J. Korean Soc. Ind. Appl. Math. Vol.28, No.3, 71–87, 2024 https://doi.org/10.12941/jksiam.2024.28.071

MF-DCCA ANALYSIS OF INVESTOR SENTIMENT AND FINANCIAL MARKET BASED ON NLP ALGORITHM

RUI ZHANG[1,2,3,4], CAIRANG JIA[1,2,3,4†], AND JIAN WANG[5,6,7†]

[1]SCOOL OF COMPUTER, QINGHAI NORMAL UNIVERSITY, XINING, QINGHAI 810016, CHINA

[2]QINGHAI PROVINCE TIBETAN INFOMATION PROCESSING ENGINEERING TECHNOLOGY RESCARCH CENTER, XINING, QINGHAI 810008, CHINA

[3]KEY LABORATORY OF TIBETAN INFORMATION PROCESSING AND MACHINE TRANSLATION OF QINGHAI PROVINCE, XINING, QINGHAI 810008, CHINA

[4]STATE KEY LABORATORY OF TIBETAN INTELLIGENT INFORMATION PROCESSING AND APPLICATION, XINING, QINGHAI, 810008, CHINA

[5]SCHOOL OF MATHEMATICS AND STATISTICS, NANJING UNIVERSITY OF INFORMATION SCIENCE AND TECHNOLOGY, NANJING, 210044, CHINA

[6]CENTER FOR APPLIED MATHEMATICS OF JIANGSU PROVINCE, NANJING UNIVERSITY OF INFORMATION SCIENCE AND TECHNOLOGY, NANJING 210044, CHINA

[7]JIANGSU INTERNATIONAL JOINT LABORATORY ON SYSTEM MODELING AND DATA ANALYSIS, NANJING UNIVERSITY OF INFORMATION SCIENCE AND TECHNOLOGY, NANJING 210044, CHINA

Email address: 13997089569@163.com, zwxxzx@163.com†, 003328@nuist.edu.cn†

ABSTRACT. In this paper, we adopt the MF-DCCA (Multifractal Detrended Cross-Correlation Analysis) method to study the nonlinear correlation between the returns of financial stock markets and investors' sentiment index (SI). The return series of Shanghai Securities Composite Index (SSEC) of China, Shenzhen Securities Component Index (SZI) of China, Nikkei 225 Index (N225) of Japan, and Standard & Poor's 500 Index (S&P500) of the United States are adopted. Firstly, we preliminarily analyze the correlation between SSEC and SI through the Pearson correlation coefficient. In addition, by MF-DCCA, we observe a power-law correlation between investors' sentiment index and SSEC stock market returns, with a significant multifractal correlation. Besides, SI series and SSEC return series have positive persistence. We compare the differences in multifractal cross-correlation between SI and stock return sequences in different markets. We found that the values of SZI-SI in terms of cross-correlation persistence and cross-correlation strength are relatively close to those of SSEC-SI, while the $H_{xy}(2)$, ΔH_{xy}, and $\Delta\alpha_{xy}$ of N225-SI and S&P500 are much smaller than those of SSEC-SI and SZI-SI. This reason is related to the fact that the investors' sentiment index originated from the Shanghai Composite Index Tieba. The SI is obtained through natural language processing method. Finally, we study the rolling of $H_{xy}(2)$ and $\Delta\alpha_{xy}$. Results indicate that the macroeconomic environment may cause fluctuations in two sequences of $H_{xy}(2)$ and $\Delta\alpha_{xy}$.

Received May 10 2024; Revised August 6 2024; Accepted in revised form September 4 2024; Published online September 25 2024.

2020 *Mathematics Subject Classification.* 40-08.

Key words and phrases. MF-DCCA, sentiment index, word vector, stock price.

†Corresponding author.

1. INTRODUCTION

Financial markets, especially stock markets are often influenced by numerous factors, among which economic factors [1] and non-economic fatcors [2] are two main stream of factors. In the domain of non-economic factors, Investor Sentiment Measures (ISM) [3] are one of the most unneglectable elements. Investor trading behavior, frequently driven by irrational decision-making, tends to result in excessive market transactions which not only fail to yield reasonable returns but often lead to substantial losses [4]. The intrinsic human emotions of greed and fear introduce instability into market dynamics. As investors are prone to emotional influences during decision-making processes, there is a tendency for irrational deviations to occur. Emotional dynamics are increasingly acknowledged as systemic risk factors with the potential to misalign the market from its rational operational trajectory, thus amplifying market risk and volatility [5]. This over-trading phenomenon is of significant concern, not just for the individual investor's capital but also for the overall stability of the stock market.

Investor Sentiment Measures are measures that cannot be straightforwardly quatitatively described. In real world, ISM are often referred as qualitative data, since ISM are reflected by statements, reviews or comments from the customers and investors [3, 6]. To quantitative and analyze the sentiments hidden behind texts and words, experts have brought forward several constructive methods, like lexical analysis, bag-of-words model and Natural Language Processing (NLP). Hippisley [7] systematically introduced lexical analysis and set forth its drawbacks. Zhang et al. [8] established a statistical framework under bag-of-words model that forgoes heuristic clustering for visual word generation, offering competitive empirical performance and two new algorithms that maintain efficacy in object categorization without relying on clustering. However, our paper selected NLP model over lexical analysis or bag-of-words model to extract sentiments behind words due to main two reasons: one [9], NLP could understand the role of the words paly in sentences; two [10], NLP pays more attention on word order and contextual relationships. For NLP model, Sun et al. [11] reviewed recent paradigm shifts in NLP, where reformulating tasks has improved model performance and shown potential for unifying various NLP tasks within single-model frameworks. Li et al. [12] categorized data augmentation (DA) methods into paraphrasing, noising, and sampling, analyzes their application in enhancing training data diversity for NLP tasks. Ruder et al. [13] presented an overview of transfer learning methods in NLP, detailing how models pre-trained on diverse data improve state-of-the-art performance across tasks and the nature of their learned representations. Müller et al. [14] introduced COVID-Twitter-BERT (CT-BERT), a model trained on COVID-19 Twitter data, evaluated for NLP tasks like classification and question-answering, and benchmarked against BERT-LARGE.

To enhance the precision of stock price forecasts and empower investors and investment institutions to mitigate risks while augmenting returns, our work study the cross-relationship between stock markets and investor sentiment measures. With the sentiments data we acquired from NLP algoritn and the stock data from the Shanghai Securities Composite Index (SSEC), we satisfy all conditions to employ a Multifratcal Detrended Cross-Correlation Analysis (MF-DCCA) [15] upon the two sets of data (stock markets and ISM). Numerous past works have

shown the feasbility and efficiency of MF-DCCA method. Hurst [16] was the first author who propposed method that characterized the properties of nonlinear time series. More and more works have been inspired to apply fractal correlation properties since then, among which Multifractal Detrended Fluctuation Analysis (MF-DFA) gradually shows its outstanding performance in financial market [17, 18, 19], dealing with complex nonlinear datasets. A number of varieties of MF-DFA also thrived through deligent works by scholars. Carbone et al. [20] presented Multifractal Detrended Moving Average (MF-DMA) algorithm investigating German financial series. Wang et al. [21] creatively brought out multifractal detrending weighted average algorithm of historical volatility (MF-DHV) for one-dimensional multifractal measure. Podobnik et al. [22] proposed Multifractal DCCA (MF-DCCA) method, exploring the cross-relationship between two nonlinear and nonstationary time series. In the domain of finanical markets, MF-DCCA is frequently applied in conducting cross-relationship analysis by scholars [23, 24, 25, 26, 27]. In this paper, we study the cross-relationship of ISM and stock markets under the scope of MF-DCCA model.

Past works have shown there exist some conncetion between ISM and stock markets. Baker et al. [28] developed a "top down" macroeconomic approach to investor sentiment, treating its origin as exogenous and demonstrating its measurable, significant effects on stock prices, especially for stocks that are hard to arbitrage or value, based on behavioral finance principles. Bandopadhyaya et al. [29] created an Equity Market Sentiment Index from public data, showing its efficacy in capturing quick market shifts due to news events and explaining a substantial part of the variability in a stock market index, underscoring sentiment's role in asset pricing dynamics. In our paper, we propose a more efficient and accurate model utilizing NLP and MF-DCCA algorithm aiming at unveiling the cross-relationship between ISM and stock markets, in hope of providing positive suggestions to investors and create healthy stock markets.

The layout of this article is shown as follows. In Section 2.3, the main method used in this article is provided. In Section 3, the experimental results are presented. Conclusions are delivered in Section 4.

2. METHODOLOGY

2.1. **MF-DFA.** Multifractal Detrended Fluctuation Analysis (MF-DFA) [30, 31] is a method used to analyze multiple fractal properties in time series. It is mainly used to reveal the fractal structure in time series at different time scales, especially in complex data with nonlinear and non-Gaussian characteristics. The steps of MF-DFA are as follows.

Step 1: The time series data to be analyzed are preprocessed and then the preprocessed time series are cumulatively summed to obtain the cumulative series $P(i)$.

Step 2: Divide into Non-Overlapping Windows. We split the sequence $P(i)$ into A_s non-overlapping windows, each of equal length, where the number of windows is determined by the total length N of the sequence and the scale s used for dividing the windows. Specifically, $A_s \equiv int(\frac{N}{s})$ calculates the number of windows based on the integer division of N by s. To ensure comprehensive processing of all data, the same division procedure is performed from the end of the sequence, resulting in a total of $2A_s$ windows in total.

Step 3: Fitting sequence. Perform least squares fitting for each window by utilizing $y_v^m(i)$ to represent the m-th order polynomial for the v-th window. Subsequently, compute the residuals between the original sequence $P(i)$ and the fitted polynomial $y_v^m(i)$. This process yields the fluctuation function for each window with the following formula:

$$F^2(s,v) = \frac{1}{s}\sum_{i=1}^{s}\left\{P[N-(v-N_s)s+i] - y_v^m(i)\right\}^2,$$

where v, $v = A_s+1, A_s+2, ..., 2A_s$.

Averaging over the $2A_s$ windows gives the q-th order fluctuation function. The fractal order q can take any real value.

$$F_q(s) = \begin{cases} \left\{\frac{1}{2A_s}\sum_{v=1}^{2A_s}[F^2(s,v)]^{\frac{q}{2}}\right\}^{\frac{1}{q}}, & if\ q \neq 0, \\ exp\left\{\frac{1}{4N_s}\sum_{v=1}^{2A_s}\ln[F^2(s,v)]\right\}, & if\ q = 0. \end{cases}$$

Step 4: Obtaining the q-order Fluctuation Function. For each window, the magnitudes of the residuals are weighted and averaged based on different values of q, resulting in the q-order fluctuation function $F_q(s)$.

Step 5: Calculating the Fractal Dimension. As the window scale s increases, $F_q(s)$ follows a power-law relationship $F_q(s) \propto s^{H(q)}$. Here, $H(q)$ is the generalized Hurst exponent, which reflects the fractal characteristics of the sequence at different scales.

2.2. **DCCA.** Detrended Cross-Correlation Analysis (DCCA) [32] is a method used to analyze the long-term correlation between two non-stationary time series. It is mainly used to study the correlation characteristics of two time series at different scales. The steps of DCCA are:

Step 1: $X(i)$ and $Y(i)$ represent two different original sequences, $\bar{x}$ and $\bar{y}$ denote the average of all values in the $X(i)$ and $Y(i)$ sequences, respectively. The sequences profile can be expressed as:

$$X(i) = \sum_{k=1}^{i}(x_k - \bar{x}), \quad i = 1, 2, \ldots, T,$$

$$Y(i) = \sum_{k=1}^{i}(y_j - \bar{y}), \quad i = 1, 2, \ldots, T.$$

Step 2: The X and Y lengths are divided into s non-overlapping parts, a total of $N_s = N/s$, and N may not be a multiple of the time scale s, so the residual case will arise. In order to account for this residual sequence, the same process is repeated starting at the other end of each sequence, so that $2N_s$ segments are obtained.

Step 3: By the least squares fit of each sequence, the local trend for each of the $2N_s$ segments is calculated, subsequently represented by $\tilde{X}$ and $\tilde{Y}$. Then calculate the difference between the original time series and the fitting polynomial.

Here, v represents the v-th window, with values ranging from 1 to $2N_s$.A When $v = 1, 2, \ldots, N_s$,

$$f^2_{DCCA}(s,v) = \frac{1}{s}\sum_{i=1}^{s}(X_{[(v-1)s+i]} - \tilde{X}_v(i))(Y_{[(v-1)s+i]} - \tilde{Y}_v(i)).$$

When $v = N_s, N_s+1, \ldots, 2N_s$,

$$f^2_{DCCA}(s,v) = \frac{1}{s}\sum_{i=1}^{s}(X_{[N-(v-N_s)s+i]} - \tilde{X}_v(i))(Y_{[N-(v-N_s)s+i]} - \tilde{Y}_v(i)).$$

Step 4: Averaging all segments gives the wave function:

$$F_{DCCA}(s) = \{\frac{1}{2N_s}\sum_{v=1}^{2N_s} f^2_{DCCA}(s,v)\}^{\frac{1}{2}}.$$

Step 5: The scaling behavior of the volatility function is determined by analyzing the logarithmic plot of $F_{DCCA}(s)$ versus s: $F_{DCCA}(s) \propto s^{\lambda}$. Different values of the scaling exponent λ indicate different degrees of correlation between the two time series $X(i)$ and $Y(i)$.

2.3. **MF-DCCA.** MF-DFA methods are limited to characterizing the multifractal nature of a single time series and lack the ability to quantify the correlation between two non-stationary time series. Considering these aspects, Zhou proposed the MF-DCCA method [33] based on the MFDFA method and DCCA, which combines the advantages of the MF-DFA and the DCCA methods. Firstly, the MF-DFA method is applied to analyze the multiple fractal characteristics of complex time series to reveal their fractal features in different time scales. Then, the DCCA method is applied to analyze the long-term dependencies in the sequences and establish the complex interactions among different sequences. Finally, the results of MF-DFA and DCCA methods are combined to optimize the accuracy and comprehension of analyzing complex time series, so as to overcome the limitations of traditional methods in interpreting multifractal characteristics and long-term dependencies. The specific calculation process is shown as follows.

Step 1. Consider two time series $\{\phi_i, i = 1, 2, \ldots, T\}$ and $\{\varphi_i, i = 1, 2, \ldots, T\}$. Then construct the cumulative summation sequence with the mean removed.

$$\mathcal{A}_i = \sum_{j=1}^{i}(\phi_j - \bar{\phi}), \quad i = 1, 2, \ldots, T,$$

$$\mathcal{B}_i = \sum_{j=1}^{i}(\varphi_j - \bar{\varphi}), \quad i = 1, 2, \ldots, T.$$

where ϕ_j denotes the jth number in a sequence, where $j = 1, 2, ..., i$. And $\bar{\phi}$ equals $\frac{1}{T}\sum_{i=1}^{T}\phi_i$ and $\bar{\varphi}$ equals $\frac{1}{T}\sum_{i=1}^{T}\varphi_i$. In other words, $\bar{\phi}$ and $\bar{\varphi}$ are the mean of the time series $\{\phi_i\}$ and $\{\varphi_i\}$, respectively.

Step 2. Divide the time series $\{\phi_i\}$ into T_s non-overlapping segments with the fixed size s, where $T_s = [T/s]$ is the number of the segment of $\{\phi_i\}$. In order to minimize information loss during the segmentation process, the same operation is repeated once more from the end of the sequence. Therefore, we can obtain $2T_s$ segments. Meanwhile, the same procedure is applied to sequence $\{\varphi_i\}$.

Step 3. In this step, the polynomial fitting can be obtained through the least square method, which is adopted to fit s points in each segment μ ($\mu = 1, 2, \ldots, T_s, T_{s+1}, \ldots, 2T_s$).

$$\mathcal{M}_\mu(m) = a_1 m^k + a_2 m^{k-1} + \cdots + a_k m + a_{k+1}, \quad m = 1, 2, \ldots, s,$$
$$\mathcal{N}_\mu(m) = b_1 m^k + b_2 m^{k-1} + \cdots + b_k m + b_{k+1}, \quad m = 1, 2, \ldots, s.$$

where $(a_1, a_2, ..., a_k, a_{k+1})$ and $(b_1, b_2, ..., b_k, b_{k+1})$ are the coefficients of polynomial, they are all constants.

Step 4. Calculate the detrended wave function for each subdomain. When $\mu = 1, 2, \ldots, T_s$,

$$\mathcal{F}^2(s, \mu) = \frac{1}{s} \sum_{m=1}^{s} |\mathcal{A}_{[(\mu-1)s+m]} - \mathcal{M}_\mu(m)||\mathcal{B}_{[(\mu-1)s+m]} - \mathcal{N}_\mu(m)|.$$

When $\mu = T_{s+1}, T_{s+2}, \ldots, 2T_s$,

$$\mathcal{F}^2(s, \mu) = \frac{1}{s} \sum_{m=1}^{s} |\mathcal{A}_{[T-(\mu-T_s)s+m]} - \mathcal{M}_\mu(m)||\mathcal{B}_{[T-(\mu-T_s)s+m]} - \mathcal{N}_\mu(m)|.$$

Step 5. The q order wave function $\mathcal{F}_q(s)$ is calculated as follows.

$$\mathcal{F}_q(s) = \begin{cases} \left\{ \frac{1}{2T_s} \sum_{\mu=1}^{2T_s} [\mathcal{F}^2(s, \mu)]^{q/2} \right\}^{1/q}, & \text{if } q \neq 0, \\ \exp\left\{ \frac{1}{T_s} \sum_{\mu=1}^{T_s} \ln[\mathcal{F}^2(s, \mu)] \right\}, & \text{if } q = 0 \end{cases}$$

$\mathcal{F}_q(s)$ represents a function dependent on segment s and fractal order q. As s increases, the series exhibits a long-range power-law correlation. In addition, $h_{\phi\varphi}(q)$ can be obtained by $\mathcal{F}_q(s) \propto s^{h_{\phi\varphi}(q)}$, which is called the generalized Hurst exponent. When $q = 2$, $h_{\phi\varphi}(2)$ is defined as the standard Hurst exponent. When $h_{\phi\varphi}(2)$ equals 0.5, the sequences $\{\phi_i\}$ and $\{\varphi_i\}$ are not mutually correlated and they have no mutual influence on each other. When $h_{\phi\varphi}(2)$ belongs to [0.5,1], the two series have long range correlation. When $h_{\phi\varphi}(2) < 0.5$, the correlations are antipersistent. In addition, the degree of multifractality of the sequences can be described by $\Delta h_{\phi\varphi}(q)$, which can be obtained by $h_{\phi\varphi_{max}}(q) - h_{\phi\varphi_{min}}(q)$. The stronger the multifractal characteristics of the sequences are, the greater $\Delta h_{\phi\varphi}(q)$ is [34].

As well, the scaling exponent $\tau_{\phi\varphi}(q)$ can be used to characterize multifractal features. The specific calculation of $\tau_{\phi\varphi}(q)$ is shown as below.

$$\tau_{\phi\varphi}(q) = q h_{\phi\varphi}(q) - 1.$$

Meanwhile, the deepest essence of multifractal features of the series is the diversity of singularity exponent $\alpha_{\phi\varphi}(q)$. The singularity exponent $\alpha_{\phi\varphi}(q)$ and multifractal spectrum $f_{\phi\varphi}(\alpha)$,

which is used to exhibit the fractal dimension of $\alpha_{\phi\varphi}$ can be obtained by the following formula.

$$\alpha_{\phi\varphi} = h_{\phi\varphi}(q) + qh'_{\phi\varphi}(q),$$
$$f_{\phi\varphi} = q[\alpha_{\phi\varphi} - h_{\phi\varphi}(q)] + 1.$$

Here, $\Delta\alpha_{\phi\varphi} = \alpha_{\phi\varphi_{max}} - \alpha_{\phi\varphi_{min}}$, which reflects the degree of non-uniformity of probability measure distribution and the complexity of the process on the whole fractal structure.

3. NUMERICAL EXPERIMENTS

The layout of this section is as follows. In Section 3.1, we collect the series of daily investor sentiment index with the data series of the Shanghai Securities Composite Index (SSEC) for the same period as well as compare them and test the correlation between them by calculating the Pearson correlation coefficient. And in Section 3.2 we analyze the multifractal cross-correlation between sentiment index and stock return series using the MF-DCCA method. The experimental environment is based on Python 3.8 and MATLAB R2020a with a computer with an Intel(R) Core(TM) i5-4430 CPU 3.00 GHz processor, using the Windows 10 operating system.

3.1. **Correlation analysis.** We first utilize a Python web crawler to obtain comments from investors in the SSE Composite Index Tieba. Then, we use the natural language processing tools [35] in Python to convert these comments into the form of word vectors. We choose a 20 dimensional word vector as the feature representation, which extracts 20 features from each text. Based on the word vectors of these features, we perform the prediction of optimistic and pessimistic moods. Figure 1 shows examples of word vectors for optimistic comments and pessimistic comments. The blue solid line in Fig. 1 represents the word vector for optimistic sentiment and the red solid line represents the word vector for pessimistic sentiment. They can be used to represent the location and characteristics of particular words in a comment in a particular vector space.

Baker and Wurgler [36] defined investor sentiment as their optimistic and pessimistic mindset towards the stock market as a whole. In addition, we define the sentiment index as follows in order to quantify the value of investor sentiment in the market on a given day:

$$SI' = \frac{N_{Positive}}{N_{Positive} + N_{Negative}},$$

where $N_{Positive}$ and $N_{Negative}$ represent the number of positive and negative sentiment predicted by investor comments on a given day, respectively, and $N_{Positive} + N_{Negative}$ is the total number of investor comments selected from the SSE Composite Index Tieba. In order to make the indicator sequence look uniformly distributed, we regularize the sequence, and the specific adjustment formula is Eq. (3.1):

$$SI = \frac{SI' - Min(SI')}{Max(SI') - Min(SI')}, \tag{3.1}$$

where $Max(SI')$ and $Min(SI')$ represent the maximum and minimum values in SI', respectively.

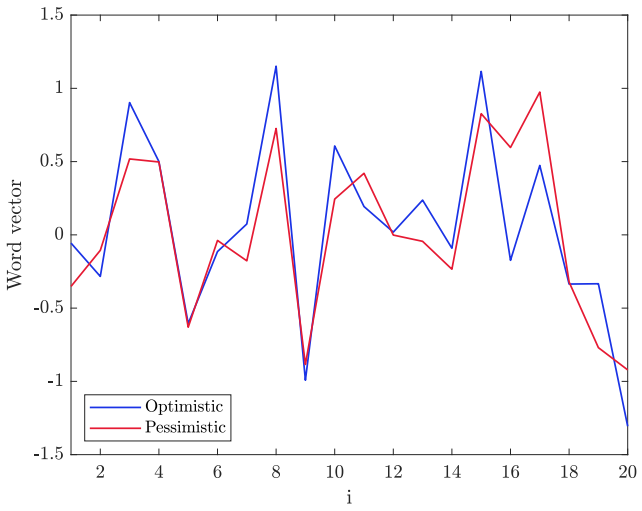

FIGURE 1. Word vector of optimistic and pessimistic comments from investors in the SSE Composite Index Tieba. A color version of the figure is available in the web version of the article.

Next, we select a time interval from 19, Oct 2017 to 16, Oct 2019 and calculate the daily investor sentiment index. The resulting time series is shown in Fig. 2 (a). We also select the data series of SSEC for the same time period, as shown in Fig 2 (b). The data are obtained from website "https://www.investing.com/". Before investigating the cross-correlations of MF-DCCA between sentiment index and stock price index, we first use Pearson correlation coefficient to test the correlations between the two series.

The Pearson correlation coefficient changes from -1 to 1, and when $r > 0$, it indicates that the two series are positively correlated; $r < 0$ indicates that two series are negatively correlated, and the larger the absolute value of r, the stronger the correlation between the two series. The calculated r and P-value of the Pearson correlation coefficient between SSE Composite Index and SI in Table 1 shows there exists a certain correlation between the two sequences.

In addition, we utilize the summation processing step in MF-DCCA for the two sequences, which involves smoothing the data to eliminate some noise and obtain the two smooth sequences shown in Fig. 3. It can be seen that the two time series exhibit strong correlations. We further calculate the Pearson correlation coefficients of the two sequences, as shown in Table 1. The results further demonstrate the cross-correlation between stock prices and investors' sentiment.

3.2. **Experiment and analysis based on MF-DCCA method.** Subsequently, we employ the MF-DCCA method to study the multifractal characteristics of the cross-correlation between

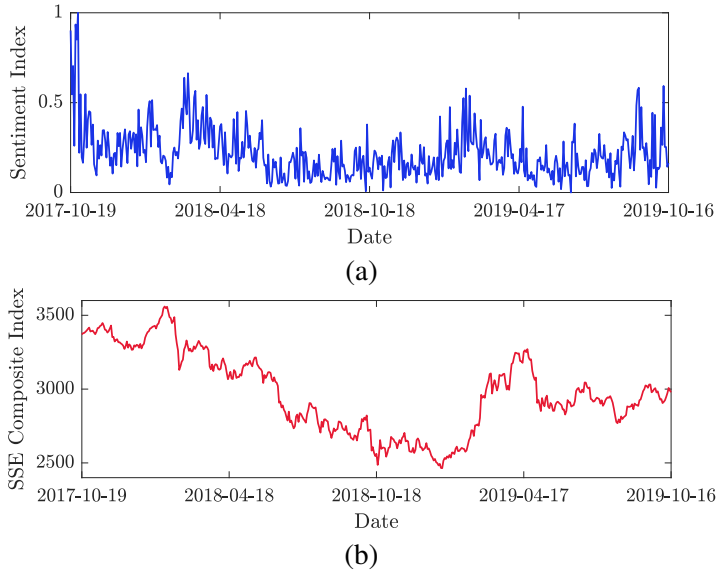

FIGURE 2. Data of (a) Sentiment index based on comments from investors in the SSE Composite Index Tieba, and (b) SSE Composite Index.

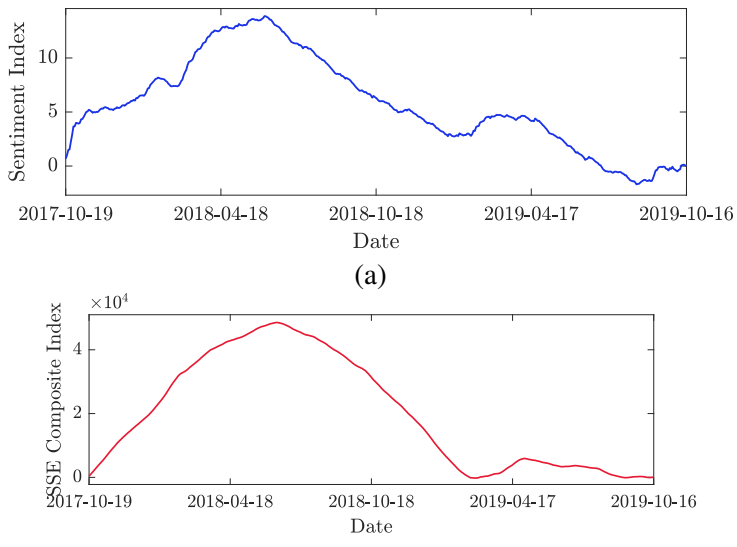

FIGURE 3. Smoothed data of (a) Sentiment index based on comments from investors in the SSE Composite Index Tieba, and (b) SSE Composite Index.

TABLE 1. Pearson correlation test.

Metrics	Fig. 2(a) & Fig. 2(b)	Fig. 3(a) & Fig. 3(b)
ρ	0.4254	0.9214
P-value	$9.7247e-23$	$2.1762e-200$

the return series of the SSE Composite Index and the SI time series. According to [15], the segment size s should be selected in an appropriate interval, due to a large s can cause the $F_q(s)$ becomes statistically unreliable. Here, we set $10 \leq s \leq 60$ with an increment of 5. In addition, $-30 \leq q \leq 30$ with an increment of 3, the fitting order is determined by 2. We depict the double log curves between $F_q(s)$ and s in Fig. 4.

We note that for the series pair $F_q(s)$ increases linearly along with the s, which indicates that there exists power-law behavior and long-range cross-correlations between the return series of SSE Composite Index and the SI time series. Besides, we also observe that with the increasing of q, the slope of the log-log curve decreases, implying that the cross-correlation between the return series of SSE Composite Index and the SI time series is multifractal.

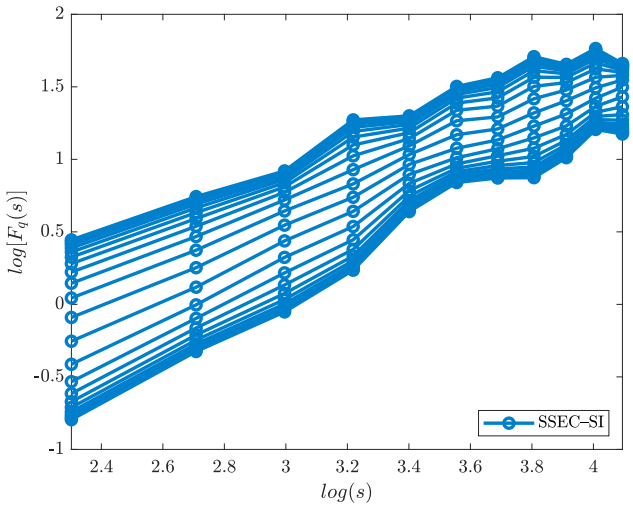

FIGURE 4. Log-log plots of fluctuation function $F_q(s)$ versus segment scale s for SSE Composite Index and the SI time series. From the bottom to the top, q varies from -30 to 30.

In order to more intuitively reflect the degree of multifractal variation between two sequences, we further calculate the generalized Hurst exponent between the two sequences, as shown in Fig. 5. We see that the value of $H_{xy}(q)$ decreases with the increase of q, further confirming the existence of multifractals. Besides, when $q = 2$, we calculate $H_{xy}(2) = 0.8267$, which is larger than 0.5, showing that the cross-correlation between SSEC return and SI time series exhibits persistence.

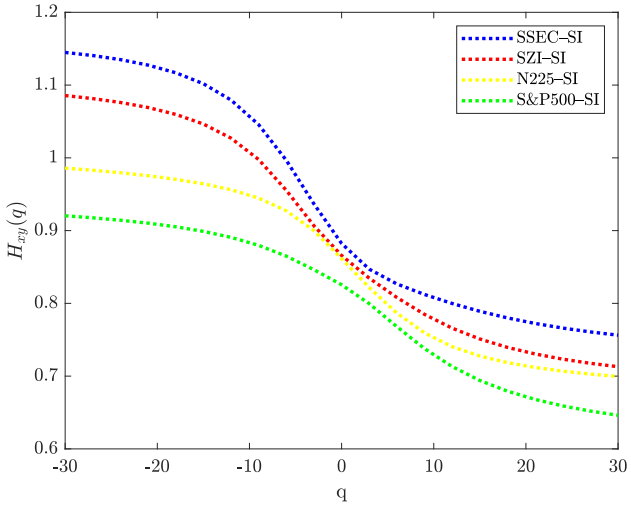

FIGURE 5. The cross-correlation of generalized Hurst exponent $H_{xy}(q)$ between the investor sentiment index derived from the Shanghai Stock Exchange Index Tieba and the different market index. A color version of thee figure is available in the web version of the article.

The multifractal strength of cross-correlation ΔH_{xy}, can be characterized by calculating the difference between the maximum and minimum values of H_{xy}. The larger the value of ΔH_{xy}, the stronger the multifractal intensity of the cross-correlation. In addition, the Renyi exponent can also describe multifractal features, and we also plot the curve of the Renyi exponent between SSEC return and SI time series in Fig. 6. A curve with greater curvature indicates a higher degree of multifractality, which means that fluctuations in the two series are more strongly correlated. We also utilize the multifractal spectrum to further check the cross-correlation of multifractality. As depicted in Fig. 7, the blue curve represents the relation between SSEC return and SI time series, and not exhibited as a point, representing the multifractality exists. Similarly to the generalized Hurst exponent, the width of multifractal spectrum can efficiently describe the degree of multifractality.

In order to demonstrate the effectiveness of multifractal correlation between SSEC and SI time series, we conducted MF-DCCA analysis on the investors' sentiment index series and other market indices simultaneously, also for a comparison. Here, Shenzhen Securities Component Index (SZI), Nikkei 225 Index (N225), and Standard & Poor's 500 Index (S&P500) are considered, and SZI is also a stock index belonging to the Chinese financial market. Before conducting MF-DCCA analysis, we first organize the data of N225 and S&P500 stock price sequences to correspond to the time of sentiment index. We calculate the generalized Hurst exponent, Renyi exponent, and multifractal spectrum of series pairs such as SZI/SI, N225/SI, S&P 500/SI, respectively. The curves are shown in Figs. 5,6,7.

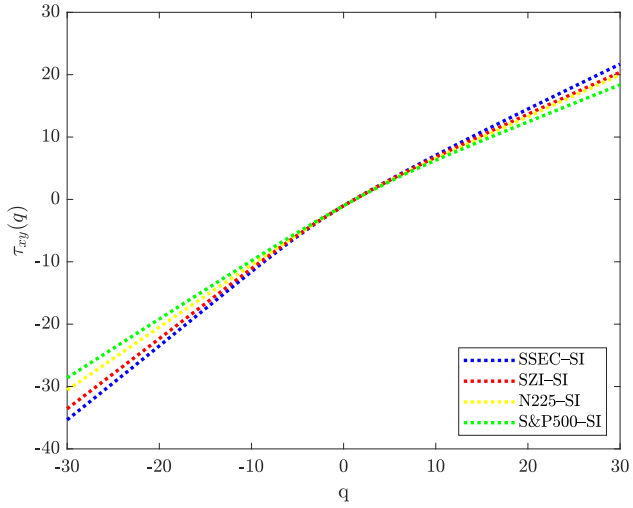

FIGURE 6. The cross-correlation of Renyi exponent $\tau_{xy}(q)$ between the investor sentiment index derived from the Shanghai Stock Exchange Index Tieba and different market index. A color version of the figure is available in the web version of the article.

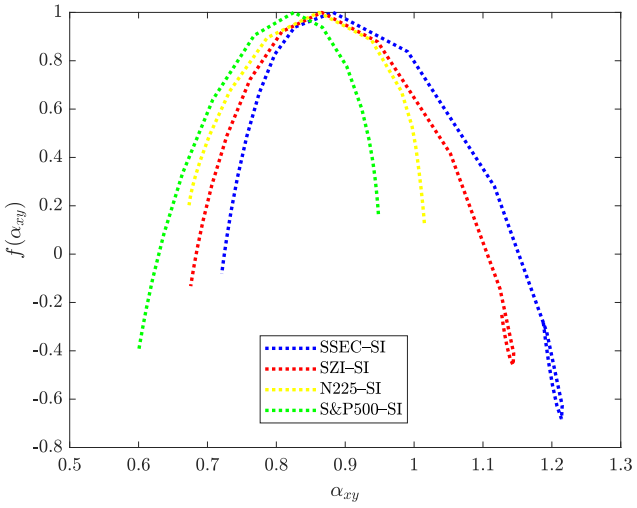

FIGURE 7. The cross-correlation of multifractal spectrum between the investor sentiment index derived from the Shanghai Stock Exchange Index Tieba and the different market index. A color version of the figure is available in the web version of the article.

Specifically, we calculate the values of $H_{xy}(2)$, ΔH_{xy}, and $\Delta\alpha_{xy}$ in Table 2. The informations reflected by these multifractal feature values are as expected. The higher multifractal cross-correlation indicates that there is a closer relationship between the multifractal properties of the two sequences. Specifically, it shows a stronger correlation between fluctuations and variability in one sequence and fluctuations and variability in another. The results reflect that the investor sentiment index and all markets index return series exhibit a positive persistence of multifractal cross-correlation, and $H_{SSEC-SI}(2) > H_{SZI-SI}(2) > H_{N225-SI}(2) > H_{S\&P500-SI}(2)$. Due to the fact that the investors' sentiment index is derived from comments in the Shanghai Stock Exchange Index Tieba, it also shows the highest performance on SSEC in reflecting the persistence of cross-correlations. In addition, SZI belongs to the financial securities market of China, thus, $H_{xy}(2)$ between SZI and SI is also relatively large, only slightly smaller than $H_{SSEC-SI}(2)$. Besides, the persistence of N225 and S&P500 markets and the investors' sentiment index series from the Shanghai Composite Index Tieba is much smaller than that of the two securities indices in the Chinese market.

At the same time, we calculate the indicator ΔH_{xy} that reflects the degree of multifractal cross-correlation between two sequences. The results show that the cross-correlation between investor sentiment index and SSEC return sequence is the strongest, and the cross-correlation between SZI and SI is slightly weaker than that of $\Delta H_{SSEC-SI}$. For two financial markets outside of the Chinese market, their values are similar and far less than the strength of the cross-correlation between the Chinese financial market and SI. The relevant information can also be obtained from Fig. 5. To further verify this conclusion, we then obtain the Renyi exponent $\tau_{xy}(q)$, an metric that characterizes the strength of multifractal cross-correlation based on the curvature size of the curve. From Fig. 6, it can be seen that the curvature of the Renyi exponent curve of SSEC and SI is the highest, followed by SZI-SI, N225-SI, S&P500-SI. $\Delta\alpha_{xy}$ of SI and the different market index are also computed, which can also reflect the strength of multifractal cross-correlation. The related results are shown in Fig. 7 and Table 2, which further validate the conclusion we have drawn above.

TABLE 2. Pearson correlation test.

Metrics	SSEC-SI	SZI-SI	N225-SI	S&P500-SI
$H_{xy}(2)$	0.8267	0.8080	0.7866	0.7679
ΔH_{xy}	0.3886	0.3726	0.2863	0.2742
$\Delta\alpha_{xy}$	0.4943	0.4689	0.3417	0.3481

To reveal the trend of cross-correlation between investors' sentiment index and stock price return series, we next use MF-DCCA for a rolling time interval analysis. This measure captures the dynamic evolution features of the cross-correlation relationship between the SSEC and SI. We set 201 trading days as the time interval. Besides, for the rolling time interval analysis, one trading day is chosen as the step size.

Based on the initial sample data, we calculate the $H_{xy}(2)$ and $\Delta\alpha_{xy}$ time series from August 10, 2018 to October 16, 2019, as shown in Fig. 8. From these data, it can be seen that the correlation between financial market returns and investors' sentiment index, as well as the

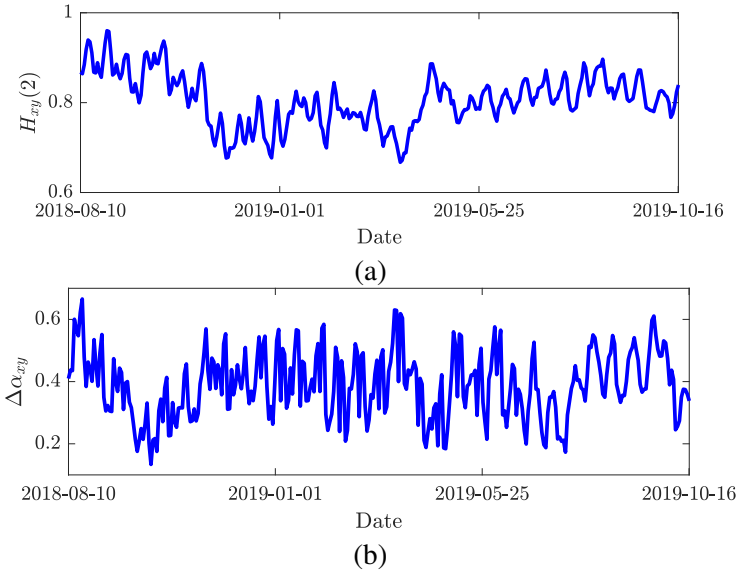

FIGURE 8. The rolling time interval analysis of the cross-correlation between SSEC and SI of (a) $H_{xy}(2)$, and (b) $\Delta\alpha_{xy}$.

persistence of the correlation, fluctuates over time, indicating that the correlation between financial market returns and sentiment may be influenced by the macroeconomic environment. In addition, the values of $H_{xy}(2)$ are all greater than 0.5, indicating that the cross-correlation has maintained a positive persistence during these periods.

In addition, we also compute the time series of $H_{xy}(2)$ and $\Delta\alpha_{xy}$ sequence values between SZI and SI, as shown in Fig. 9. We note a strong consistency of the correlation between SZI financial market returns and investors' sentiment index with that of the SSEC-SI behaves, which applies to both $H_{xy}(2)$ and $\Delta\alpha_{xy}$.

4. CONCLUSIONS

In this study, we adopted the natural language processing method to convert the personal comment text in the Shanghai Stock Exchange Index Tieba into word vectors and predicted the positive or negative attributes of the text. Then, the investor sentiment index was constructed for as a time series. Firstly, the correlation between the sentiment index and the financial market was preliminarily confirmed through the Pearson correlation coefficient. Furthermore, we used MF-DCCA to study the multifractal cross-correlation between sentiment index and stock return series. The empirical results showed that there is a cross-correlation between SSEC return series and SI, and it has positive persistence. In addition, in order to display the degree of cross-correlation between the two time series more clearly, we used three stock market return sequences such as SZI, N225, and S&P500 for comparison. The calculation results suggested that except for SZI, which belongs to China, the multifractal cross-correlation

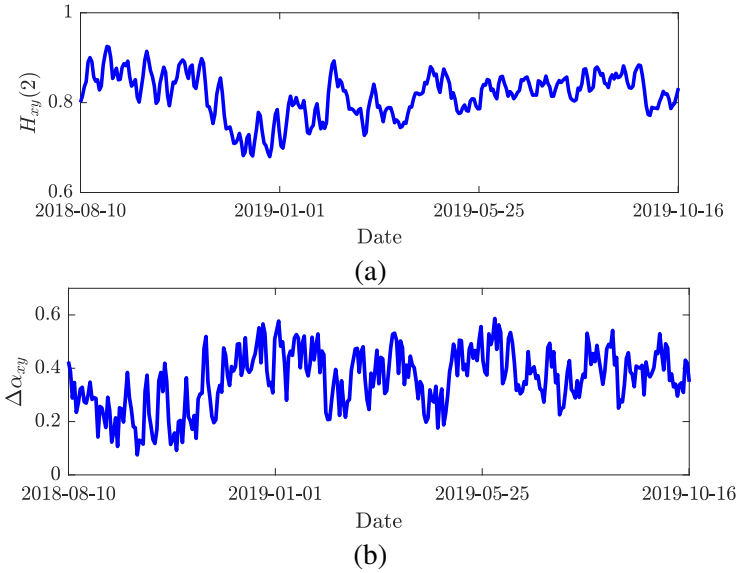

FIGURE 9. The rolling time interval analysis of the cross-correlation between SZI and SI of (a) $H_{xy}(2)$, and (b) $\Delta\alpha_{xy}$.

and persistence of the other two financial markets are much smaller than those of SSEC-SI. SZI-SI is only slightly weaker than SSEC-SI. In addition, we also calculated the $H_{xy}(2)$ and $\Delta\alpha_{xy}$ series over time using the rolling time interval analysis, and the results showed that the macroeconomic environment may affect the persistence and strength of the multifractal cross-correlation between the SSEC return series and the SI series. Finally, we also conducted the rolling time interval analysis of SZI-SI and found that the trends of SSEC-SI and SZI-SI in the indicators of rolling time interval analysis were basically consistent. Our research can help investors better adapt to the changing market environment and create long-term value for investors.

ACKNOWLEDGMENT

The corresponding author Jian Wang expresses thanks for the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant Nos. 22KJB110020).

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests regarding the publication of this article.

REFERENCES

[1] Cheng, A. C. *Economic factors and stock markets: Empirical evidence from the UK and the US*, International Journal of Finance& Economics, **1(4)** (1996), 287–302.

[2] Momani, G. F., & Alsharari, M. A. *Impact of economic factors on the stock prices at Amman stock market (1992-2010)*, International Journal of Economics and Finance, **4(1)** (2012), 151–159.
[3] Qiu, L., & Welch, I. *Investor sentiment measures* (2004).
[4] Shiller, R. J. *Measuring bubble expectations and investor confidence*, The Journal of Psychology and Financial Markets, **1(1)** (2000), 49–60.
[5] Beer, F., & Zouaoui, M. *Measuring stock market investor sentiment*, Journal of Applied Business Research (JABR), **29(1)** (2013), 51–68.
[6] Bochkay, K., & Dimitrov, V. *Qualitative management disclosures and market sentiment*, Available at SSRN 2538812 (2014), 1–45.
[7] Hippisley, A. *Lexical Analysis*, Handbook of natural language processing 2, 2010.
[8] Zhang, Y., Jin, R., & Zhou, Z. H. *Understanding bag-of-words model: a statistical framework*, International journal of machine learning and cybernetics, **1** (2010), 43–52.
[9] Guthrie, L., Pustejovsky, J., Wilks, Y., & Slator, B. M. *The role of lexicons in natural language processing*, Communications of the ACM, **39(1)** (1996), 63–72.
[10] David, P., Hawes, T., Hansen, N., & Nolan, J. J. *Considering context: reliable entity networks through contextual relationship extraction*, Proceedings of SPIE - The International Society for Optical Engineering, **9851** 2016.
[11] Sun, T. X., Liu, X. Y., Qiu, X. P., & Huang, X. J. *Paradigm shift in natural language processing*, Machine Intelligence Research, **19(3)** (2022), 169–183.
[12] Li, B., Hou, Y., & Che, W. Data augmentation approaches in natural language processing: A survey. Ai Open, 3 (2022), 71–90.
[13] Ruder, S., Peters, M. E., Swayamdipta, S., & Wolf, T. *Transfer learning in natural language processing*, Proceedings of the conference of the North American chapter of the association for computational linguistics: Tutorials, 2019.
[14] Müller, M., Salathé, M., & Kummervold, P. E. *Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter*, Frontiers in Artificial Intelligence, **6** (2023), 1023281.
[15] Kantelhardt J.W., Zschiegner S.A., Koscienlny-Bunde E., Bunde A., Havlin S., Stanley H.E., *Multifractal detrended fluctuation analysis of nonstationary time series*, Physica A, **316** (2002), 87–114.
[16] Hurst, H. E. *Long-term storage capacity of reservoirs*, Transactions of the American society of civil engineers, **116(1)** (1951), 770–799.
[17] Wu, X., Xu, H., Wu, S., Huang, M., & Wang, J. *Nonlinear Dynamic Analysis of the US Defense Stock Markets under the Russia–Ukraine Conflict*, Fluctuation and Noise Letters, **23** (2023). 2450004.
[18] Rizvi, S. A. R., Dewandaru, G., Bacha, O. I., & Masih, M. *An analysis of stock market efficiency: Developed vs Islamic stock markets using MF-DFA*, Physica A: Statistical Mechanics and its Applications, **407** (2014), 86–99.
[19] Wang, J., & Shao, W. *Multifractal analysis with detrending weighted average algorithm of historical volatility*, Fractals, **29(05)** (2021), 2150193.
[20] Carbone, A., Castelli, G., & Stanley, H. E. *Time-dependent Hurst exponent in financial time series*, Physica A: Statistical Mechanics and its Applications, **344(1-2)** (2004), 267–271.
[21] Wang, J., & Shao, W. *Multifractal analysis with detrending weighted average algorithm of historical volatility*, Fractals, **29(05)** (2021), 2150193.
[22] Podobnik, B., Jiang, Z. Q., Zhou, W. X., & Stanley, H. E. *Statistical tests for power-law cross-correlated processes*, Physical Review E, **84(6)** (2011), 066118.
[23] Feng, S., Liu, H., & Yang, Y. *Research on the Risk of the Online Lending Market in China: A New Perspective Based on MF-DCCA*, Emerging Markets Finance and Trade, **58(7)** (2022), 1860–1870.
[24] Ruan, Q., Bao, J., Zhang, M., & Fan, L. *The effects of exchange rate regime reform on RMB markets: A new perspective based on MF-DCCA*, Physica A: Statistical Mechanics and its Applications, **522** (2019), 122–134.
[25] Wang, J., Shao, W., & Kim, J. *Cross-correlations between bacterial foodborne diseases and meteorological factors based on MF-DCCA: A case in South Korea*, Fractals, **28(03)** (2020), 2050046.

[26] Yan, Y., Shao, W., & Wang, J. *Comparison of Price-Volume Correlation for Some Cryptocurrencies Based on MF-ADCCA*, Fluctuation and Noise Letters, **21(03)** (2022), 2250040.

[27] Wang, J., Jiang, W., Yan, Y., Shao, W., Wu, X., & Hua, Z. *Exploring the asymmetric multifractal characteristics of price–volume cross-correlation in the Chinese rebar futures market based on MF-ADCCA*, Fluctuation and Noise Letters, **22** (2023). 2350029.

[28] Baker, M., & Wurgler, J. *Investor sentiment in the stock market*, Journal of economic perspectives, **21(2)** (2007), 129–151.

[29] Bandopadhyaya, A., & Jones, A. L. *Measuring investor sentiment in equity markets*, Journal of Asset Management, **7** (2006), 208–215.

[30] Zhang, Q., Xu, C. Y., Yu, Z., Liu, C. L., & Chen, Y. D. *Multifractal analysis of streamflow records of the East River basin (Pearl River), China*, Physica A: Statistical Mechanics and its Applications, **388(6)** (2009), 927-934.

[31] Wang, J., Shao, W., & Kim, J. *Ecg classification comparison between mf-dfa and mf-dxa*, Fractals, **29(02)** (2021), 2150029.

[32] Podobnik, B., & Stanley, H. E. *Detrended Cross-Correlation Analysis: A New Method < format> for Analyzing Two Nonstationary Time Series*, Physical review letters, **100(8)** (2008), 084102.

[33] Zhou, W. X. *Multifractal detrended cross-correlation analysis for two nonstationary signals*, Physical Review E—Statistical, Nonlinear, and Soft Matter Physics, **77(6)** (2008), 066211.

[34] Zunino, L., Tabak, B. M., Figliola, A., Pérez, D. G., Garavaglia, M., & Rosso, O. A. *A multifractal approach for stock market inefficiency*, Physica A: Statistical Mechanics and its Applications, **387(26)** (2008), 6558-6566.

[35] Zhang, R., Jia, C., & Wang, J. *Text emotion classification system based on multifractal methods*, Chaos, Solitons & Fractals, **156** (2022), 111867.

[36] Baker, M., & Wurgler, J. *Investor sentiment and the cross-section of stock returns*, The journal of Finance, **61(4)** (2006), 1645-1680.