Scholar Publishing Group
Socio-Economic Statistics Research
https://doi.org/10.38007/SESR.2025.060220
ISSN 2790-2722 Vol. 6, Issue 2: 185-193

Research on Financial Systemic Risk Measurement Based on Investor Sentiment and Network Text Mining

Jiahe Sun

Tepper School of Business, Carnegie Mellon University, Pittsburgh 15213, Pennsylvania, U.S.

Keywords: Financial systemic risk measurement, investor sentiment, multi-scale feature extraction, graph neural network, hybrid investment strategy

Abstract: The research on financial systemic risk measurement based on investor sentiment and network text mining focuses on the core value of financial time series prediction in policy formulation, investment decision-making, and risk management. It addresses the challenges posed by its high volatility, multi-scale characteristics, and nonlinear relationships to traditional methods. The research background points out that traditional statistical models such as ARIMA and GARCH are limited by linear assumptions and stationarity requirements, making it difficult to capture the nonlinear dependencies and dynamic factors of multivariate sequences; Although deep learning models such as LSTM and Transformer enhance non-linear processing capabilities, they suffer from black box problems and insufficient interpretability, as well as inadequate utilization of textual information such as investor sentiment. The research method innovatively integrates multi-scale feature extraction (extreme symmetric mode decomposition, ESMD), transfer entropy causal relationship modeling, and graph neural network (GCN). By using transfer entropy to construct causal relationship graphs between variables, interpretability is enhanced. Combining ESMD denoising with FEDformer's long-term prediction ability to strengthen temporal adaptability, risk warning indicators are finally constructed and their application value is verified. Research has shown that the proposed model significantly outperforms five deep models, including LSTNet, MTGNN, and Transformer, in stock price sequence prediction (validated by DM test and RMSE, MAE, and other indicators); Based on the predicted results of stock selection, the mean variance, mean absolute deviation, mean CVaR, and entropy enhanced mean CVaR model are used for portfolio allocation. Under the constraints of prohibiting short selling and transaction costs, the "predicted stock selection+portfolio" hybrid strategy can achieve higher returns. This study integrates investor sentiment text mining and network information to broaden the risk measurement data source, improve model interpretability and temporal adaptability, and provide a scientific framework for accurate measurement of financial systemic risks. Future research can be expanded to non numerical data quantification, cross domain prediction, and multi-objective optimization of investment portfolio construction, further optimizing risk measurement and allocation effects.

1. Introduction

This study focuses on the field of financial systemic risk measurement[1], which is rooted in the core value of financial time series forecasting in policy-making, investment decision-making, and risk management. Its high volatility, multi-scale characteristics, and nonlinear relationships require high demands for forecasting methods. The challenges in previous literature mainly lie in: traditional statistical models such as ARIMA and GARCH are limited by linear assumptions and stationarity requirements, making it difficult to capture the nonlinear dependencies of multivariate sequences and dynamic factors such as investor sentiment; Although deep learning models such as LSTM and Transformer enhance non-linear processing capabilities, they suffer from black box problems and insufficient interpretability, as well as inadequate utilization of investor sentiment information in network text mining; In the prediction of multivariate time series, the extraction and quantification of complex correlations between variables (such as the causal relationship between investor sentiment and market indicators) still need breakthroughs, and high-dimensional and high noise data limit the accuracy of risk measurement. The motivation for work comes from combining investor sentiment (quantified through text mining) with online text information to construct a more accurate financial systemic risk measurement model, addressing the limitations of traditional methods in dynamic risk capture. The paper aims to propose a financial systemic risk measurement framework based on investor sentiment and network text mining, integrating multi-scale feature extraction[2](such as ESMD), transmission entropy causal relationship modeling, and graph neural network (GCN) to achieve dynamic identification and quantification of risk factors. The core contributions include: innovating the integration of investor sentiment text mining and network information, and expanding the sources of risk measurement data; Constructing causal relationship diagrams between variables using transfer entropy to enhance model interpretability; Combining ESMD denoising with FEDformer's long-term prediction capability to enhance the temporal adaptability of risk measurement; Verify the superiority of the model in systematic risk prediction through DM testing and multiple evaluation indicators; Finally, based on the predicted results, risk warning indicators are constructed to empirically demonstrate their application value in financial institutions and market regulation.

2. Correlation theory

2.1 An innovative time series prediction model based on ESMD-FEDformer fusion

Pole symmetric mode decomposition (ESMD) is a new development of empirical mode decomposition (EMD) algorithm. As an adaptive digital signal decomposition method, it optimizes the final residual mode to the "adaptive global mean" of the entire data through the "least squares" idea, thereby determining the optimal number of filters for nonlinear non-stationary time series analysis. The specific steps are: selecting a time series$\mathrm{Y(u)}$， Set the filtering frequency$L_b$and Mark the local maximum and minimum points and connect adjacent extreme points with n remaining extreme points, and mark the midpoint $N_j(1 \leq j \leq o-1)$, fill in the midpoint of the left and right boundaries$N_0$and N_OUse cubic spline interpolation to obtain the sum of curves for the midpoint of odd and even coordinates, and calculate the average curve $Q^* = (I_1 + I_2)/2$, repeat the process until the screening times are less than the preset allowable error, and get the first modal component $E_1(u)$,Continue repeating until the trend residual$\mathrm{T(u)}$obtain all decomposition results by reaching the number of remaining extremum points.By calculating the variance ratio$\mathrm{R = g/g_0}$, when R the minimum decomposition effect is optimal, and the final time series can be represented as

$$\mathrm{Y(u)} = \sum \mathrm{E_r(u)} + \mathrm{C(u)}$$

Frequency enhanced decomposition transformer (FEDformer) [4]combines Transformer with seasonal trend decomposition method, using Transformer to capture the global pattern and detailed structure of time series, and proposes incremental frequency Transformer. Its complexity is linearly related to sequence length, including frequency enhancement module (FEB), frequency enhanced attention mechanism (FEA), and mixed expert module (MOEDecomp) [5].

2.2 Collaborative methodology for financial knowledge graph representation and classification

Transmitting entropy, as an extension of information entropy, quantifies the causality between two variables through asymmetric computation. Its discrete expression is

$$U(z \to y) = \sum\nolimits_{yo,yo+h,zo} Q(yo+h, yo, zo)\log\frac{Q(yo+h, yo, zo)Q(yo)}{Q(yo, zo)Q(yo+h, yo)}$$

Suitable for discrete data and capable of accurately capturing linear and nonlinear causal relationships, it has more advantages than Granger causality analysis. Graph Convolutional Networks (GCNs) achieve feature learning and transfer of non Euclidean graph structured data through graph convolution operations, but deep networks are prone to oversmoothing problems (node feature convergence). The causal relationship matrix between nodes generated by entropy transfer can be directly used as the adjacency matrix input of GCN, and the two work together to construct a complete framework from causal relationship modeling to feature transfer, providing theoretical support and methodological innovation for the prediction and risk measurement of complex time series.

3. Research method

3.1 Transfer Entropy GCN-ESMD-FEDformer Prediction Model

The model framework is based on the construction of graph convolutional networks using transfer entropy. The process of processing financial time series data is as follows: firstly, pole symmetric mode decomposition (ESMD) is used to decompose the multivariate sequence, obtaining intrinsic mode functions (IMF) and residual sequences at different time scales. These components are mapped into graph node features (each dimension corresponds to a node), effectively reducing noise interference; Secondly, by calculating the transfer entropy matrix between the decomposed components and quantifying the causal relationships between variables, a node adjacency matrix is generated as the input for the graph structure; Subsequently, the graph convolutional network (GCN) [6] is used to update node features by sharing node information through information propagation mechanisms, filtering key features through attention mechanisms, enhancing inter node connections, and reducing noise impact; Finally, the time relationship of node embedding is established by combining the frequency enhanced decomposition transformer (FEDformer), and the sequence length is normalized by residual connections and standard convolutional layers to output the final predicted value. This design achieves efficient and accurate prediction of complex financial time series through the collaboration of multi-scale feature extraction, causal relationship modeling, and time relationship modeling.

3.2 Entity and Relationship Model Design and Data Preprocessing

This empirical study focuses on multiple time series prediction methods and verifies the effectiveness of the model through a three-stage framework. At the data level, the trading data of four representative stocks in a certain index from December 2003 to December 2023 were selected.

The data was divided into a training set, a validation set, and a testing set using a 6:2:2 ratio, and the impact of dimensionality was eliminated using maximum minimum normalization. By using the ESMD algorithm to decompose the opening price, closing price, highest price, and lowest price, each stock generates 10 IMF components and 1 residual term, forming 16 sets of decomposition results. The experimental setup adopts a universal operating system, mainstream graphics cards, and deep learning frameworks, and optimizes hyperparameters through grid search: the optimal sliding window size for stock A is 7, and stock B is 5; In the combination of hidden dimension and output dimension of graph convolutional network (GCN), stock A has the lowest root mean square error (RMSE) in 3 × 4 dimensions [6]. The model prediction results are visually displayed through the fitting curves of the closing prices of four stocks, verifying the generalization ability of the ESMD GFormer model driven by transfer entropy in financial time series prediction. Further validate the contribution of each module through ablation experiments (as shown in Table 1)

Table 1. Comparison of ablation experiment results

Stock code	Types of indicators for medium and long-term loan ratios	Method1	Method2	Method3	Method4	Proposed
600028	RMSE	0.8164	0.6528	0.7602	0.7281	0.5743
600028	MAE	0.7677	0.6001	0.7054	0.6756	0.5193
600030	RMSE	0.5974	0.4983	0.6142	0.6481	0.4532
600030	MAE	0.5412	0.4470	0.5667	0.5899	0.3961
600031	RMSE	0.3139	0.2614	0.3273	0.3642	0.2209
600031	MAE	0.2677	0.2167	0.2769	0.3101	0.1755
600036	RMSE	0.3653	0.2206	0.3134	0.2921	0.1683
600036	MAE	0.2971	0.1787	0.2672	0.2167	0.1285

The complete model outperforms the four variant methods in terms of RMSE/MAE indicators for all four stocks. Among them, replacing the transfer entropy matrix with the all 1 matrix (Method1) leads to a significant increase in prediction error; Removing the time relationship modeling module (Method2) or directly using the original data (Method3) both lead to a decrease in performance; Wavelet decomposition replacing multi-scale decomposition method (Method 4) is not as effective as the original method. Experiments have shown that multi-scale feature extraction, causal relationship modeling, spatial feature fusion, and temporal relationship modeling collectively constitute key elements for improving model performance. In summary, through systematic experimental design and multidimensional verification, this paragraph confirms the effectiveness of the proposed method in stock price prediction tasks and the synergistic effects of various components, providing new methodological support for financial time series prediction. The chart numbering and experimental framework remain unchanged to ensure the complete presentation of research logic and visualization results.

3.3 Comparative Analysis of Financial Time Series Prediction Models

This article focuses on financial time series prediction tasks and systematically compares the core mechanisms and performance of traditional statistical models and deep learning models. The ARIMA model processes non-stationary sequences through a differential autoregressive moving average process and uses BIC criteria for parameter selection, making it suitable for capturing linear trends; DeepAR, as an RNN based autoregressive model, uses long short-term memory or gated recurrent unit architecture to learn periodic features and models future sequences through conditional probability distribution. LSTNet adopts a combination architecture, combining convolutional modules to extract local dependencies, recursive modules to capture temporal features, attention mechanisms to allocate weights, and autoregressive modules to enhance linear feature recognition, achieving rolling prediction of multivariate time series. The Transformer model is based on the Encoder Decoder structure and utilizes multi head attention mechanism, feedforward neural network, and normalization layer to process sequence data, making it suitable for long-range dependency modeling. MTGNN, as a spatiotemporal graph neural network, generates adjacency matrices through adaptive graph learning layers, integrates inter node information through graph convolution modules, expands the field of view through temporal convolution modules, and optimizes feature extraction through residual and skip connections, effectively capturing spatiotemporal correlations and nonlinear trends of multivariate sequences. Through comparative experiments, it has been verified that each model exhibits different advantages in financial time series prediction, providing a benchmark reference for the performance evaluation of combined models.

4. Results and discussion

4.1 Construction of Investment Portfolio Optimization Model and Multi dimensional Empirical Evaluation

Modern portfolio theory constructs an optimization framework through core mathematical methods that quantify returns and risks. The Mean Variance (MV) model is based on weighted sum of expected asset returns to measure returns and variance/standard deviation to measure risk. It minimizes risk or vice versa under a given return through quadratic programming, subject to constraints such as no short selling and a sum of weights of 1. However, its application is limited by the assumption of normal distribution of returns and complex calculations. To overcome limitations, the Mean Absolute Deviation (MAD) model replaces variance with absolute deviation, transforming the problem into linear programming and suitable for non normal distribution scenarios; The Conditional Value at Risk (CVaR) model focuses on tail risk and minimizes the expected loss beyond VaR at a specific confidence level through linear programming, becoming a representative measure of risk consistency; Furthermore, the mean CVaR entropy model introduces information entropy to supplement uncertainty, and optimizes the objective through a weighted combination of entropy and CVaR (weights determined by investors' risk aversion level μ). The entropy value reflects investors' understanding of the problem - the higher the entropy, the less information and the more concentrated the investment. When the entropy is log (o)/o, it corresponds to equally weighted diversified investment, thus incorporating system uncertainty considerations in risk measurement.Empirical research has shown that the proposed ESMD-GFormer model based on transfer entropy exhibits advantages in predicting multiple stocks: its predictive metrics (such as RMSE, MAE) are superior to Transformer, LSTNet, DeepAR, ARIMA, and other methods, and the training time efficiency is significant; The statistical significance of the differences between models was further verified through DM testing, and the deep learning model outperformed traditional methods in terms of stability. In the application of stock price prediction, this model selects ten stocks with high returns and low correlation for predicting the returns of the 50 constituent stocks of the Shanghai Stock Exchange. The correlation coefficient matrix shows a high degree of diversification, effectively achieving the goal of risk diversification and providing scientific support for quantitative investment. These models together form the core framework for portfolio optimization, adapting to diverse investment needs through different risk measures and constraints.

4.2 Model experiment

Modern portfolio theory quantifies returns and risks through mathematical methods and constructs optimization models to balance the relationship between the two. The Mean Variance (MV) model, as the foundational theory, measures portfolio returns by weighted sum of expected asset returns, and risk by variance or standard deviation. It minimizes risk or maximizes return under a given return through quadratic programming, subject to constraints such as no short selling and a weight sum of 1. However, its application is limited by the assumption of a normal distribution of returns and complex calculations. To overcome limitations, the Mean Absolute Deviation (MAD) model replaces variance with absolute deviation and transforms the problem into linear programming. It is suitable for non normal distribution scenarios and is equivalent to MV when returns are normally distributed. The Conditional Value at Risk (CVaR)错误!未找到引用源。 model focuses on tail risk and minimizes the expected loss beyond VaR at a specific confidence level through linear programming, becoming a representative measure of risk consistency. Furthermore, the mean CVaR entropy model introduces information entropy to supplement uncertainty measures, optimizing the objective through a weighted combination of entropy and CVaR (weights determined by investors' risk aversion level μ). Entropy reflects investors' understanding of the problem - the higher the entropy, the less information and more concentrated the investment. When the entropy is log (o)/o, it corresponds to equally weighted diversified investment, thus incorporating consideration of system uncertainty in risk measurement. These models together form the core framework for portfolio optimization, adapting to diverse investment needs through different risk measures and constraints.

4.3 Effect analysis

This paragraph focuses on the application of financial time series forecasting results in portfolio allocation, and empirically verifies the effectiveness of mixed strategies. Research on the ESMD GFormer model based on transfer entropy to predict the top ten stocks with the highest returns and low correlation in the Shanghai Stock Exchange 50 from January 2020 to December 2023. Four models, namely mean variance (MV), mean absolute deviation (MAD), mean CVaR, and mean CVaR entropy based on information entropy, were used for configuration, with no consideration/consideration of transaction cost scenarios. When transaction costs are not considered, each model generates the minimum risk portfolio at different risk-free return levels (s0)错误!未找到引用源。 (as shown in Table 2):

As s0 increases, actual returns may not necessarily increase synchronously, and the effectiveness of the strategy under short selling constraints is significant - for example, when s0=0.020, the MV model assigns all weights to the third stock (x3=1.000), while the Mean CVaR entropy model 错误!未找到引用源。 achieves higher returns through decentralized allocation (x3=0.6736, x8=0.2036), and this model outperforms traditional MV models in most s0 levels, indicating more accurate risk measurement under non normal distribution assumptions. After considering transaction costs, the allocation weights and risk indicators (τ q) are significantly adjusted: transaction costs directly reduce returns, and expected returns and costs need to be balanced - for example, when s0=0.000, the MV model weights are adjusted from a dispersed state without cost scenarios to a centralized allocation (x3=0.1672, x5=0.1833), while the risk (τ q=0.0032) increases; Model comparison shows that the MV and Mean CVaR entropy models have lower risk under the same s0, which conforms to the low-risk, low return rule. Empirical evidence shows that a hybrid strategy combining predictive stock selection and investment portfolio can provide returns higher than market indices. When considering trading costs, it is necessary to balance returns and costs through optimization strategies such as risk management and transaction cost control. This validates the applicability of the model in practical allocation and provides practical support for measuring systemic financial risks.

Table 2.Column chart comparing the squared errors of various models without considering costs

Model name	S0 value	Square error (sq)
MV Model	0	0.00030
MV Model	0.005	0.00036
MV Model	0.010	0.00127
MV Model	0.015	0.00118
MV Model	0.020	0.00051
MV Model	0.025	0.00031
MAD Model	0	0.00042
MAD Model	0.005	0.00079
MAD Model	0.010	0.00135
MAD Model	0.015	0.00104
MAD Model	0.020	0.00052
MAD Model	0.025	0.00037
Mean-CVaR Model	0	0.00044
Mean-CVaR Model	0.005	0.00081
Mean-CVaR Model	0.010	0.00138
Mean-CVaR Model	0.015	0.00109
Mean-CVaR Model	0.020	0.00052
Mean-CVaR Model	0.025	0.00038
Mean-CVaR-Entropy Model	0	0.00077
Mean-CVaR-Entropy Model	0.005	0.00150
Mean-CVaR-Entropy Model	0.010	0.00113
Mean-CVaR-Entropy Model	0.015	0.00051
Mean-CVaR-Entropy Model	0.020	0.00050

5. Conclusion

This paper proposes a multivariate time series prediction graph neural network model based on multi-scale time feature extraction and attention mechanism, and verifies its effectiveness in financial time series prediction. The model extracts temporal features of multivariate time series at different time scales through extreme symmetric mode decomposition as node features of the graph. It combines transfer entropy to construct a node adjacency matrix to identify causal relationships between variables. A graph convolutional neural network is used to generate node embeddings containing spatial relationships. Finally, the temporal relationships of node embeddings are established through FEDformer to achieve multivariate prediction. Empirical research was conducted on four stocks listed in the Shanghai Stock Exchange, and the significance and overall predictive ability of each module of the model were verified through simulation and ablation experiments. By comparing five deep models including LSTNet, MTGNN, and Transformer, combined with DM testing and evaluation indicators such as RMSE and MAE, it is shown that this model performs better in stock price sequence prediction. After selecting stocks based on the predicted results, four investment portfolio models including mean variance, mean absolute deviation, mean CVaR, and mean CVaR based on information entropy were used for allocation under the constraints of not allowing short selling and transaction costs. Empirical evidence shows that the mixed strategy of "predicted stock selection+investment portfolio" can achieve higher returns. Future research can be expanded in three aspects: firstly, incorporating non numerical data quantification methods such as technical indicators and emotional factors to improve prediction accuracy; The second is to extend the model to other fields such as power load forecasting and exchange rate forecasting; The third is to introduce multi-objective optimization algorithms and timing strategies in portfolio construction to optimize the effectiveness of multi-level allocation.

References

[1] Chen L. Research on Systemic Financial Risk Measurement and Early Warning Model Based On TVP-SV-VAR Model. Procedia Computer Science, 2025, 262(000):868-877.

[2] Hu Z, Luo K, Liu Y. Classification of motor imagery based on multi-scale feature extraction and fusion-residual temporal convolutional network. 2025.

[3] [Cao B, Jiang X, Leong D, et al. EMD-Fuzzy: An Empirical Mode Decomposition Based Fuzzy Model for Cross-Stimulus Transfer Learning of SSVEP. 2025.

[4] Liu X, Qiu B, Cao J, et al. Freqformer: Image-Demoir\\'eing Transformer via Efficient Frequency Decomposition. 2025.

[5] Liang S, Gu Y. Multi-Modal Dysarthria Severity Assessment Using Dual-Branch Feature Decoupling Network and Mixed Expert Framework. 2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2024:126-130.

[6] Liu, Y. (2025). The Importance of Cross-Departmental Collaboration Driven by Technology in the Compliance of Financial Institutions. Economics and Management Innovation, 2(5), 15-21.

[7] Huang, J. (2025). Research on Resource Prediction and Load Balancing Strategies Based on Big Data in Cloud Computing Platform. Artificial Intelligence and Digital Technology, 2(1), 49-55.

[8] Jiang, Y. (2025). Application and Practice of Machine Learning Infrastructure Optimization in Advertising Systems. Journal of Computer, Signal, and System Research, 2(6), 74-81.

[9] Zou, Y. (2025). Automated Reasoning and Technological Innovation in Cloud Computing Security. Economics and Management Innovation, 2(6), 25-32.

[10] Qi, Y. (2025). Data Consistency and Performance Scalability Design in High-Concurrency Payment Systems. European Journal of AI, Computing & Informatics, 1(3), 39-46.

[11] An, C. (2025). Study on Efficiency Improvement of Data Analysis in Customer Asset Allocatior. Journal of Computer, Signal, and System Research, 2(6), 57-65.

[12] Huang, J. (2025). Optimization and Innovation of AI-Based E-Commerce Platform Recommendation System. Journal of Computer, Signal, and System Research, 2(6), 66-73.

[13] Zhang, X. (2025). Optimization of Financial Fraud Risk Identification System Based on Machine Learning. Journal of Computer, Signal, and System Research, 2(6), 82-89.

[14] Wang, Y. (2025). Exploration and Clinical Practice of the Optimization Path of Sports Rehabilitation Technology. Journal of Medicine and Life Sciences, 1(3), 88-94.

[15] Sheng, C. (2025). Innovative Application and Effect Evaluation of Big Data in Cross-Border Tax Compliance Management. Journal of Computer, Signal, and System Research, 2(6), 40-48.

[16] Sheng, C. (2025). Research on the Application of AI in Enterprise Financial Risk Management and Its Optimization Strategy. Economics and Management Innovation, 2(6), 18-24.

[17] *Tu, X. (2025). Optimization Strategy for Personalized Recommendation System Based on Data Analysis. Journal of Computer, Signal, and System Research, 2(6), 32-39.*

[18] *Zhu, P. (2025). The Role and Mechanism of Deep Statistical Machine Learning In Biological Target Screening and Immune Microenvironment Regulation of Asthma. arXiv preprint arXiv:2511. 05904.*

[19] *Liu, B. (2025). Design and Implementation of Data Acquisition and Analysis System for Programming Debugging Process Based On VS Code Plug-In. arXiv preprint arXiv:2511. 05825.*

[20] *Design and Implementation of a Cloud Computing Security Assessment Model Based on Hierarchical Analysis and Fuzzy Comprehensive Evaluation*