the first time. Informed traders tend to use large trades to minimize transaction costs and maximize profits from their informed trading activities. To hide individual information, informed traders tend to trade in midsized trades, preferring large trades to block ones. Therefore, large trading can better reflect the trend of major funds than block trading. Then, with the development of social media, online investor forums, and big data analytics, it is of great significance to investigate large trading and capture its pattern by using online investor sentiment derived from stock comments. According to the analysis results, online investor sentiment can be seen as a predictor of the trading behavior of large trading investors and therefore can have great values in many areas. First, for individual investors, investor sentiment can be incorporated into the actual investment portfolio to assist them to make more efficient decisions. Second, for regulatory authorities such as China Securities Regulatory Commission, the establishment of investor sentiment monitoring system is feasible and useful, which can play an important role in improving the optimization of market regulation and promoting the stable and healthy operation of the stock market.

Although we develop an online investor sentiment index with a supervised machine learning method and analyze the relations between investor sentiment and stock movements, our work can be improved in the future. Firstly, in this study, we only use the titles of online comments to develop an investor sentiment index, and, in the future, the whole article of each comment should be analyzed. Secondly, high-frequency data can be used in future studies.

Data Availability

The data used in this study can be accessed via https://github.com/ArronWang77/stock/.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

References

[1] Q. Li, J. Wang, F. Wang, P. Li, L. Liu, and Y. Chen, "The role of social sentiment in stock markets: a view from joint effects of multiple information sources," *Multimedia Tools and Applications*, vol. 76, no. 10, pp. 12315–12345, 2017.

[2] Q. Wang, W. Xu, and H. Zheng, "Combining the wisdom of crowds and technical analysis for financial market prediction using deep random subspace ensembles," *Neurocomputing*, vol. 299, pp. 51–61, 2018.

[3] T. Li, J. van Dalen, and P. J. van Rees, "More than just noise? examining the information content of stock microblogs on financial markets," *Journal of Information Technology*, vol. 33, no. 1, pp. 50–69, 2018.

[4] C. Xie and Y. X. Wang, "Does online investor sentiment affect the asset price movement? evidence from the Chinese stock market," *Mathematical Problems in Engineering*, vol. 2017, Article ID 2407086, 11 pages, 2017.

[5] H. K. Sul, A. R. Dennis, and L. I. Yuan, "Trading on twitter: using social media sentiment to predict stock returns," *Decision Sciences*, vol. 48, no. 3, pp. 454–488, 2017.

[6] T. Renault, "Intraday online investor sentiment and return patterns in the U.S. stock market," *Journal of Banking and Finance*, vol. 84, pp. 25–40, 2017.

[7] N. Oliveira, P. Cortez, and N. Areal, "The impact of microblogging data for stock market prediction: using Twitter to predict returns, volatility, trading volume and survey sentiment indices," *Expert Systems with Applications*, vol. 73, pp. 125–144, 2017.

[8] F. Audrino, F. Sigrist, and D. Ballinari, "The impact of sentiment and attention measures on stock market volatility," *International Journal of Forecasting*, vol. 36, no. 2, pp. 334–357, 2019.

[9] X. Chu, C. Wu, and J. Qiu, "A nonlinear Granger causality test between stock returns and investor sentiment for Chinese stock market: a wavelet-based approach," *Applied Economics*, vol. 48, no. 21, pp. 1915–1924, 2016.

[10] S.-H. Kim and D. Kim, "Investor sentiment from internet message postings and the predictability of stock returns," *Journal of Economic Behavior & Organization*, vol. 107, pp. 708–729, 2014.

[11] Z. Zhu, X. Song, and W. Huang, "Top management team demography, investor sentiment, and the investment levels of listed companies," *Asia-Pacific Journal of Accounting & Economics*, vol. 25, no. 3-4, pp. 481–495, 2018.

[12] C. G. Ding, H.-J. Wang, M.-C. Lee, W.-C. Hung, and C.-P. Lin, "How does the change in investor sentiment over time affect stock returns?," *Emerging Markets Finance and Trade*, vol. 50, no. 2, pp. 144–158, 2014.

[13] V. Singal and J. Tayal, "Risky short positions and investor sentiment: evidence from the weekend effect in futures markets," *Journal of Futures Markets*, vol. 22, 2019.

[14] L. Sun, M. Najand, and J. Shen, "Stock return predictability and investor sentiment: a high-frequency perspective," *Journal of Banking & Finance*, vol. 73, pp. 147–164, 2016.

[15] T. Dimpfl and V. Kleiman, "Investor pessimism and the German stock market: exploring Google search queries," *German Economic Review*, vol. 20, no. 1, 2019.

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

[17] W. Cheng and J. Lin, "Investors' bullish sentiment of social media and stock market indices," *Journal of Management Science*, 2013.

[18] H. Chen, P. De, Y. Hu, and B.-H. Hwang, "Wisdom of crowds: the value of stock opinions transmitted through social media," *Review of Financial Studies*, vol. 27, no. 5, pp. 1367–1403, 2014.

[19] Q. Li, T. Wang, P. Li, L. Liu, Q. Gong, and Y. Chen, "The effect of news and public mood on stock movements," *Information Sciences*, vol. 278, pp. 826–840, 2014.

[20] T. Loughran and B. McDonald, "When is a liability not a liability? textual analysis, dictionaries, and 10-Ks," *The Journal of Finance*, vol. 66, no. 1, pp. 35–65, 2011.

[21] M. Eickhoff and J. Muntermann, "Stock analysts vs. the crowd: mutual prediction and the drivers of crowd wisdom," *Information and Management*, vol. 53, no. 7, pp. 835–845, 2016.

[22] A. García, "Global financial indices and twitter sentiment: a random matrix theory approach," *Physica A: Statistical Mechanics and Its Applications*, vol. 461, pp. 509–522, 2016.

[23] A. Kanavos, N. Nodarakis, S. Sioutas, A. Tsakalidis, D. Tsolis, and G. Tzimas, "Large scale implementations for twitter sentiment classification," *Algorithms*, vol. 10, 2017.

[24] S. Borovkova and I. Tsiamas, "An ensemble of LSTM neural networks for high-frequency stock market classification," *Journal of Forecasting*, vol. 38, no. 6, pp. 600–619, 2019.

[25] J. Zhao, W.-Y. Kuo, and T.-C. Lin, "Does cognitive limitation affect investor behavior and performance? evidence from limit order clustering," *Review of Financial Studies*, vol. 28, pp. 838–875, 2014.

[26] Y. Adachi, M. Masuda, and F. Takeda, "Google search intensity and its relationship to the returns and liquidity of Japanese startup stocks," *Pacific-Basin Finance Journal*, vol. 46, pp. 243–257, 2017.

[27] D. B. Daiane, d. C. Newton, G. Marco, and C. Jéssica, "Personality traits and investor profile analysis: a behavioral finance study," *PLoS One*, vol. 14, 2019.

[28] D. Yan, G. Zhou, X. Zhao, Y. Tian, and F. Yang, "Predicting stock using microblog moods," *China Communications*, vol. 13, no. 8, pp. 244–257, 2016.

[29] W. Y. Lee, C. X. Jiang, and D. C. Indro, "Stock market volatility, excess returns, and the role of investor sentiment," *Journal of Banking & Finance*, vol. 26, no. 12, pp. 2277–2299, 2002.

[30] M. Schmeling, "Investor sentiment and stock returns: some international evidence," *Journal of Empirical Finance*, vol. 16, no. 3, pp. 394–408, 2009.

[31] D. Huang, F. Jiang, J. Tu, and G. Zhou, "Investor sentiment aligned: a powerful predictor of stock returns," *Review of Financial Studies*, vol. 28, no. 3, pp. 791–837, 2015.

[32] T. Hu and A. Tripathi, "Impact of social media and news media on financial markets," in *Proceedings of the Thirty Seventh International Conference on Information Systems*, Dublin, Ireland, December 2016.

[33] A. Bouteska, "The effect of investor sentiment on market reactions to financial earnings restatements: lessons from the United States," *Journal of Behavioral and Experimental Finance*, vol. 24, 2019.

[34] P. Corredor, E. Ferrer, and R. Santamaria, "The impact of investor sentiment on stock returns in emerging markets: the Case of central European markets," *Eastern European Economics*, vol. 53, no. 4, pp. 328–355, 2015.

[35] Y. Jiang, B. Mo, and H. Nie, "Does investor sentiment dynamically impact stock returns from different investor horizons? Evidence from the US stock market using a multi-scale method," *Applied Economics Letters*, vol. 25, no. 7, pp. 472–476, 2018.

[36] Z.-X. Ni, D.-Z. Wang, and W.-J. Xue, "Investor sentiment and its nonlinear effect on stock returns-New evidence from the Chinese stock market based on panel quantile regression model," *Economic Modelling*, vol. 50, pp. 266–274, 2015.

[37] Z. He, L. He, and F. Wen, "Risk compensation and market returns: the role of investor sentiment in the stock market," *Emerging Markets Finance and Trade*, vol. 55, no. 3, pp. 704–718, 2019.

[38] M. Baker and J. C. Stein, "Market liquidity as a sentiment indicator," *Journal of Financial Markets*, vol. 7, no. 3, pp. 271–299, 2004.

[39] M. Baker and J. Wurgler, "Investor sentiment in the stock market," *Journal of Economic Perspectives*, vol. 21, no. 2, pp. 129–151, 2007.

[40] D. Ryu, H. Kim, and H. Yang, "Investor sentiment, trading behavior and stock returns," *Applied Economics Letters*, vol. 24, no. 12, pp. 826–830, 2017.

[41] X. Li, J. A. Hendler, and J. L. Teall, "Investor attention on the social web," *Journal of Behavioral Finance*, vol. 17, no. 1, pp. 45–59, 2016.

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

[43] K. Joseph, M. Babajide Wintoki, and Z. Zhang, "Forecasting abnormal stock returns and trading volume using investor sentiment: evidence from online search," *International Journal of Forecasting*, vol. 27, no. 4, pp. 1116–1127, 2011.

[44] A. Siganos, E. Vagenas-Nanos, and P. Verwijmeren, "Facebook's daily sentiment and international stock markets," *Journal of Economic Behavior & Organization*, vol. 107, pp. 730–743, 2014.

[45] Y. Sun and G. Ibikunle, "Informed trading and the price impact of block trades: a high frequency trading analysis," *International Review of Financial Analysis*, vol. 54, pp. 114–129, 2017.

[46] Z.-M. Wang, C. Chiao, and Y.-T. Chang, "Technical analyses and order submission behaviors: evidence from an emerging market," *International Review of Economics & Finance*, vol. 24, pp. 109–128, 2012.

[47] P.-H. Hung, "Investor sentiment, order submission, and investment performance on the Taiwan Stock Exchange," *Pacific-Basin Finance Journal*, vol. 39, pp. 124–140, 2016.

[48] C. Chiarella, X.-Z. He, L. Shi, and L. Wei, "A behavioural model of investor sentiment in limit order markets," *Quantitative Finance*, vol. 17, no. 1, pp. 71–86, 2017.

[49] Y. Xiao, B. Li, and Z. Gong, "Real-time identification of urban rainstorm waterlogging disasters based on Weibo big data," *Natural Hazards*, vol. 94, no. 2, pp. 833–842, 2018.

[50] S. Kim, H. Park, and J. Lee, "Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: a study on blockchain technology trend analysis," *Expert Systems with Applications*, vol. 152, p. 12, 2020.

[51] D. D. Wu, L. Zheng, and D. L. Olson, "A decision support approach for online stock forum sentiment analysis," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 8, pp. 1077–1087, 2014.

[52] A. Gross-Klussmann, S. Konig, and M. Ebner, "Buzzwords build momentum: global financial Twitter sentiment and the aggregate stock market," *Expert Systems with Applications*, vol. 136, pp. 171–186, 2019.