Data Science in Finance and Economics

http://www.aimspress.com/journal/dsfe

DSFE, 1(3): 235–252.
DOI: 10.3934/DSFE.2021013
Received: 22 August 2021
Accepted: 30 October 2021
Published: 03 November 2021

Research article

What reflects investor sentiment? Empirical evidence from China

Zimei Huang[1] and Zhenghui Li[2,*]

[1] School of Economics and Statistics, Guangzhou University, Guangzhou 510006, China
[2] Guangzhou Institute of International Finance, Guangzhou University, Guangzhou 510006, China

* **Correspondence:** Email: lizh@gzhu.edu.cn.

Abstract: Investor sentiment tends to show systemic bias on the market, and exerts a significant impact on future market fluctuations, which tends to form an amplified feedback effect. This paper selects three different types of data, namely the emotional text data, the volatility of the stock price and the turnover rate, and other multi-index comprehensive data. Then, this paper formulates different types of investor sentiment indexes through different types of data. From fitting effect of three different types investor sentiment, three different types of investor sentiment index and stock price index correlation to compare the reliability of investor sentiment index. The findings are as follows: First, from the perspective of model fitting, the emotional text-based sentiment index performs better and the model is more robust. Second, from the perspective of market correlation, the text-based sentiment index has the strongest correlation with the stock market. Based on these, the investor sentiment index compiled based on emotional text data more fully reflects investor sentiment.

Keywords: investor sentiment; R vine copula; DCC-GARCH

JEL Codes: E22, G30

1. Introduction

Investor sentiment is a thermometer to measure the stock market, which exerts an important impact on predicting market direction (Hengelbrock et al., 2013) and dynamic asset pricing (Labidi and Yaakoubi, 2016; Luo et al., 2021; Yang and Wu, 2019). Investor sentiment reflects investor’s psychological expectation (Dimic et al., 2018), and investor’s psychological factors and behavioral characteristics exerts an important impact on stock prices(Chue et al., 2019). Investor sentiment not only directly affects the decision-making behavior of investors (Yang and Zhou, 2015), but also plays a crucial role in the stock market bubble (Kim et al., 2014; Laborda and Olmo, 2014; Qadan and Aharon, 2019), financial market volatility (Massa and Yadav, 2015; Shu and Chang, 2015) and so on. High investor sentiment results in stock prices fluctuation and volatile stock market. When investor sentiment is low, the stock market tends to return to stability, investors return to confidence. Therefore, research on investor sentiment is conducive to people's better understanding of the law of market operation and stock price volatility, and provides strong support for investors' behavior decision-making and market supervision (Hirshleifer et al., 2020; Molchanov and Stangl, 2018; Stambaugh et al., 2012).

The existing investor sentiment measurement methods are of three categories: survey method, market variable method and text data measurement method. Besides that, Chan et al. (2017) examined the validity of investor sentiment proxies. One strand of literature employed survey method to measure investor sentiment. The survey method collects individuals' views and attitudes on the current or future economic conditions and the trend of the financial market through questionnaires such as telephone, email and so on. Then investigator aggregate these questionnaire results into an index. On the one hand, the survey method can be based on the judgment of investors on the future trend of the stock market. On the other hand, it can also be based on the views or confidence of investors on the future economic and investment prospects. For example, the consumer confidence index of the University of Michigan is a classic representative of the survey method. Although the survey method directly measures investor sentiment, its implementation cost is high, the frequency of constructing sentiment index is low, and the time span is short. Besides that, it cannot reflect the real investors sentiment in the decision-making process. Liston (2016) used data obtained from the American Association of Individual Investors (AAII) survey to construct individual investors sentiment and examined the impact of investor sentiment on sin stock returns.

Another strand of the literature used some of market variable that reflect economic fundamentals to some extent to represent investor sentiment (Baker et al., 2012; Ben-Rephael et al., 2012; DeVault et al., 2019; Huang et al., 2015). Baker and Wurgler (2006) formed a composite investor sentiment index. To elevate individual proxies arbitrarily and to iron out idiosyncratic variation, Baker and Wurgler (2007) commented on several sentiment proxies and then choose trading volume, the dividend premium, the closed-end fund discount, the number and first-day returns on Initial Public Offerings (IPOs) and the equity share in new issues six indicators, then employed principal components analysis method to construct investor sentiment index. Blau (2016) used the same sentiment proxies to construct investor sentiment index to support that more optimism among investors may strengthen investors' skewness preferences. Unlike sentiment measure of Baker and Wurgler, Kim and Ryu (2020) selected relative strength index (RSI), psychological line index (PLI), trading volume (TV), and adjusted turnover ratio (ATR), employed their own firm-specific sentiment measure to capture firm-level characteristics and illuminate the trading behavior of each investor group. Besides that, Sturm (2014) proposed a turning point method to measure investor sentiment. He defined and tested a turning point methodology that investors use prior highs and lows in prices as reference points from which to make their trading decisions. The results indicated that turning points do have value and the turning point methodology effectively captures investor sentiment.

Besides that, a large amount of literature focuses on text data to measure investor sentiment (Chan et al., 2017; Gao et al., 2019; García, 2013). Tetlock (2007) constructed an investor sentiment index by analyzing daily variation in the Wall Street Journal's "Abreast of the Market" column span from 1984 to 1999. They found that measures of media content can serve as a proxy for investor sentiment. DA et al. (2011) used search frequency in Google SVI to construct a new measure of investor attention.

They found that SVI is correlated with but different from existing proxies of investor attention, which measures the attention of retail investors and captures investor attention in a more timely fashion. Da et al. (2015) used daily Internet search volume from millions of households and constructed a Financial and Economic Attitudes Revealed by Search (FEARS) index to reveal investor sentiment. The results are broadly consistent with theories of investor sentiment, which found that FEARS can predict short-term return reversals, temporary increases in volatility, and mutual fund flows out of equity funds and into bond funds. Based on internet searches in Google and Baidu, Amstad et al. (2020) proposed a new Covid-19 risk attitude (CRA) index for 61 markets and found that CRA index does a good job at capturing investors' attitudes toward pandemic-related risks. Furthermore, Qadan and Nama (2018) compared the investor sentiment index that captured by nine proxies: the adjusted version of Baker and Wurgler's Sentiment Index, the Economic Policy Uncertainty Index, the Financial Stress Index, the Volatility Index (VIX), the Oil Volatility Index (OVX), the Conference Board's Consumer Confidence Index (CCI), the University of Michigan's Consumer Sentiment Index (CSI), the American Association of Individual Investors' Sentiment Survey (AAII) with Google's search volume index (SVI), they found that daily search query data from Google Trends can establish oil shocks Granger-cause the attention of retail investors.

Our empirical analysis contributes to the extant literature on two folds. On the one hand, this paper constructs different types of investor sentiment indexes through different types of data and find that the text-based sentiment index performs better and the model is more robust than the other two investor sentiment indexes from the perspective of model fitting. On the other hand, this paper compares the relationship between three different types of investor sentiment index and stock price index. The results indicate that the text-based sentiment index has the strongest correlation with the stock market from the perspective of market correlation.

The remainder of the paper is organized as follows. Section 2 describes the investor sentiment measurement based on different types of data. Then, Section 3 presents fitting characteristics comparison of different investor sentiment. Comparison of correlation characteristics of different investor sentiment was presented in Section 4. Section 5 is conclusion.

2. Investor sentiment measurement based on different types of data

In this section, investor sentiment measurement methods based on different types of data are introduced. We divided investor sentiment measurement methods into three types: based on emotional text data of investor sentiment (ET), based on the range volatility data of stock index price (RV) and multi-index comprehensive index (MC).

2.1. Measurement based on text data of investor sentiment

First, we construct investor sentiment index based on emotional text data. The key to the construction method is to select the appropriate search keyword set, which can accurately and comprehensively reflect the psychological characteristics of investors (Kruse, 2020). Media Coverage can affect investor sentiment (Zou et al., 2018). In this paper, Baidu trend was selected to investigate the trend of China's investor sentiment. At present, the most widely used network search data is Google search data. However, in China, due to the influence of network restrictions and habits, the most widely used search engine is Baidu. Therefore, the application of Baidu search index to study the trend of investor sentiment in China is more in line with the actual situation. Based on the search volume of Internet users in Baidu, Baidu index take keywords as the statistical object, analyze and calculate the weighted sum of search frequency of each keyword in Baidu web page search. Therefore, using Baidu index is appropriate to construct investor sentiment index of China.

Three steps are needed when construct investor sentiment index based on text data. First is selecting keywords. On the one hand, keywords are required to have a high correlation with the stock market. On the other hand, it is necessary to keep the keywords in a period of time, and rich in change to realize the dynamic monitoring of the stock market. According to these requirements, we select 43 keywords related to investor sentiment in the stock market. The keywords of Baidu index are presented in Appendix. The emotional types are of two categories: positive emotion and negative emotion.

Second step is to analyze word frequency of related keywords. In view of Baidu in China search engine market holds absolute advantage, this paper intends to use the Baidu index data of each keyword as the index of the number of searches of each keyword. According to the list of keywords, input each keyword into the Baidu Index to view the time series data of the search volume of this keyword. Because Baidu index does not provide the downloading function of search data, this paper obtains the daily data time series of keywords in batches based on the crawler program written.

Finally, composite investor sentiment index. The Baidu indexes of each positive and negative keyword were used as the score of their positive and negative sentiment. The total score of positive sentiment and negative sentiment are the sum of the Baidu index of positive emotion keywords and the sum of the Baidu index of negative emotion keywords, respectively. Then divide the total positive sentiment score by the total negative sentiment score and subtract 1 to get the final investor sentiment index. The final investor sentiment index based on text data (ET) can be expressed as follows.

$$\mathrm{ET} = \frac{S_{pos}}{S_{neg}} - 1 \tag{1}$$

where S_{pos} represents the total score of positive sentiment and S_{neg} denotes the total score of negative sentiment. ET greater than 1 indicates that positive sentiment is higher than negative sentiment. While ET less than 1 indicates that negative sentiment is higher than positive sentiment.

The data comes from Baidu index official website, which span from January 04, 2011 to May 21, 2021.

2.2. Measurement based on the range volatility data of stock price index

The measurement based on the range volatility data of stock price index is the spread of Shanghai Securities Composite Index times the turnover rate. The spread of Shanghai Securities Composite Index is used to measure market liquidity. Under normal circumstances, the smaller the spread, the higher the liquidity. Turnover rate refers to the frequency of stock turnover in the market within a certain period of time, which is one of the indicators reflecting the strength of stock liquidity and the one of most important technical indexes to reflect the market trading activity. A high turnover rate generally means that the stock is liquid and easy to get in and out of the market. The higher the turnover rate of a stock, the more actively the stock is traded and the more willing people are to buy. On the contrary, the lower the turnover rate of a stock, the less people pay attention to the stock. In other word is Herding. During herding, investors tend to ignore their private information or beliefs in favor of imitating the behavior of other investors, whether this behavior is rational or not (Chang et al., 2019).

Combining the turnover rate with the trend of stock prices, we can make certain predictions and judgments about the future stock prices. There is a certain internal relationship between the rise and fall of stock prices and the size of their trading volume. Volume and price in the same direction refers to the stock price and volume of the same direction of change. The rise in stock price and the rise in trading volume is a sign that the market continues to look good. Stock prices fell, the volume of the subsequent reduction, indicating that the seller is optimistic about the future, hold the position to sell. Volume-price divergence refers to the opposite trend between the stock price and volume. The stock price rises while the trading volume decreases or stays the same, indicating that the rising trend of the stock price is not supported by the trading volume, and this rising trend is difficult to maintain. A fall in stock prices but a rise in trading volume is a harbinger of a downturn, indicating investors are selling out in fear of disaster. This is in line with the results of Liu (2015), which found that stock market is more liquid when sentiment indices rise and market trading volume also increases when investor sentiment is higher. Therefore, investors can analyze the relationship between them, judge the stock situation to decide buy or sell stocks. To some extent, this indicator reflects investor sentiment. The data comes from Wind database.

2.3. Multi-index comprehensive index measurement

In this paper, four potential variables including price to earnings ratio, turnover rate, closed-end fund discounts and premiums rate and consumer confidence index are selected to construct investor sentiment index. The reasons for selecting these potential variables are as follow. The higher the price to earnings ratio is, the more optimistic investors are about the trend of the current securities market and the higher their mood is. The stocks with higher market turnover rate have higher liquidity, which can reflect the investors' desire to trade them, and also have a positive relationship with investor sentiment. Closed-end fund discounts and premiums rate represents the difference between the net asset value of a fund's actual security holdings and the fund's market price (Baker and Wurgler, 2007), which can reflect investor sentiment to some extent. When retail investors are bearish, the discount rate increases. The consumer confidence index represents the willingness of consumers to consume goods and services. Stock market investors prefer the consumer confidence index with upward growth.

The data of consumer confidence index are obtained from Eastmoney website. The data of price to earnings ratio, turnover rate, closed-end fund discounts and premiums rate are come from Wind database. The data span from January 04, 2011 to April 30, 2021, a total of 2511 days of data.

After selecting these potential variables of investor sentiment index, we need to construct the investor sentiment index through R vine Copula. Vine Copula has widely applied to financial field, especially investigate the nonlinear relationship of multi-dimensional variables. The common C vine and D vine require a specific dependence between the variables. However, Regular Vine proposed by Bedford and Cooke (2001), which reflects the dependent structural relationship of multi-dimensional variables through the minimum or maximum spanning tree structure diagram, does not have strict requirements on the data dependence. R vine is highly practical for describing the complex correlation between financial sequences. This can make up for the deficiency of the traditional linear construction method of investor sentiment index. In addition, Copula function is essentially a connection function, which connects the edge distributions of multiple single variables to obtain a joint distribution, and then analyzes the correlation among multiple variables as a whole. R vine Copula method can flexibly describe the correlation between multi-dimensional variables. Through different Copula functions, the correlation between variables with different correlation structures is studied. Therefore, Copula theory should be more predictable in both the construction of investor sentiment index and the characterization of the nonlinear spillover effect between investor sentiment and stock market returns.

Next, selecting price to earnings ratio (PER), turnover rate (TR), closed-end fund discounts and premiums rate (CEFDPR) and consumer confidence index (CCI), we utilize R vine copula method to construct investor sentiment index. Given the four potential variables of investor sentiment index be $X = \{PER, TR, CEFDPR, CCI\}$, and let the marginal density of the K_{th} potential variable X_k be $f_k(k = 1, \cdots, 4)$. The R vine distribution of four potential variables can be indicated by the joint probability density function $f(PER, TR, CEFDPR, CCI)$ of the random vector $X = \{PER, TR, CEFDPR, CCI\}$ in Formula (2). Formula (2) can be expressed as follows:

$$f(PER, TR, CEFDPR, CCI) = (\prod_{k=1}^{4} f_k(x_k)) \cdot \\ (\prod_{i=1}^{3} \prod_{e \in E_i} c_{j(e),k(e)|D(e)} (F(x_{j(e)}), F(x_{j(e)}|x_{D(e)}))), \tag{2}$$

where E_i is the set of three edges in undirected tree structure composed of four potential variables. $e = j(e), k(e)|D(e)$ denotes an edge in E_i, $c_{j(e),k(e)|D(e)}$ represents the corresponding Pair Copula connecting function. $j(e)\ and\ k(e)$ are the two conditional nodes connected to edge e. $D(e)$ denotes a condition set, that is, when the third potential variable exists, the interaction relationship between two potential variables will change accordingly.

The first step is to use ARMA-GARCH-partial T model to find the optimal edge distribution function (Li et al., 2019). The estimated results of ARMA-GARCH (1,1)-t model are represented in Table 1.

From Table 1, we can know that Copula edge distribution modeling is reasonable. The presupposes of Copula modeling is that the time series does not have heteroscedasticity and autocorrelation, so except for LB test of TR and CEFDPR, the P value of LB, LB2 and ARCH-LM test is basically insignificant (greater than 5%), which can prove that the modeling of Copula edge distribution is reasonable.

Then, vine Copula method is employed to construct the investor sentiment index. And the final investor sentiment index needs to be calculated according to the joint distribution of four variables. Among many combinations, the combination with the highest overall dependence was more accurate in depicting investor sentiment. Therefore, the maximum spanning tree algorithm was utilized to select the R-Vine structure diagram that maximized the absolute sum of Kendell of each layer of tree. The R-Vine structure diagram are shown in Figure 1.

Figure 1 shows the R vine structure diagram of each variable, from which it can be seen that turnover rate, closed-end fund discounts and premiums rate and consumer confidence index are indirectly related through price to earnings ratio.

After the structure diagram is obtained, the optimal connection function is selected for each edge according to AIC criterion, and the distribution of the third layer tree is the joint distribution of the four indexes. The investor sentiment index can be calculated according to the joint distribution.

Table 1. Estimated results of ARMA-GARCH (1,1)-t model.

	PER	TR	CEFDPR	CCI
mean				
ϕ_0	21.945***	7.227***	0.000	100.281***
	(136.800)	(18.819)	(1.369)	(800.553)
ϕ_1	0.100***	0.976***	—	2.007***
	(2618.525)	(185.33)		(7958.908)
ϕ_2	—	—	—	−1.007***
				(−6940.045)
φ_1	—	-0.468***	0.030*	0.552***
		(-21.890)	(1.650)	(37.479)
φ_2	—	−0.098***	—	0.179***
		(−5.211)		(24.564)
φ_3	—	−0.038*	—	—
		(−1.922)		
φ_4	—	−0.047**	—	—
		(−2.481)		
variance				
ω	0.001***	0.031**	0.000	0.000***
	(3.874)	(2.324)	(0.311)	(7.125)
α	0.082***	0.072***	0.066	0.957***
	(4.919)	(4.628)	(1.149)	(18.667)
β	0.905***	0.913***	0.925***	0.042***
	(60.043)	(45.886)	(16.943)	(2.813)
skew	0.954***	1.532***	0.952***	1.070***
	(42.938)	(25.300)	(34.740)	(22.762)
shape	3.291***	5.396***	5.107***	3.509***
	(13.669)	(9.078)	(3.778)	(44.249)
LogLike	921.368	−4119.569	8221.902	7571.889
LB	3.822	32.700	6.509	1.071
	[0.255]	[0.000]	[0.030]	[1.000]
LB2	0.083	9.142	4.411	0.078
	[1.0]	[0.076]	[0.521]	[1.000]
ARCH-LM	0.059	7.338	4.158	0.055
	[0.999]	[0.073]	[0.324]	[0.999]

Notes: Parameter estimates are shown in the table above, and t statistics are shown in brackets. *, ** and *** indicate that the significance level of 10%, 5% and 1%, respectively. LB and LB2 represent logarithmic likelihood values and Ljung-box statistics of sequence correlation in the residual model and square residual model, respectively. ARCH-LM refers to Engle's LM test of ARCH effect in residual error, and P values are shown in square brackets.

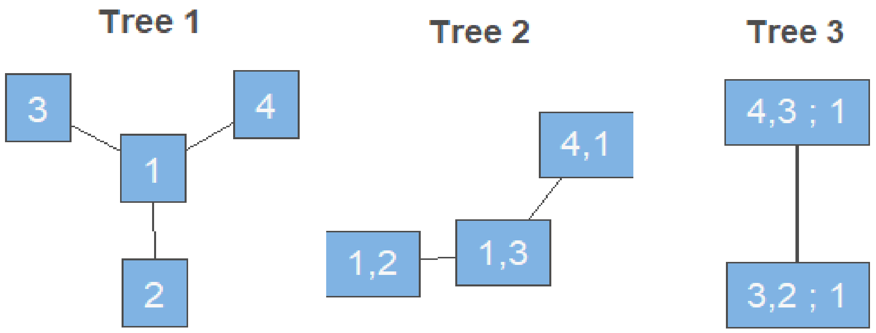

Figure 1. R vine structure diagram. Notes: 1, 2, 3 and 4 represent price to earnings ratio, turnover rate, closed-end fund discounts and premiums rate and consumer confidence index, respectively.

2.4. Descriptive statistics

Table 2 reports the descriptive statistics for three investor sentiments. The minimum value of ET was −93, the maximum value was 123.890, and the mean value was −27.4231, indicating that investor sentiment fluctuates greatly, which is in line with the changing trend of investor sentiment in the stock market. The minimum value of RV was 0.485, the maximum value was 103.543, and the mean value was 4.755. The maximum value of MC was 60.147, the minimum value was 0.028, and the mean value was 4.491. The descriptive statistics indicate that investor sentiment based on ET method is more modest than the other two measurement methods.

Table 2. Descriptive statistics of three investor sentiment.

Variable	Obs.	Mean	Std. Dev.	Min.	Max.
ET	2523	−27.423	13.846	−93.000	123.890
RV	2523	4.755	8.319	0.485	103.543
MC	2511	4.491	7.732	0.028	60.147

Notes: ET is investor sentiment based on emotional text data; RV represents investor sentiment based on the range volatility data of stock index price; MC denotes investor sentiment based on multi-index comprehensive index.

3. Fitting characteristics comparison of different investor sentiment

This section provides a precise description of fitting characteristics comparison of three different investor sentiments. In Subsection 3.1, we introduce the Artificial Neural Networks (ANN) model and Long Short Term Memory (LSTM) model. Then, we utilize three scientific evaluation indicators of time series prediction to assess the fitting performance under three investor sentiment measurement methods in Subsection 3.2. Finally, fitting effects of three Investor Sentiment measurement methods are compared in Subsection 3.3.

3.1. Model construction

To compare the fitting performance of three different investor sentiments, we choose ANN model and LSTM model to evaluate their performance.

ANN model is a kind of non-programmed, adaptive and brain-style information processing, which simulates the processing of complex information by the nervous system of the human brain. Its essence is to obtain a parallel and distributed information processing function through the transformation and dynamic behavior of the network, and imitate the information processing function of the human brain nervous system at different degrees and levels. Firstly, ANN model can map the nonlinear relationship between input and output. ANN model can discover complex changing trend of investor sentiments. Secondly, artificial neural networks model can cope with optimum long-term investor sentiments change in noisy, uncertain, and complex stock market environments. Therefore, as a nonlinear method in the field of artificial intelligence, ANN model can deal with nonlinear, discontinuous and high-frequency multidimensional data, and can be used in investor sentiments changing trend forecasting. ANN has been commonly used in forecasting price movements (Pabuçcu et al., 2020).

LSTM neural network model is an advanced kind of deep machine learning neural network based on recurrent neural network (RNN) model, which is proposed by Hochreiter and Schmidhuber (1997). LSTM model has a good memory for past information of investor sentiments and can eliminate irrelevant data. Besides that, LSTM model is capable of handling the gradient vanishing and gradient explosion problems faced by RNN and long-term dependencies. LSTM neural network propose a more flexible learning process for the feedback error attribution of RNN neural network, which establishes a long time delay between input, feedback and gradient outbreak prevention. This architecture forces its internal state in a particular memory unit to maintain a constant stream of errors, which will not quickly enter the local optimal solution, so that gradients can neither explode nor disappear. LSTM neural network model can fully extract historical information of investor sentiments, and consider the characteristics of current data information of investor sentiments. Furthermore, LSTM neural network model has great advantages in mining the long-term dependence of investor sentiments sequence data and can automatically search for nonlinear features and complex patterns of investor sentiments. Therefore, we utilized LSTM neural network model to compare the fitting performance of three investor sentiments.

3.2. Evaluation metrics of model fitting effect

In this paper, three scientific evaluation indicators of time series prediction are selected to assess the fitting performance under three investor sentiment measurement methods. The selected scientific evaluation indicators are mean squared error (MSE), mean absolute error (MAE) and standard deviation of absolute percentage error (SDAPE). The mathematical formulas of these evaluation metrics are as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i(t) - \hat{y}_i(t))^2 , \tag{3}$$

$$AE = \frac{1}{n}\sum_{i=1}^{n}|y_i(t) - \hat{y}_i(t)| , \tag{4}$$

$$SDAPE = \sqrt{\left(\frac{1}{n}\right)\sum_{i=1}^{n}\left(\left|\frac{\hat{y}_i(t) - y_i(t)}{y_i(t)}\right| - MAPE\right)^2}, \tag{5}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i(t) - y_i(t)}{y_i(t)}\right| * 100\%, \tag{6}$$

where n is the number of investor sentiment in the training set, $\hat{y}_i$ denotes the predictive value of investor sentiment and y_i is the true value of investor sentiment.

The performance of investor sentiment measurement methods was evaluated by these three scientific evaluation indicators. The value range of MSE is [0,+∞). When the predicted value is exactly consistent with the real value, MSE is equal to 0, that is, the investor sentiment measurement methods performs well. The larger the error, the greater the MSE value. So MAE are the same. Furthermore, the value of SDAPE is in the interval [0, 1]. The smaller values of SDAPE represent better performance of investor sentiment measurement methods.

3.3. The fitting effects comparison of different investor sentiment

The performance of three investor sentiment measurement methods can be expressed in Table 3. Columns 3, 4 and 5 in Table 3 reports mean squared error (MSE), mean absolute error (MAE) and Standard Deviation of Absolute Percentage Error (SDAPE) of fitting effect of three investor sentiment measurement methods, respectively.

Table 3. Comparation of the fitting effects of three indicators.

Measurement Indicator	Fitting Method	MSE	MAE	SDAPE
ET	ANN	0.002	0.033	0.026
	LSTM	0.006	0.053	0.040
RV	ANN	0.072	0.058	0.055
	LSTM	0.122	0.039	0.041
MC	ANN	0.025	0.103	0.125
	LSTM	0.031	0.148	0.111

Notes: ET is investor sentiment based on emotional text data; RV represents investor sentiment based on the range volatility data of stock index price; MC denotes investor sentiment based on multi-index comprehensive index.

It can be concluded from Table 4 that among the fitting effects of the three investor sentiment measurement indexes, investor sentiment index based on emotional text data performs best, followed by the investor sentiment index based on the range volatility data of stock index price and multi-index comprehensive index. Specifically speaking, firstly, from the aspect of MSE, the measurement result of ET is the smallest. The MSE of ANN and LSTM neural network model are 0.002 and 0.006, respectively. Followed by the MSE of MC, are 0.025 and 0.031, respectively. The MSE of RV was the largest, which are 0.072 and 0.122, respectively. Secondly, in terms of forecast deviation, according to the MAE, when ANN model is used, MAE of ET is the smallest; when LSTM model is used, MAE of RV is the smallest, followed by MAE of ET. Therefore, we believe that ET index performs better.

Thirdly, from the perspective of prediction accuracy, according to the SDAPE, as for ET, both the SDAPE of ANN and LSTM neural network is smallest, which are 0.026 and 0.040, respectively. Then, as for RV, SDAPE of ANN and LSTM neural network are 0.055 and 0.041, respectively. The SDAPE of ANN and LSTM neural network based on MC is the largest, which are 0.125 and 0.111, respectively. Therefore, we believe that in terms of skewness and accuracy, both ANN and LSTM neural network, the estimation results of ET are better than those based on RV and MC.

4. Comparison of correlation characteristics of different investor sentiment

In this section, we further investigate the correlation characteristics of different investor sentiment. First, we introduce dynamic conditional correlational autoregressive Conditional Heteroscedasticity (DCC-GARCH) model in Subsection 4.1. Second, Subsection 4.2 utilize DCC-GARCH model to analyze the dynamic correlation between investor sentiment and Shanghai Composite Index.

4.1. DCC-GARCH model

DCC-GARCH model can describe the dynamic correlation between time series well. DCC-GARCH model measures the dynamic correlation of two or more different time series data, which offers flexibility to simultaneously model the multivariate conditional volatility of financial assets and their time-varying linkage (Dong et al., 2019). DCC-GARCH model, which was proposed by Engle (2003), is an improved CCC-GARCH model. The DCC-GARCH model relaxes the assumption that the fluctuation coefficient of the correlation of time series data is constant in the CCC-GARCH model, and considers that the fluctuation of time series data is time-varying in actual circumstances. On the theoretical basis of CCC-GARCH model, a variable conditional coefficient is added. It is believed that the investor sentiment time series is time-varying, the fluctuation of the current period is related to the fluctuation of the previous period, and the fluctuation will change with the change of time, and the conditional correlation coefficient is dynamic. Therefore, DCC-GARCH model can realize the dynamic correlation analysis between investor sentiment index and Shanghai Composite Index.

The form of the DCC-GARCH model is as following:

$$\begin{gathered} \gamma_t = \mu_t + \varepsilon_t, \\ \varepsilon_t \sim N(0, H_t), \\ H_t = (h_{ii,t}) = D_t R_t D_t, \\ D_t = diag(h_{11,t}^{1/2}, , h_{nn,t}^{1/2}), \\ h_{ii,t} = \omega_i + \sum_{i=1}^{p} \alpha_i \, \varepsilon_{i,t-i}^2 + \sum_{i=1}^{q} \beta_i h_{i,t-1} \end{gathered} \tag{7}$$

where H_t is the conditional covariance matrix, D_t denotes the conditional standard deviation calculated by the univariate GARCH model as the principal diagonal matrix of diagonal elements. R_t represents dynamic correlation coefficient matrix. $h_{ii,t}$ are the elements of D_t. α and β are the coefficient before residual squared term (ARCH term) and the lag term coefficient of the conditional variance (GARCH term) of GARCH model, respectively. α represents the influence of current fluctuation on future fluctuation. The higher the value is, the greater the influence of current fluctuations on future fluctuations will be. β denotes the influence of the fluctuation value of the previous period on the current value. The higher the value is, the easier it is to be affected by its previous value. $\alpha + \beta$ represents the extent to which current volatility trends sustain future volatility. The smaller the value is, the faster the fluctuation trend of the current residual will disappear in the future. The higher the value is, it indicates that the fluctuation trend of current residual will disappear slowly in the future.

$$\begin{gathered}
R_t = diag\{Q_t\}^{-1} Q_t diag\{Q_t\}^{-1}, \\
Q_t = (q_{ij,t}) = (1 - \sum_{m=1}^{M} \alpha_m^* - \sum_{n=1}^{N} \beta_n^*)\bar{Q} + \sum_{m=1}^{M} \alpha_m^* e_{t-m} e_{t-m}' + \sum_{n=1}^{N} \beta_n^* Q_{t-n} \\
e_t = D_t^{-1} \varepsilon_t \\
\bar{Q} = \frac{1}{T} \sum_{t=1}^{T} e_t e_t'
\end{gathered} \tag{8}$$

The dynamic correlation coefficient in the DCC-GARCH model can be expressed as:

$$\begin{aligned}
\rho_{12,t} &= \frac{q_{12,t}}{\sqrt{q_{11,t} q_{22,t}}} \\
&= \frac{(1-\alpha_m^*-\beta_n^*)\overline{q_{12}}+\alpha_m^* e_{1,t-1} e_{2,t-1}+\beta_n^* q_{12,t-1}}{\sqrt{((1-\alpha_m^*-\beta_n^*)\overline{q_{11}}+\alpha_m^* e_{1,t-1}^2+\beta_n^* q_{11,t-1})((1-\alpha_m^*-\beta_n^*)\overline{q_{22}}+\alpha_m^* e_{2,t-1}^2+\beta_n^* q_{22,t-1})}},
\end{aligned} \tag{9}$$

where α^* and β^* are the dynamic correlation coefficient of DCC-GARCH model. α^* represents the influence of the standardized residuals of the previous stage on the dynamic correlation coefficient. β^* denotes the influence of the correlation coefficient of the previous stage on the correlation coefficient of this stage. $\alpha^* + \beta^*$ is the attenuation coefficient of the model, representing the persistence of correlation between two time series. The greater the value, the stronger the persistence of correlation. m and n are the lag orders determined by the univariate GARCH model. $\bar{Q}$ is the unconditional variance matrix of normalized residual sequence e_t of exponential regression equation. e_t denotes the residual after standardization. R_t is the dynamic correlation coefficient matrix.

The estimated steps of DCC-GARCH model are divided into two steps. First, the GARCH model is used to estimate the investor sentiment index and the Shanghai Composite Index respectively, and the standardized residual sequence is obtained. When we determine the order of GARCH model, we usually choose the order according to the minimization principle of Akakike information criterion (AIC) and Bayesian information criterion (BIC). Different GARCH models are used to fit the data, and the model with the minimum AIC and BIC values of each model is selected to model the data. The smaller the values of these two criteria are, the better the model is, and thus the degree of data fitting is high. According to AIC criterion and BIC criterion minimization principle, this paper determines that the lag order p and q of GARCH model are both 1. Then, the standardized residual sequence is used to estimate the dynamic correlation between investor sentiment and Shanghai Composite Index.

4.2. Dynamic correlation results of different investor sentiment

We further employ DCC-GARCH model to investigate the dynamic correlation between investor sentiment index and the Shanghai Composite Index. The estimated results of DCC-GARCH model are shown in Table 4. Columns 2 and 3 report ARCH term and GARCH term, respectively. Columns 4 and 5 in Table 4 reports the alpha and beta value of DCC. Panel A shows the dynamic correlation between ET and Shanghai Composite Index. Panel B and C presents the dynamic correlation between RV, MC and Shanghai Composite Index, respectively.

Table 4. Estimated results of DCC-GARCH model.

Variable	ARCH term	GARCH term	DCC alpha	DCC beta
Panel A: Shanghai Composite Index and ET				
Shanghai Composite Index	0.089^{***}	0.910^{***}	0.003	0.996^{***}
	(0.022)	(0.050)	(0.004)	(0.005)
ET	0.500^{***}	0.387^{***}		
	(0.003)	(0.034)		
Panel B: Shanghai Composite Index and RV				
Shanghai Composite Index	0.236^{***}	0.758^{***}	0.062^{***}	0.389
	(0.041)	(0.031)	(0.019)	(0.269)
RV	0.251	0.817^{***}		
	(1474.046)	(0.021)		
Panel C: Shanghai Composite Index and MC				
Shanghai Composite Index	0.079^{***}	0.932	0.253^{***}	0.402^{***}
	（0.015）	(188.907)	(0.000)	(0.000)
MC	162.835	0.988^{***}		
	(57065.200)	(0.000)		

Notes: The parentheses are the standard errors of the coefficient estimates. *, **, and *** represent the significance level of 10%, 5%, and 1%, respectively. ET is investor sentiment based on emotional text data; RV represents investor sentiment based on the range volatility data of stock index price; MC denotes investor sentiment based on multi-index comprehensive index.

Dynamic correlation relationships exist between investor sentiment index and the Shanghai Composite Index. Firstly, the results of univariate estimation show that the earnings rate of Shanghai Composite Index and ET are affected by conditional variance and prior fluctuations, while RV and MC are mainly affected by prior fluctuations. Secondly, the results of DCC estimation show that the correlation coefficients between the returns of Shanghai Composite Index and ET, RV and MC all show significant dynamic change characteristics.

Specifically, in Panel A of Table 4, the coefficient of DCC beta is significantly positive, while the coefficient of DCC alpha is not significant. It shows that the correlation coefficient between the Shanghai Composite Index and ET presents a significant dynamic change characteristic, and the correlation coefficient is mainly affected by the previous fluctuations. In Panel B, the coefficient of DCC alpha is significantly positive, while the coefficient of DCC beta is not significant. It indicates that the correlation coefficient between the return rate of Shanghai Composite Index and RV presents a significant dynamic change characteristic, and the correlation coefficient is mainly affected by the correlation coefficient of the previous period. The coefficients of DCC alpha and DCC beta in Panel C are both significantly positive. It manifests that the correlation coefficient between the return rate of Shanghai Composite Index and MC presents a significant dynamic change characteristic, and the correlation coefficient is affected by the earlier fluctuation and the earlier correlation coefficient at the same time.

There is a dynamic correlation between ET and the earnings rates of Shanghai Composite Index, and the degree of dynamic correlation is very persistent. Figure 2 shows the changing trend of the correlation coefficient between ET and the return rate of the Shanghai Composite Index under dynamic conditions graphically. As can be seen from Figure 2, the correlation between ET and the return rate of Shanghai Composite Index is not fixed, but changes with time. On the whole, the dynamic relationship between ET and the return rate of Shanghai Composite Index fluctuates greatly, but they are basically positively correlated, indicating a good correlation between ET and Shanghai Composite Index. From the perspective of fluctuation characteristics, the correlation coefficient before 2015 gradually decreased from the highest value of 0.48 to around 0.08 in 2015. In December 2014, the correlation coefficient of dynamic conditions between the two was the smallest, and was much smaller than that in other time periods. In subsequent years, the correlation between the two fluctuated considerably. Sudden events will bring shocks to investor sentiment. But there is a certain period of stability between these shocks, during which investor sentiment will fluctuate less, which will bring different influences to the earnings rates of Shanghai Composite Index. And the correlation between the two will fluctuate for many times. For example, there was a bull market in the first half of 2015, and the SSE index rose rapidly from March of that year. With the stimulation of a series of policies such as supporting emerging industries, One Belt And One Road strategy and reform of state-owned enterprises, investor sentiment continued to rise, and the Shanghai Composite Index broke through the high of 3478 on March 3, 2015. It peaked at 5178 in June. After A four-month surge, the policies of suspending IPO, raising margin ratio and lowering reserve ratio were introduced one after another, investor sentiment fell, and the A-share market began to plunge from July to September. After the sharp rise and fall, the market sentiment fell rapidly, and in September, there were three thousand-share declines. In October, it introduced the bonus tax adjustment, the circuit breaker mechanism, the CFIX lowered the opening limit and increased the intraday turnover. Therefore, ET exists dynamic correlation with Shanghai Composite Index.

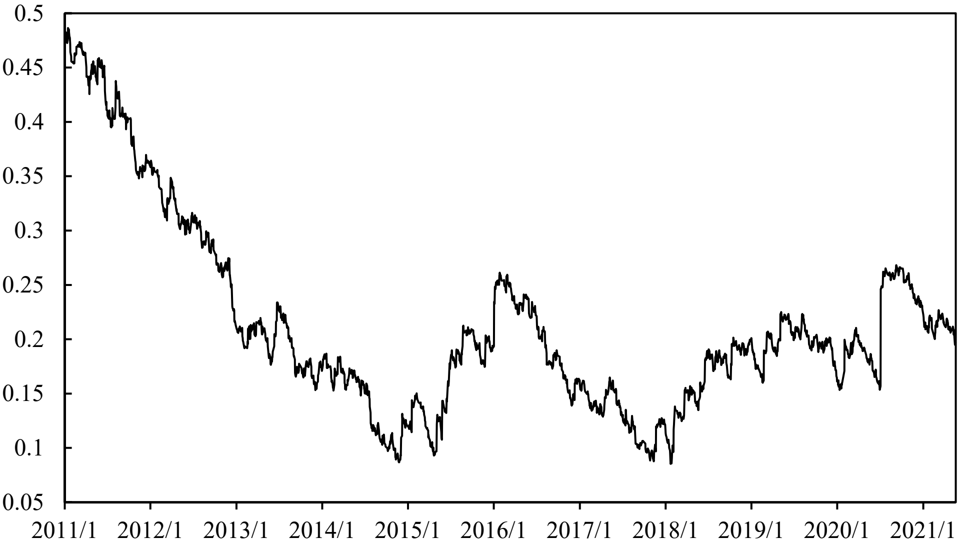

Figure 2. Dynamic conditional correlation coefficient between ET and Shanghai Composite Index.

The dynamic correlation relationship between RV and the earnings rates of Shanghai Composite Index fluctuates greatly and is weak. Figure 3 intuitively shows the changing trend of the dynamic correlation coefficient between RV and the earnings rates of the Shanghai Composite Index under dynamic conditions. As can be seen from Figure 3, in terms of the overall trend, the dynamic relationship between RV and the earnings rates of Shanghai Composite Index also fluctuates greatly, but the correlation between them is weak, indicating that the correlation relationship between RV and the return rate of stock market is poor. From the perspective of volatility characteristics, the correlation coefficient between RV and the earnings rates of Shanghai Composite Index basically fluctuates around zero value.

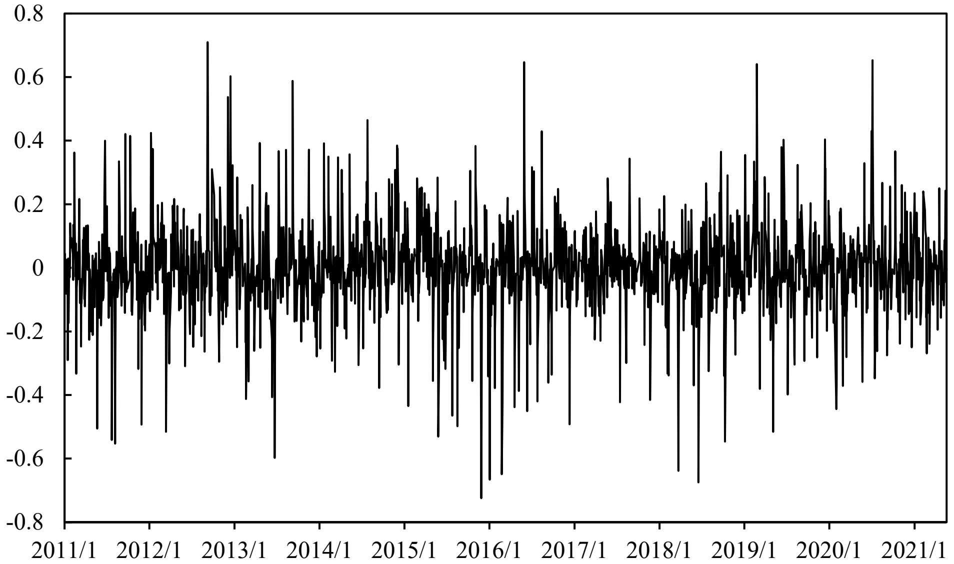

Figure 3. Dynamic conditional correlation coefficient between RV and Shanghai Composite Index.

The dynamic correlation relationship between MC and the earnings rates of Shanghai Composite Index is stable but weak. Figure 4 intuitively shows the changing trend of the correlation coefficient between MC and earnings rates of the Shanghai Composite Index under dynamic conditions. As can be seen from Figure 4, from the overall trend, the dynamic relationship between MC and earnings rates of Shanghai Composite Index is relatively stable, with a correlation coefficient of about 0.08. However, there shows a weak correlation between MC and earnings rates of the Shanghai Composite Index. From the perspective of volatility characteristics, except for a few periods in February 2012, February 2019 and March 2019, the correlation coefficient between MC and earnings rates of the Shanghai Composite Index is basically around 0.08.

Combined with Figures 2–4, it can be seen that the dynamic correlation between ET and the return rate of Shanghai Composite Index is the strongest. In addition, ET investor sentiment comes from the Baidu index, which reflects the trend of investors' attention and their investment behavior. When there is a special event, investor sentiment constructed based on the Baidu index react in a timely manner. Therefore, we believe that the investor sentiment index based on Baidu index is the best.

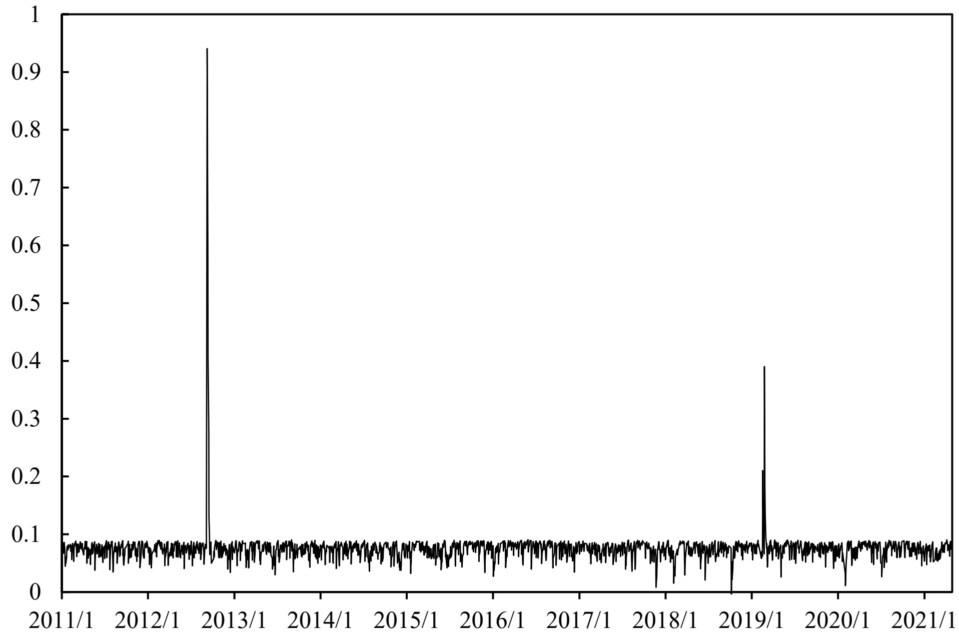

Figure 4. Dynamic conditional correlation coefficient between MC and Shanghai Composite Index.

5. Conclusions

In this paper, three different types of data, namely the emotional text data, the volatility of the stock price, the turnover rate and other multi-index comprehensive data are selected to construct investor sentiment indexes. Then, we compare the fitting effect and dynamic correlation relationship with Shanghai Composite Index. We can draw some conclusions as follows.

Firstly, from fitting effect of three different types investor sentiment, investor sentiment index based on emotional text data performs best and the model is more robust, followed by the investor sentiment index based on the range volatility data of stock index price and multi-index comprehensive index.

Secondly, from the perspective of market correlation, the emotional text-based sentiment index has the strongest dynamic correlation relationship with Shanghai Composite Index. On the one hand, the earnings rate of Shanghai Composite Index and ET are affected by conditional variance and prior fluctuations, while RV and MC are mainly affected by prior fluctuations. On the other hand, there are significant dynamic change characteristics between the returns of Shanghai Composite Index and ET, RV and MC. Based on these, the investor sentiment index compiled based on text data more fully reflects investor sentiment.

Therefore, the sentiment index based on emotional text is more appropriate to reflect investor sentiment of stock market.

Funding

This research was funded by the Innovation Research for the Postgraduates of Guangzhou University, grant number 2020GDJC-D05.

Acknowledgments

Authors would like to thank Guangzhou University for sponsoring this research. Besides that, authors would like to thank the editor and anonymous reviewers for their patient and valuable comments on earlier versions of this paper.

Conflicts of interest

All authors declare no conflicts of interest in this paper.

References

Amstad M, Cornelli G, Gambacorta L, et al. (2020) Investors' risk attitudes in the pandemic and the stock market: New evidence based on internet searches. BIS Bulletin No. 25. Available from: https://ssrn.com/abstract=3654374.

Baker M, Wurgler J (2006) Investor sentiment and the cross-section of stock returns. *J Financ* 61: 1645–1680.

Baker M, Wurgler J (2007) Investor sentiment in the stock market. *J Econ Perspect* 21: 129–152.

Baker M, Wurgler J, Yuan Y (2012) Global, local, and contagious investor sentiment. *J Financ Econ* 104: 272–287.

Bedford T, Cooke RM (2001) Probability density decomposition for conditionally dependent random variables modeled by vines. *Ann Math Artif Intell* 32: 245–268.

Ben-Rephael A, Kandel S, Wohl A (2012) Measuring investor sentiment with mutual fund flows. *J Financ Econ* 104: 363–382.

Blau BM (2016) Skewness preferences, asset prices and investor sentiment. *Appl Econ* 49: 812–822.

Chan F, Durand RB, Khuu J, et al. (2017) The validity of investor sentiment proxies. *Int Rev Financ* 17: 473–477.

Chang CY, Shie FS, Yang SL (2019) The relationship between herding behavior and firm size before and after the elimination of short-sale price restrictions. *Quant Financ Econ* 3: 526–549.

Chue TK, Gul FA, Mian GM (2019) Aggregate investor sentiment and stock return synchronicity. *J Bank Financ* 108: 105628.

Da Z, Engelberg J, Gao P (2015) The sum of all fears investor sentiment and asset prices. *Rev Financ Stud* 28: 1–32.

DeVault L, Sias R, Starks L (2019) Sentiment metrics and investor demand. *J Financ* 74: 985–1024.

Dimic N, Neudl M, Orlov V, et al. (2018) Investor sentiment, soccer games and stock returns. *Res Int Bus Financ* 43: 90–98.

Dong H, Liu Y, Chang J (2019) The heterogeneous linkage of economic policy uncertainty and oil return risks. *Green Financ* 1: 46–66.

Engle RF (2003) Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *J Bus Econ Stat* 20: 339–350.

Gao Z, Ren H, Zhang B (2019) Googling investor sentiment around the world. *J Financ Quant Anal* 55: 549–580.

García D (2013) Sentiment during recessions. *J Financ* 68: 1267–1300.

Hengelbrock J, Theissen E, Westheide C (2013) Market response to investor sentiment. *J Bus Financ Account* 40: 901–917.

Hirshleifer D, Jiang D, DiGiovanni YM (2020) Mood beta and seasonalities in stock returns. *J Financ Econ* 137: 272–295.

Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9: 1735–1780.

Huang D, Jiang F, Tu J, et al. (2015) Investor sentiment aligned: A powerful predictor of stock returns. *Rev Financ Stud* 28: 791–837.

Kim JS, Ryu D, Seo SW (2014) Investor sentiment and return predictability of disagreement. *J Bank Financ* 42: 166–178.

Kim K, Ryu D (2020) Does sentiment determine investor trading behaviour? *Appl Econ Lett* 28: 811–816.

Kruse P (2020) Spreading entrepreneurial news—investigating media influence on social entrepreneurial antecedents. *Green Financ* 2: 284–301.

Labidi C, Yaakoubi S (2016) Investor sentiment and aggregate volatility pricing. *Q Rev Econ Financ* 61: 53–63.

Laborda R, Olmo J (2014) Investor sentiment and bond risk premia. *J Financ Mark* 18: 206–233.

Li T, Zhong J, Huang Z (2019) Potential dependence of financial cycles between emerging and developed countries: Based on arima-garch copula model. *Emerging Mark Financ Trade* 56: 1237–1250.

Liston DP (2016) Sin stock returns and investor sentiment. *Q Rev Econ Financ* 59: 63–70.

Liu S (2015) Investor sentiment and stock market liquidity. *J Behav Financ* 16: 51–67.

Luo C, Li Z, Liu L (2021) Does investor sentiment affect stock pricing? Evidence from seasoned equity offerings in China. *Nat Account Rev* 3: 115–136.

Massa M, Yadav V (2015) Investor sentiment and mutual fund strategies. *J Financ Quant Anal* 50: 699–727.

Molchanov A, Stangl J (2018) Investor sentiment and industry returns. *Int J Financ Econ* 23: 546–570.

Pabuçcu H, Ongan S, Ongan A (2020) Forecasting the movements of bitcoin prices: An application of machine learning algorithms. *Quant Financ Econ* 4: 679–692.

Qadan M, Aharon DY (2019) Can investor sentiment predict the size premium? *Int Rev Financ Anal* 63: 10–26.

Qadan M, Nama H (2018) Investor sentiment and the price of oil. *Energy Econ* 69: 42–58.

Shu HC, Chang JH (2015) Investor sentiment and financial market volatility. *J Behav Financ* 16: 206–219.

Stambaugh RF, Yu J, Yuan Y (2012) The short of it: Investor sentiment and anomalies. *J Financ Econ* 104: 288–302.

Sturm RR (2014) A turning point method for measuring investor sentiment. *J Behav Financ* 15: 30–42.

Tetlock PC (2007) Giving content to investor sentiment: The role of media in the stock market. *J Financ* 62: 1139–1168.

Yang C, Wu H (2019) Chasing investor sentiment in stock market. *N Am J Econ Financ* 50: 100975.

Yang C, Zhou L (2015) Investor trading behavior, investor sentiment and asset prices. *N Am J Econ Financ* 34: 42–62.

DA Z, Engelberg J, Gao P (2011) In search of attention. *J Financ* 66: 1461–1499.

Zou L, Cao KD, Wang Y (2018) Media coverage and the cross-section of stock returns: The chinese evidence. *Int Rev Financ* 19: 707–729.