A Novel Correlation Model between Investor Sentiment and Trading Behavior Based on Attention Mechanism with Time-Varying Information

Shan Li[1,a], Mengxiang Sun[1,b], Xinge Liu[1,c*], Li Zeng[2,d]

[1]School of Mathematics and Statistics, Central South University, Changsha, 410083, China
[2]Information Technology Department I, Shenzhen Stock Exchange, Shenzhen, 518038, China

[a]lishan5600@163.com, [b]2179243196@qq.com
[c*]liuxgjiayou@126.com, [d]lzeng@szse.cn

Abstract. The security market has the dual characteristics of emerging and transforming, and its stability has a far-reaching impact on the healthy development of social economy. At present, there are few researches on the correlation analysis between investor sentiment and trading behavior in securities market, and there is a lack of systematic and in-depth theoretical analysis. In this paper, statistical model (ARMA-GARCH-Copula) and deep learning model (Encoder+CNN) are combined to accurately describe and study the correlation between investor sentiment and transaction behavior from multiple perspectives. A novel correlation model between investor sentiment and trading behavior based on attention mechanism with time-varying information (CMISTB) is proposed. The CMISTB model can capture the nonlinear and volatility characteristics of time series data, flexibly analyze the dependence between investor sentiment and trading behavior time series data, and reduce the prediction bias caused by the limitation of a single method. The CMISTB model can deal with multi-variable time series data effectively, which provides a powerful tool for complex financial risk management. This work contributes to the in-depth study of irrational fluctuations in the market, provides theoretical and empirical support for maintaining the steady development of the financial market, and provides an important reference and guidance for further exploring the market behavior.

Keywords: investor sentiment, trading behavior, correlation analysis, Copula, Encoder.

1 Introduction

The study of investor sentiment (IS) and trading behavior (TB) has always been one of the key issues in the financial field [1]. Changes in investor sentiment may lead to increased uncertainty in trading behavior, thus affect the process of trading decisions. However, the existing correlation analysis between IS and TB remains in the stage of qualitative analysis and policy analysis, lacking scientific model construction and application research, and unable to meet the increasingly complex analysis needs.

L. Liu et al. (eds.), *Proceedings of the 3rd International Conference on Financial Innovation, FinTech and Information Technology (FFIT 2024)*, Advances in Computer Science Research 118,
https://doi.org/10.2991/978-94-6463-572-0_7

Therefore, quantitative research on the dynamic relationship between IS and TB will help explain irrational market fluctuations, maintain the healthy and balanced development of financial markets, and provide theoretical and empirical basis for further exploration of market behavior.

Correlation analysis of time series has always been a hot topic in financial statistics. Pearson, Spearman and other correlation coefficients are all traditional calculation methods. These methods are always limited to many constraints and susceptible to extreme values, and cannot measure the dependent structure between variables systematically and comprehensively. Compared with traditional methods, Copula function has a looser restriction on edge distribution. It does not require the same edge distribution, and can join any edge distribution to construct a joint distribution. Using copula function to calculate the correlation between variables has been favored by many researchers. Laih [2] proposed the GARCH-copula model to calculate the correlation coefficient by applying the sequence alignment technology in the four-stage algorithm, providing a new way of thinking for the processing and analysis of financial sequences. Li et al. [3] calculate the correlation of different indices with asymmetric copula. Shih et al. [4] proposed a new method for calculating correlation ratio based on copula function and proved the effectiveness. Copula function can adapt to the complex and changeable financial market, which has a broader application prospect.

Deep learning algorithms have made good progress in analyzing financial time series, which can dynamically and adaptively capture complex relationships between different variables and learn a compact representation of data. Treena et al. [5] analyzed financial image data basing on Siamese type neural network and using the similarity distribution of images to complete the prediction of S&P 500. Yang et al. [6] built a DL-EWP model based on deep learning and Elliott wave principle, and used financial datasets of different markets to predict future trends. In addition, some researchers have combined deep learning models with financial statistical models, using fusion models to analyze financial data. Kim et al. [7] combined LSTM and several GARCH to construct a GEW-LSTM model based the security market index, and the MAE of GEW-LSTM was 0.0107. Ni et al. [8] combined RDNN and DCC-GARCH models to build a DCDNN model to analyze the correlation of stock markets in different securities markets. Combining deep learning models with statistical models to learn the characteristics of financial data from multiple perspectives can provide more information for the analysis and decision-making of the securities market.

This paper proposes a novel correlation model between investor sentiment and trading behavior based on attention mechanism with time-varying information, called CMISTB model. First, the time-varying information of time series is obtained based on the ARMA-GARCH-Copula (AGC) model, and then the Encoder+CNN correlation model is constructed based on the time-varying information to complete the overall construction of the CMISTB model. The model uses ARMA-GARCH (AG) method to modify the time series, which effectively avoids the influence of autocorrelation and heteroscedasticity of time series on the model. In addition, Copula function modeling can overcome the disadvantages of the traditional normal hypothesis and linear correlation, and can analyze the time-varying information between variables more accurately, which can adapt to the complex and changeable financial market. The combination of statistical model and deep learning model can dynamically and adaptively capture the complex relationship between IS and TB, revealing the correlation mechanism between time series.

2 Materials and Methods

2.1 ARMA-GARCH-Copula Model Based on Stock Market Time Series

Financial time series generally have the characteristics of serial autocorrelation, time-varying, clustering, etc. This paper uses AG model to modify the series and obtain an independent and equally distributed residual series without serial autocorrelation and unconditional heteroscedasticity. Then, different methods can be used to analyze the dependent structure of the residual sequence. Because Copula function can capture the relevant information of various risk factors more extensively and effectively. Therefore, Copula technique is introduced to analyze the dependent structures of different time series on the basis of AG model. The AG model [9] is expressed as follows:

$$\begin{cases} r_t = \mu + \sum_{i=1}^{r} \varphi_i r_{t-i} + \varepsilon_t - \sum_{j=1}^{s} \theta_j \varepsilon_{t-j} \\ \varepsilon_t \left| I_{t-1} = \sqrt{h_t} \eta_t, \quad \eta_t \sim iid \right. \\ h_t^2 = \alpha_0 + \sum_{m=1}^{p} \alpha_m \varepsilon_{t-m}^2 + \sum_{n=1}^{q} \beta_n h_{t-n}^2 \end{cases} \tag{1}$$

where $\{r_t\}$ represents financial time series, μ is a constant representing the mean, $\{\varepsilon_t\}$ represents the residual sequence, $\alpha_0 > 0$, $\alpha_m \geq 0$, $\beta_n \geq 0$, h_t represents the conditional variance of ε_t. I_{t-1} represents the information set up to time t, r, s, p, q are non-negative integers.

Copula function [3] is more flexible in describing correlation, and can be used to integrate the standardized residual sequence of edge distribution. Let H be the joint distribution function of the random vector (X_1, X_2), and H_1 and H_2 are the edge distribution functions of X_1, X_2, respectively. Sklar's theorem ensures the existence of Copula function $C:[0,1]^2 \to [0,1]$, satisfying:

$$H(x, y) = C(H_1(x), H_2(y)) \tag{2}$$

2.2 Correlation Model between IS and TB Based on Encoder+CNN Mechanism

This paper constructs a correlation model based on Encoder+CNN mechanism. Specifically, a novel attention mechanism (AM) is designed in the Encoder structure to obtain the IS and TB correlation coefficient, and then adds the CNN as the output layer to complete the construction of the overall model. The attention weight [10] is used to characterize the correlation coefficient between IS and TB, as shown in (3).

$$\underset{1\times1}{a_l} = align\left\{ score(\underset{d_q\times1}{q}, \underset{d_k\times1}{k_l}); \underset{n\times1}{e} \right\} \tag{3}$$

where $a_l \in R^l$ represents the corresponding weight vector of the v_i. $align$ represents alignment processing. $e = [e_1, e_2, \cdots, e_n] \in R^{N\times1}$ is the attention score vector, $\underset{1\times1}{e_l} = score(\underset{d_q\times1}{q}, \underset{d_k\times1}{k_l})$. This paper uses time-varying information obtained from AGC model as the q. $\underset{d_k\times n}{K} = \underset{d_k\times d_f}{W_k} \times \underset{d_f\times n}{F}$, $\underset{d_v\times n}{V} = \underset{d_v\times d_f}{W_v} \times \underset{d_f\times n}{F}$.

3 Empirical Analysis

3.1 Datasets

The quantification of investor sentiment is always one of the difficulties in the securities market. In recent years, the phenomenon of A-share industry indexes rising and falling with each other has weakened. Due to the differences in macro policies and market confidence in various industries, investor sentiment towards various industries is different. It is of great significance to find appropriate indicators to describe industry sentiment for analyzing the correlation between sentiment and transaction behavior in different industries. Considering the anchoring effect of finance, this paper constructs the Net Percent of New High Minus New Low (NH-NL) to characterize the industry sentiment [11], as shown in (4):

$$NH - NL = \frac{H - L}{N} \tag{4}$$

where H represents the number of stocks that hit a new annual high, L represents the number of stocks that hit a new annual low, and N represents the number of stocks that have been listed in the industry for more than 1 year. NH-NL means a closing price greater than/less than the highest/lowest closing price of the range over the past 52 weeks to a week ago.

This paper uses the stocks data in the stock market from 2017.4 to 2023.3, in which the industry classification is divided according to CITIC Level 1 (30 industries). Volume (TrdVol) is a true response to the attractiveness of a stock to investors, and is the total number of transactions between buyers and sellers of a stock. In this paper, the trading volume time series of Shanghai Composite Index is defined as trading behavior time series.

3.2 Results of the ARMA-GARCH-Copula Model

The autocorrelation and heteroscedasticity of financial time series often affect the model performance. The AGC model is used to obtain the time-varying information of the dataset. The AG model is used to correct the autocorrelation and heteroscedasticity of the time series, and the residual sequence of the series is obtained. The normalized residual sequence is then converted to a uniform distribution on [0,1] by probability integral transformation. Finally, the Copula model is selected to extract the time-varying information between different time series. In this paper, Gaussian Copula, Student Copula and Clayton Copula are used to obtain the time-varying information between sequences. Taking media industry sentiment and trading behavior as an example, the time-varying information under different copula functions is shown in Figure 1.

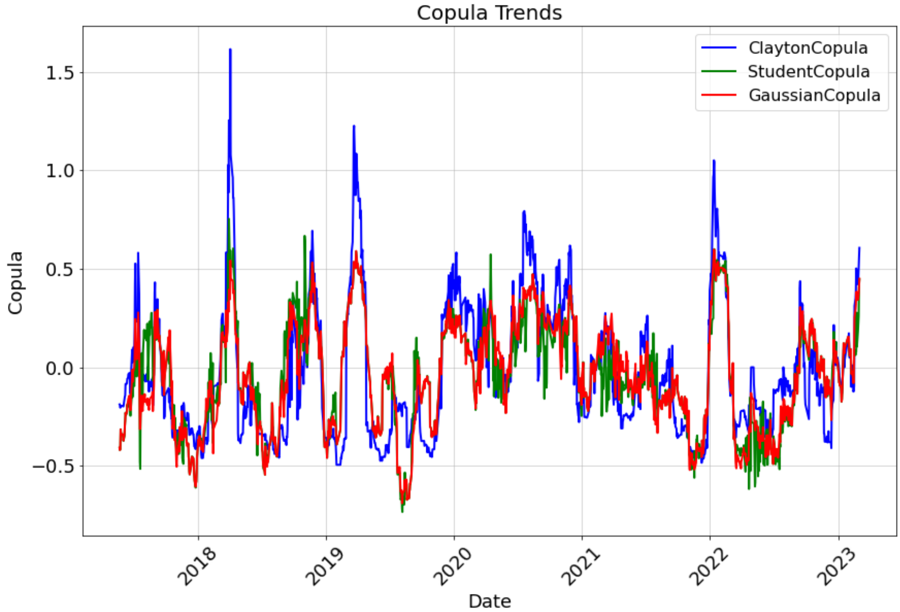

Fig. 1. Time-varying information of time series under different functions

3.3 Results of the Correlation Model Based on Encoder+CNN Mechanism

Based on the statistical model and Encoder+CNN structure, this paper constructs the correlation model between IS and TB, and obtains their correlation coefficient. The evaluation metrics use $RMSE$, MSE, MAE and ACC, where $ACC = \frac{correct\ predictions}{total\ samples}$, correct predictions samples indicate that the MSE of the sample is less than the threshold. The final evaluation results of the model are shown in Table 1.

Table 1. Evaluation metrics of the model.

	Train			Test		
Metrics	Clayton	Student	Gaussian	Clayton	Student	Gaussian
ACC_1	0.98336	0.992076	0.994453	0.950355	0.964539	0.978723
ACC_2	0.912837	0.957211	0.961173	0.87234	0.914894	0.943262
MSE	0.22693	0.161497	0.095842	0.798278	0.220872	0.12926
RMSE	0.476372	0.401867	0.309584	0.893464	0.46997	0.359528
MAE	0.298407	0.240034	0.217508	0.425766	0.2922	0.238379

Note: ACC_1: The threshold is 1, and ACC_2: The threshold is the standard deviation of the logarithm of trading behaviors.

Table 1 shows the best performance of the model when using Gaussian Copula function as the extraction method for time-varying information. For the testing set, when using the Gaussian Copula function, the model's ACC_1 increased by 2.8% and 1.4% compared to Clayton Copula and Student Copula, respectively. Using Gaussian Copula function achieved the highest ACC_1 and ACC_2 values on both the training and testing sets, while MSE, RMSE, and MAE also reached the lowest. Therefore, using Gaussian Copula function as the extraction method for time-varying information can achieve optimal model performance.

By designing AM in the CMISTB model, it is possible to assign appropriate weights to different investor sentiment variables during the learning process, thereby more accurately evaluating their impact on trading behavior. This model can effectively capture the correlation between IS and TB. Furthermore, the CMISTB model combines statistical methods with deep learning to accurately quantify the correlation between IS and TB from multiple perspectives, improving the effectiveness of the data and enhancing the performance of the model. Figure 2 shows the correlation coefficient between IS and TB in some test samples (different trading days).

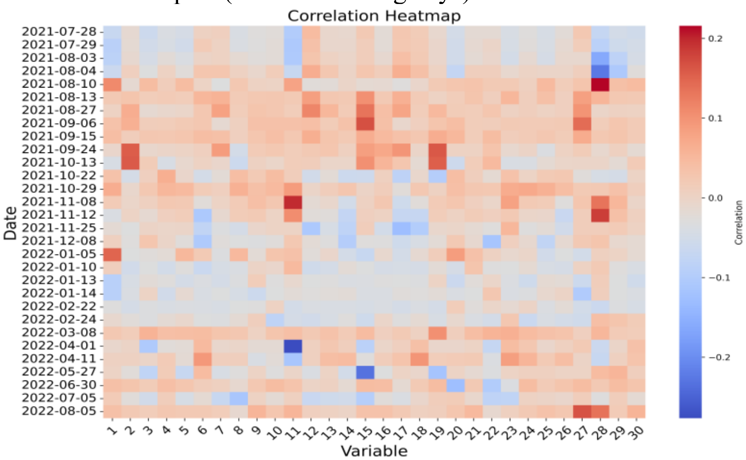

Fig. 2. Correlation coefficient between investor sentiment and trading behavior.

Figure 2 shows the impact of IS on TB in different industries. The results indicate that the correlation coefficient between IS and TB is concentrated between -0.2 and 0.2, and most trading days do not highlight the impact of industry investor sentiment on trading behavior. However, in a few trading days, investor sentiment in certain industries has a significant impact on trading behavior, which is consistent with the current industry phenomenon in the securities market. In summary, the CMISTB model proposed can effectively find the correlation coefficient between IS and TB.

4 Conclusion

The existing models for the correlation between IS and TB remain at the stage of qualitative analysis and policy analysis, lacking scientific model construction, application, and transmission mechanism research, which cannot meet the increasingly refined analysis needs. This paper proposes a novel correlation model between IS and TB based on attention mechanism with time-varying information. This model combines the advantages of time series modeling and deep learning to reduce prediction bias caused by the limitations of a single method. It can flexibly capture the dependency relationship between IS and TB time series data, providing a more comprehensive perspective for correlation measurement. However, how to fully utilize the temporal features in financial data to provide effective information for correlation models and improve model performance remains a question that needs to be continuously explored in the future.

Acknowledgements

This work is partly supported by the National Key Research and Development Program of China under Grant No. 2022YFC3303303, the National Natural Science Foundation of China (NSFC) under Grant No. 72293574 and the Natural Science Foundation of Hunan Province under Grant No. 2022JJ30677.

References

1. Yang C.P., Zhou L.Y. (2015) Investor trading behavior, investor sentiment and asset prices. North American Journal of Economics and Finance, 34: 42–62. https://doi.org /10. 1016/ j.najef.2015.08.003.
2. Laih Y.W. (2014) Measuring rank correlation coefficients between financial time series: A GARCH-copula based sequence alignment algorithm. European Journal of Operational Research, 232:375–382. https://doi.org /10.1016/j.ejor.2013.07.028
3. Li X., Hou B. (2022) Correlation analysis of financial assets based on asymmetric copula. Frontiers in Applied Mathematics and Statistics, 8: 1-16. https: // doi. org/ 10. 3389/ fams. 2022.1005956.
4. Shih J.H., Emura T. (2021) On the copula correlation ratio and its generalization. Journal of Multivariate Analysis, 182:1-14. https://doi.org/10.1016/j.jmva.2020.104708.

5. Treena B., Menzer O., Ward J., SenGupta I. (2022) A Novel Implementation of Siamese Type Neural Networks in Predicting Rare Fluctuations in Financial Time Series. Risks, 10:1-16. https://doi.org/10.3390/risks10020039.
6. Yang D.X., Mu S.D., Liu Y.J., Gu J.J., G.; Lien C.L. (2023) An Improved Deep-Learning-Based Financial Market Forecasting Model in the Digital Economy. Mathematics, 11:1466-1484. https://doi.org/10.3390/math11061466.
7. Kim H.Y., Won C.H. (2018) Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. Expert Systems with Applications, 103: 25–37. https://doi.org/10.1016/j.eswa.2018.03.002.
8. Ni J., Xu Y. (2023) Forecasting the Dynamic Correlation of Stock Indices Based on Deep Learning Method. Computational Economic, 61:35–55. https://doi.org/10.1007/s10614-021-10198-3.
9. Lee W. Y., Jiang C. X., Indro D.C. (2002) Stock market volatility, excess returns, and the role of investor of sentiment. Journal of Banking and Finance, 26(12):2277-2299. https://doi.org/10.1016/S0378-4266(01)00202-3.
10. Brauwers G., Frasincar, F. A general survey on attention mechanisms in deep learning, IEEE Trans. Knowl. Data Eng., 35 (2023) 3279-3298. https: //doi. org/ 10. 1109/ TKDE. 2021. 3126456.
11. Yang S.S. (2023). Industry Index Top and Bottom Signals: Net New High Ratio ((NH-NL)%). Research Institute, Huafu Securities.