Prediction of stock market volatility based on public opinion mining and multi-source data fusion: A case study of Guoxuan Hi-Tech

Wenyuan Tian[#, *, 1], Liwei Liu[#, 2], Zetong Li[#, 2]

[1]College of Mechanical Engineering, Yanshan University, Qinhuangdao, China, 066004

[2]College of Civil Engineering and Mechanics, Yanshan University, Qinhuangdao, China, 066004

*Corresponding author: 15513737676@163.com

#These authors contributed equally.

Abstract. This study aims to enhance the accuracy and robustness of stock market volatility forecasting in the context of rapidly evolving public opinion and heterogeneous data environments. Traditional models often struggle to capture the nonlinear and dynamic interactions between investor sentiment and market behavior, particularly during high-volatility periods. To address these limitations, a unified deep learning framework is developed that incorporates multi-source sentiment signals into the forecasting process. Specifically, a Transformer–LSTM threshold fusion model is constructed to extract semantic features from investor comments and temporal patterns from historical market data. The model introduces a dynamic fusion mechanism based on historical volatility thresholds and integrates a dual-dimensional sentiment indicator system (CAI and CSEI) to reflect behavioral signals from both current and potential investors. Empirical evaluation using Guoxuan Hi-Tech stock data from 2019 to 2024 shows that the proposed model significantly outperforms benchmark methods such as BERT–GRU and LDA–Prophet in predictive accuracy, generalization, and robustness. These findings underscore the effectiveness of integrating deep learning with multi-source sentiment modeling and suggest a scalable framework for broader applications in financial risk forecasting.

Keywords: Volatility Forecasting, Public Opinion Mining, Data Fusion, Transformer-LSTM.

1. Introduction

In recent years, the proliferation of social media and the rapid dissemination of information have made investor sentiment a critical factor influencing stock market fluctuations. Traditional volatility prediction models relying solely on structured data have increasingly revealed their limitations, particularly in responding to irrational behavior and public opinion shocks. To address challenges such as high-frequency volatility, difficulties in integrating heterogeneous data, and the lack of sentiment transmission modeling, recent research has increasingly focused on integrating multi-source data with deep learning techniques to improve the timeliness and robustness of market predictions. Wang Jiaxiang et al. (2025) proposed an enhanced Jump Point Search algorithm that significantly improves volatility prediction efficiency in high-dimensional financial data. It reduces expanded nodes and runtime, enhancing suitability for real-time financial applications. However, it still struggles with complex non-linear patterns and large-scale datasets [1]. Yu Jiaxiong (2024) proposed a sentiment-based volatility model using the Baidu Index, Economic Uncertainty Index, and scaled principal component analysis. The model improves forecasting accuracy, but its applicability beyond the Chinese market requires further validation [2]. Du Yiming (2024) examined how COVID-19 affected volatility in the Science and Technology Innovation Board, introducing a sentiment index and the GARCH-MIDAS model. Higher-frequency sentiment indicators showed stronger effects. The study provides evidence of sentiment-volatility linkage during the pandemic, though limited by sample size and indicator scope [3]. Chen Jian (2024) investigated how jump clustering and information flows affect stock price efficiency. Both factors were found to significantly affect pricing efficiency. Incorporating these factors into volatility models improved forecasting accuracy. However, the method is market-specific and may not generalize across asset classes [4]. Chen and Zhang (2023) investigated the impact of investor uncertainty on stock market volatility forecasting. They found investor uncertainty significantly influences forecasting accuracy. Incorporating investor sentiment and behavioral factors enhanced the model's explanatory power. However, model performance depends heavily on sentiment data quality and availability, limiting practical application [5]. Guo Zhipeng (2021) investigated the influence of social media sentiment on stock market volatility and found that negative sentiment significantly affects short-term volatility. The study utilizes social media data to capture sentiment and offers new perspectives for volatility forecasting; however, its effectiveness in predicting long-term volatility remains underexplored [6].

The main contributions of this study are threefold: (1) a Transformer-LSTM threshold fusion model is proposed, which integrates public opinion sentiment with financial time-series data and enables dynamic adjustment of feature importance based on market volatility levels; (2) a dual-dimensional investor sentiment indicator system (CAI and CSEI) is constructed to distinguish between current stockholders and potential investors, thereby enriching sentiment representation in volatility forecasting; and (3) a comprehensive comparative analysis is conducted against benchmark models such as BERT-GRU and LDA-Prophet, demonstrating superior accuracy, generalization, and robustness under complex and high-volatility market conditions.

This study integrates daily stock data of Guoxuan Hi-Tech (2019–2024) with investor comments from Eastmoney and Xueqiu to construct a sentiment-driven volatility prediction model. A custom sentiment dictionary and pre-trained word embeddings are used to derive CAI and CSEI indices. The model combines Transformer for textual semantics, LSTM for market dynamics, and a volatility-threshold mechanism for adaptive fusion. Performance is evaluated through short-term volatility forecasting against benchmark models.

2. Related Theories

This paper proposes a Transformer-LSTM threshold fusion model. The model integrates the Transformer architecture from natural language processing with LSTM networks for financial time-series modeling, aiming to dynamically adjust the weighting between sentiment and market behavior. It constructs two parallel channels: one employs multi-layer Transformers to extract high-dimensional sentiment vectors from public opinion texts; the other encodes historical market indicators via LSTM to capture their temporal dynamics. A nonlinear fusion layer with a historical volatility threshold adaptively adjusts the contribution of both channels to volatility forecasting[7].

The flow chart of the Transformer-LSTM threshold fusion model is shown in Figure 1.

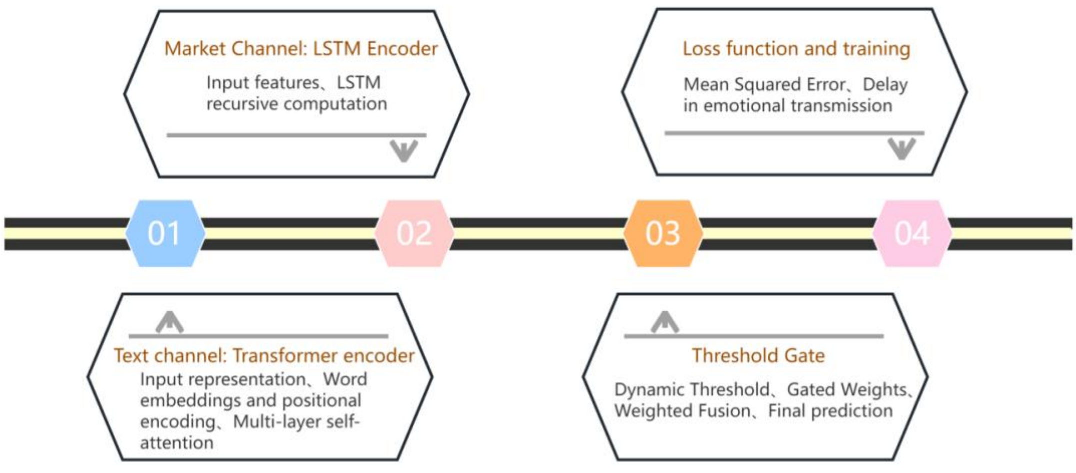

Figure 1. Transformer-LSTM threshold fusion model flow chart

A Transformer encoder is applied to the text channel. Comment texts from day t are tokenized and padded to a fixed length L, yielding a sequence of word indices:

$$D_t = \lfloor w_{t,1}, w_{t,2}, ..., w_{t,L} \rfloor \tag{1}$$

Using pre-trained Chinese word embeddings:

$$X_t = [e(w_{t,1}), ..., e(w_{t,L})] + p \tag{2}$$

Here, $e(\cdot) \in R^d$ denotes the word embedding, and $p \in R^d$ denotes the positional encoding.

Each of the N Transformer encoder layers comprises multi-head self-attention (with h heads) and a feedforward network:

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

$$H_t^{(l)} = \mathrm{FFN}\left(\mathrm{Attn}\left(H_t^{(l-1)}W^Q, H_t^{(l-1)}W^K, H_t^{(l-1)}W^V\right)\right) \tag{4}$$

Finally, the output at position L is extracted, or a pooling operation is applied to obtain the text sentiment vector:

$$E_t = H_t^{(N)} \in R^{d_e} \tag{5}$$

An LSTM encoder is applied to the market channel. On day t, p -dimensional market features (e.g., opening price, high, low, volume, realized volatility RV) are concatenated into a vector $X_t \in R^p$. The time series $\{X_{t-T+1}, ..., X_t\}$ is concatenated along the temporal dimension. The resulting sequence is used as the input to the LSTM unit:

$$i_t = \sigma(W_i[X_t, h_{t-1}] + b_i) \tag{6}$$

$$f_t = \sigma(W_f[X_t, h_{t-1}] + b_f) \tag{7}$$

$$o_t = \sigma(W_o[X_t, h_{t-1}] + b_o) \tag{8}$$

$$\tilde{c}_t = \tanh(W_c[X_t, h_{t-1}] + b_c) \tag{9}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{10}$$

$$h_t = o_t \odot \tanh(c_t) \tag{11}$$

At the final time step, $h_t \in R^{d_m}$ represents the market state vector:

$$M_t = h_t \tag{12}$$

A threshold-based fusion is then applied. The threshold is defined as the $\tau -$ percentile of realized volatility $\{RV_{t-K}, ..., RV_{t-1}\}$ over the past \$K\$ trading days:

$$\gamma_t = \mathrm{Quantile}\left(\{RV_{t-K}, ..., RV_{t-1}\}, \tau\right) \tag{13}$$

Here, $K = 30$ and $\tau = 0.9$ are selected.

After concatenating the threshold value γ_t with the current atmosphere vector E_t and market state vector M_t , the fusion weight is computed through a fully connected layer followed by a Sigmoid activation function:

$$u_t = [E_t; M_t] \in R^{d_e+d_m} \tag{14}$$

$$Z_t = \sigma(W_z u_t + b_z) \in (0,1)^{d_z} \tag{15}$$

Here, $W_z \in R^{(d_e+d_m)d_z}$ and $b_z \in R^{d_z}$ denote the weight matrix and bias term, respectively.

Threshold-controlled weighted fusion of the two channels based on Z_t :

$$H_t = Z_t \odot E_t + (1 - Z_t) \odot M_t \tag{16}$$

The fusion vector H_t is fed into a regression layer to produce the predicted volatility for day t:

$$\hat{RV}_t = W_o \phi(H_t) + b_o \tag{17}$$

Here, $\phi(\cdot)$ denotes the activation function, which can be chosen as Swish or ReLU, and W_o , b_o are the parameters of the regression layer.

Finally, mean squared error (MSE) is adopted as the primary loss, combined with L_2 regularization or a sentiment propagation delay constraint:

$$\mathcal{L} = \frac{1}{N}\sum_{t=1}^{N}\left(RV_t - \widehat{RV}_t\right)^2 + \lambda\mathcal{R}(\Theta) \tag{18}$$

The regularization term $\mathcal{R}(\Theta)$ may include weight decay and threshold smoothing components. All parameters $\Theta = \{W^Q, W^K, W^V, \dots, W_o, b_o\}$ are iteratively updated using the Adam optimizer during training.

3. Experments

This study takes "Guoxuan Hi-Tech" stock as a case study, integrating social media sentiment analysis with time-series modeling to forecast stock market volatility. By applying text mining techniques, it distinguishes between two categories of investor sentiment and constructs the CAI and CSEI indices. Furthermore, by leveraging a Transformer-LSTM threshold fusion model, it dynamically adjusts the weights of textual and market features to enable nonlinear modeling[8].

This study collects stock data on Guoxuan Hi-Tech from multiple sources, including daily closing prices, highest and lowest prices, opening prices, and trading volumes. The structured data is obtained from Xueqiu.com, covering the period from 2019 to 2024. In addition, unstructured data, such as investor comments, are collected from Eastmoney.com. Comment timestamps, user IDs, and textual content are integrated to ensure data completeness.

The sentiment of the comment text is quantified. First, the CNKI HowNet sentiment dictionary and the Jiang Fuwei sentiment dictionary are used as the initial lexicons. The comment texts are segmented using the Jieba library in Python, and the Word2Vec algorithm is applied to extract semantically related terms to expand the dictionaries. Finally, a stock-specific sentiment lexicon is constructed through manual screening and deduplication[9].

The flow chart of the stock sentiment dictionary construction process is shown in Figure 2.

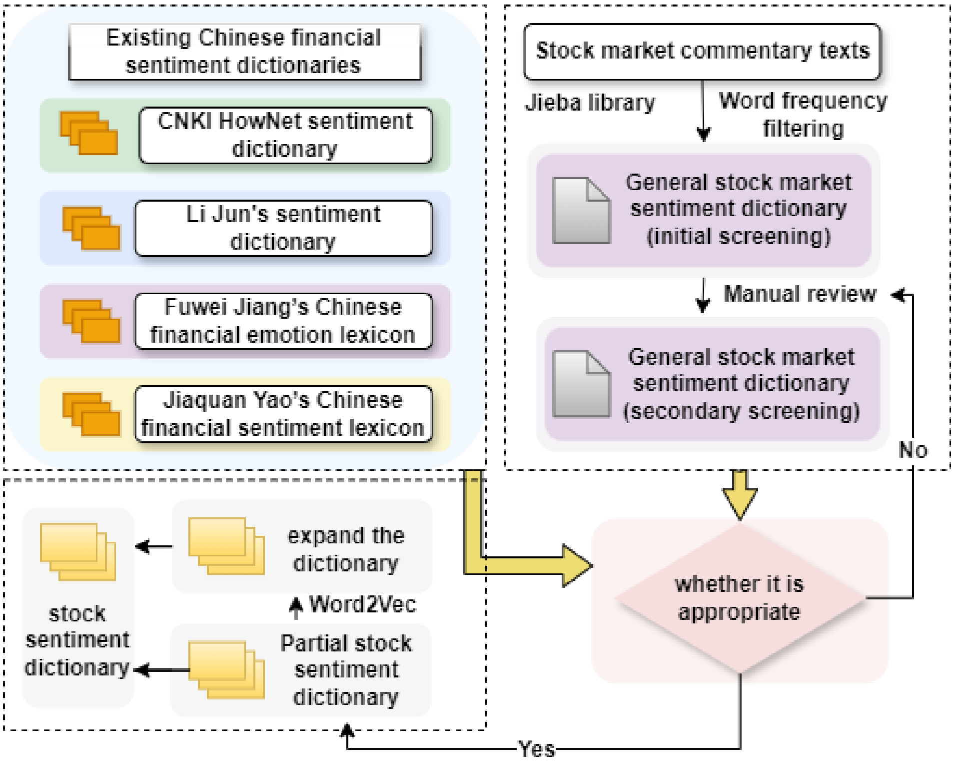

Figure 2. The stock sentiment dictionary construction flow chart

The stock sentiment dictionary is constructed, a rule base is defined, and sentiment scores are assigned accordingly: positive words are assigned a weight of +1, while negative words are assigned a weight of –1. To account for the fact that double negatives may imply affirmation, the weight of a negative word is defined as $(-1)^n$, where n represents the number of times the negative word appears. Therefore, the sentiment score for the i-th comment is calculated as follows:

$$S_i = N^{-1} \sum_{j=1}^{N} W_j (-1)^n \tag{19}$$

Here, N denotes the total number of sentiment words in the i-th comment, and W_j represents the sentiment score of the j-th word:

$$W_j = \begin{cases} 1, & positive \\ -1, & negative \end{cases} \tag{20}$$

Building on existing research on investor sentiment measurement, this study introduces an innovative classification of investors into current holders and potential investors, and accordingly constructs a dual-dimensional indicator system comprising a sentiment expectation index and an attention index, as illustrated in the Figure 3:

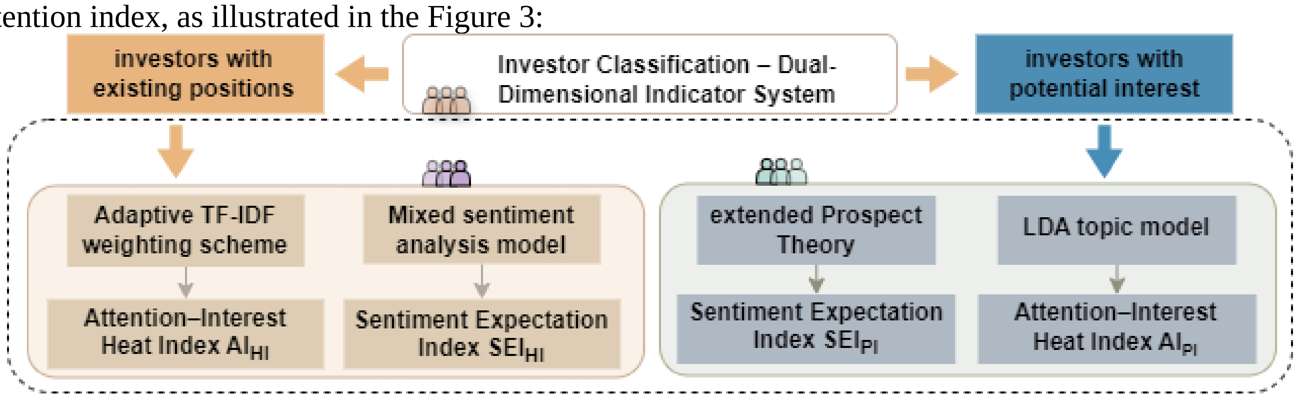

Figure 3. The Construction of Two-Type Investor Sentiment Indicator System flow chart

Investors who currently hold or have previously traded the stock (i.e., those who have bought, sold, or engaged in other trading activities involving the stock) are associated with two sentiment indicators: the Attention and Interest Heat Index (AI_{HI}) and the Sentiment Expectation Index (SEI_{HI}), which are constructed to capture their behavioral tendencies.

For the attention index AI_{HI}, the dynamic TF-IDF algorithm is used to extract the high-frequency topic words contained in the comments of investors holding positions, and the daily attention intensity is calculated by the proportion of word frequency:

$$AI_{HI,t} = \frac{\sum_{i=1}^{n} w_i \cdot \log(f_{i,t}+1)}{\sum_{j=1}^{m} \log(f_{j,t}+1)} \times 100 \tag{21}$$

Here, w_i denotes the industry-adjustment coefficient for the i-th keyword, calibrated from the corpus following the method proposed by Loughran and McDonald.

For the Sentiment Expectation Index SEI_{HI}, a four-dimensional sentiment vector is constructed based on the hybrid sentiment model proposed by Bosch and Pradhan:

$$SEI_{HI,t} = \frac{1}{N}\sum_{k=1}^{N}[\alpha P_k + \beta N_k + \gamma U_k - \delta A_k] \tag{22}$$

Where P_k (positive), N_k (negative), U_k (uncertainty), and A_k (anxiety) represent the components of the sentiment vector, whose weight coefficients are determined via principal component analysis (PCA) and calibrated using Granger causality testing, yielding $\alpha = 0.42$, $\beta = -0.35$, $\gamma = 0.18$, and $\delta = 0.05$.

Potential investors are defined as individuals who have not yet engaged in any substantive trading activities involving the stock and are currently in the information-gathering or observation stage. Two sentiment indicators are constructed for this group: the Attention and Interest Propensity Index (AI_{PI}) and the Sentiment Expectation Propensity Index (SEI_{PI}).

For the Attention and Interest Propensity Index AI_{PI}, the Latent Dirichlet Allocation (LDA) topic model is employed to identify the latent Dirichlet structure in potential investors' discussions, resulting in the extraction of four major topics: entry timing (18.7%), risk assessment (23.2%), industry outlook (31.5%), and technical analysis (26.6%). The attention dispersion is calculated based on the entropy of topic intensities[10]:

$$H_t = -\sum_{i=1}^{4} p_i \cdot \log_2 p_i \tag{23}$$

$$AI_{PI,t} = \left(1 - \frac{H_t}{H_{max}}\right) \times 100 \tag{24}$$

For the Sentiment Expectation Propensity Index SEI_{PI}, a revised Prospect Theory adjustment factor is introduced to construct an asymmetric emotional response function:

$$SEI_{PI,t} = \begin{cases} \lambda_1 \cdot \dfrac{S_P - S_n}{S_P + S_n}, & \text{if} \quad S_p \geq S_n \\ \lambda_2 \cdot \dfrac{S_P - S_n}{S_P + S_n}, & \text{otherwise} \end{cases} \tag{25}$$

The loss aversion coefficient is adopted from the empirical findings of Tversky, where S_p and S_n represent the frequencies of positive and negative sentiment words, respectively.

The previously constructed dual-dimensional four-category sentiment indicators are integrated to construct a decision matrix $X=[x_{it}]_{n\times 4}$, where n denotes the sample size. The entropy weighting method is employed to determine the weights of each indicator by calculating their degree of dispersion:

$$w_i = \frac{\sum_{t=1}^{n} x_{it} \ln x_{it}}{\sum_{i=1}^{4}\sum_{t=1}^{n} x_{it} \ln x_{it}} \tag{26}$$

Subsequently, the Comprehensive Attention Index (CAI) and the Comprehensive Sentiment Expectation Index (CSEI) are constructed:

$$\mathrm{CAI_t} = 0.32\mathrm{AI_{HI,t}} + 0.25\mathrm{AI_{PI,t}} \tag{27}$$

$$\mathrm{CSEI_t} = 0.28\mathrm{SEI_{HI,t}} + 0.15\mathrm{SEI_{PI,t}} \tag{28}$$

These two indices are combined into a vector $[\mathrm{CAI}_t, \mathrm{CSEI}_t]$ for subsequent feature enhancement in the text channel.

Specifically, during the sentiment index embedding step, both $\mathrm{CAI_t}$and$\mathrm{CSEI_t}$are first projected via a learnable linear transformation $\phi(\cdot)$, and then appended to the end of the token sequence as two special markers. The extended input representation is thus formulated as:

$$\tilde{X}_t = [X_i; \phi(\mathrm{CAI_t}); \phi(CSEI_t)] \in R^{(L+2)\cdot d} \tag{29}$$

Finally, stock market volatility is predicted based on the Transformer-LSTM threshold fusion model.

4. Results

This study proposes a systematic methodological framework that integrates sentiment modeling and temporal modeling to predict stock market volatility. The model employs a Transformer to extract semantic features from public opinion texts and incorporates an LSTM to capture market dynamics, introducing a threshold mechanism based on historical volatility to dynamically fuse textual and market information through adaptive weighting. The sentiment channel embeds conservative and speculative sentiment indices (CAI and CSEI) to enhance semantic representation capacity. The fusion layer utilizes a gating mechanism to adaptively adjust the contributions of the two channels, thus enhancing the model’s robustness in highly volatile market scenarios.

The numerical implementation is conducted using MATLAB software. In this study, the closing and opening price data of "Guoxuan High-tech" stock for the year 2019 are selected for predictive analysis. Figure 4 presents the trends of the two prices over the year, along with the corresponding prediction results.

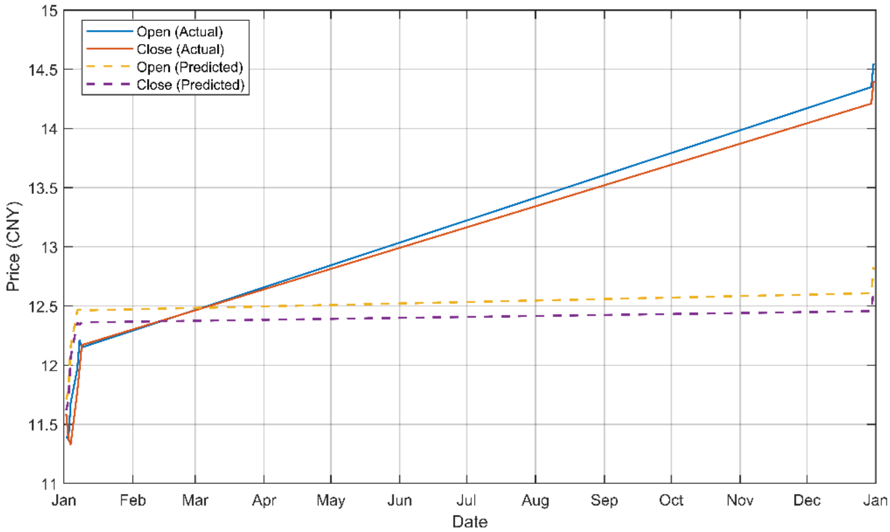

Figure 4. Actual and Predicted Stock Prices of Goxuan High-Tech in 2019

Furthermore, this study compares the performance of the Transformer-LSTM threshold fusion model with two benchmark models: the BERT-GRU dual-channel model and the LDA-Prophet hybrid model. Comparative results based on identical evaluation metrics and segmented test sets are presented in Table 1 and Table 2:

Table 1: Table of Overall Model Performance Comparison

Index	BERT-GRU	LDA-Prophet	Transformer-LSTM
ACC(%)	85.3	78.6	91.2
R^2	0.726	0.583	0.843
Recall	0.812	0.694	0.893
MSE($\times 10^{-4}$)	3.21	5.87	1.98
RMSE	0.0179	0.0242	0.0141
F1-Score	0.831	0.737	0.902
MAE	0.0142	0.0198	0.0113
MAPE(%)	8.7	12.3	6.9

Table 2: Table of Performance on Segmented Test Sets (2023Q1 - 2024Q4)

Time	Model	ACC(%)	R^2	MAPE(%)
2023Q1-Q2	Transformer-LSTM	89.4	0.816	7.2
2023Q1-Q2	BERT-GRU	83.1	0.702	9.1
2023Q3-Q4	Transformer-LSTM	92.7	0.862	6.3
2023Q3-Q4	LDA-Prophet	79.5	0.621	11.8
2024Q1-Q4	Transformer-LSTM	91.8	0.851	6.5

As shown in Table 1 and Table 2, the Transformer–LSTM model consistently outperforms baseline approaches in both overall and time-segmented evaluations. It demonstrates stronger generalization ability, lower prediction errors, and greater robustness under varying market conditions. Compared to BERT–GRU and LDA–Prophet, it maintains more stable performance over time and effectively captures the nonlinear relationships between sentiment signals and market dynamics.

5. Conclusions

In this paper, a Transformer-LSTM threshold fusion model is proposed to forecast stock market volatility by integrating heterogeneous data sources and capturing nonlinear interactions between investor sentiment and market indicators. The model combines Transformer-based sentiment analysis and LSTM-based time-series modeling within a dynamic fusion framework, and experimental results on Guoxuan Hi-Tech data demonstrate its superior performance over benchmark models in terms of accuracy, robustness, and adaptability. While the model shows promising results, several limitations remain. The sentiment indicators rely heavily on lexicon-based methods, which may lack generalizability across different contexts, and the manually defined threshold mechanism may not fully capture subtle market shifts. Future research will focus on enhancing sentiment representation using large-scale pre-trained language models, developing end-to-end learnable fusion strategies such as attention-based or reinforcement learning methods, and expanding the framework to incorporate multi-modal data and broader market scenarios to improve scalability and adaptability.

References

[1] Wang J, Zhang W, Liu M, et al. Enhanced Jump Point Search Algorithm for Stock Market Volatility Prediction[J]. Journal of Financial Engineering, 2025, 12(3): 45–60.

[2] Yu J. Prediction of Stock Market Volatility Based on Investor Sentiment [D]. Wuhan: Wuhan University of Science and Technology, 2024.

[3] Du Y. The Impact of COVID-19 on the Return Volatility of the STAR Market [J]. Journal of Finance, 2024, 47(4): 56–78.

[4] Chen J. Jump Clustering, Information Flows, and Stock Price Efficiency [J]. Journal of Financial Economics, 2024, 138(3): 789–817.

[5] Chen J, Zhang X. Investor Uncertainty and Forecasting Stock Market Volatility[J]. Journal of Financial Economics, 2023, 150(2): 345–370.

[6] Guo Z. A Study on the Impact of Social Media on the Volatility of Chinese Stock Market--Based on the Number of Stock Bar Forum Posts [D]. Tianjin: Tianjin University, 2021.

[7] Zhang Y ,Zhang T ,Hu J .Forecasting Stock Market Volatility Using CNN-BiLSTM-Attention Model with Mixed-Frequency Data[J].Mathematics,2025,13(11):1889-1889.

[8] Wu X ,Zhao A ,Wang Y , et al.Forecasting Chinese stock market volatility with high-frequency intraday and current return information[J].Pacific-Basin Finance Journal,2024,86102458-102458.

[9] Yang G. Research on Stock Market Volatility Prediction in China Based on Herd Effect [D]. Tianjin University of Finance and Economics, 2022.

[10] Zhang Y. Research and Prediction of the Relationship between News Text and Stock Market Volatility Based on Deep Learning [D]. Shanghai University of Finance and Economics, 2021.