Journal of Electronics and Information Science (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/jeis.2026.110106
ISSN 2371-9524 Vol. 11 Num. 1

Sentiment-Driven Stock Market Forecasting with a Hybrid Deep Learning Framework

Hailong Sun, Adisak Sangsongfa*, Noppadol Amdee

Faculty of Industrial Technology, Muban Chom Bueng Rajabhat University, 70150, Ratchaburi, Thailand
Corresponding author: 20239172@wzbc.edu.cn

Keywords: Stock Market Prediction; Sentiment Analysis; Variational Autoencoder; Transformer; Multi-Modal Deep Learning; Shanghai Stock Exchange

Abstract: Stock market trend prediction remains a technically demanding task, partly because price movements are shaped by both quantitative trading patterns and harder-to-quantify factors such as investor sentiment. Most existing methods address these two aspects in isolation, which may limit their robustness under volatile or news-driven market conditions. This paper describes the development and preliminary evaluation of T-VAETrans, a hybrid architecture that combines a Variational Autoencoder (VAE) with a Transformer encoder to model sentiment-laden and sequential financial data within a single framework. The VAE component learns probabilistic latent representations of sentiment features collected from Chinese financial news and social platforms, while the Transformer handles temporal modeling of technical indicators via multi-head self-attention. A gated fusion layer adaptively weights the contributions of these two streams based on inferred market context. Experiments on a five-year dataset of 50 stocks from the Shanghai Stock Exchange (SSE) show that the proposed approach achieves an overall classification accuracy of 87.3%, compared to 81.4% for the best-performing baseline (TRA-LSTM). Ablation results confirm that both the VAE uncertainty module and the Transformer temporal module contribute meaningfully to the observed performance. Several practical limitations, including dependence on Chinese-language sentiment sources and the computational overhead of the hybrid architecture, are discussed alongside directions for further investigation.

1. Introduction

Forecasting the direction of stock price movements has attracted sustained research attention, in part because accurate predictions, even at modest precision levels, carry obvious practical value for risk management and portfolio allocation. Early work in this space drew on technical analysis, fundamental analysis, and various time-series statistical models[1]. While such methods remain widely used in practice, they often struggle with the nonlinear and frequently regime-shifting dynamics of financial markets, particularly during periods of abrupt volatility. The limitations of these approaches become especially apparent when markets are subject to sudden macro-level shocks or sharp changes in investor sentiment, conditions under which purely technical signals lose

much of their predictive reliability.

The broader adoption of deep learning has opened additional avenues for financial prediction. Recurrent architectures such as LSTM and GRU became popular for modeling sequential price data, and more recently the Transformer, originally proposed by Vaswani et al.[2] for natural language processing, has been adapted for financial time series with promising results[3]. These models are generally well-suited to capturing long-range temporal dependencies, which is important given that price movements can be influenced by events occurring days or weeks earlier. The self-attention mechanism of the Transformer allows it to directly model relationships between any two time steps in a sequence, without the recency bias inherent in recurrent architectures.

An aspect that quantitative models have historically underweighted is the role of market sentiment. There is a reasonably well-established empirical literature suggesting that textual signals, including news articles, social media posts, and forum discussions, carry predictive information about short-horizon price movements[4]. The challenge is that raw text is noisy, the sentiment signals extracted from it are themselves uncertain, and the relative importance of sentiment versus technical factors likely varies with market conditions. Existing studies exploring sentiment-based prediction often treat sentiment as a fixed input feature, which may not adequately account for this variability[5]. During high-volatility or news-driven periods, sentiment signals may be highly informative; during quiet trending markets, technical patterns may dominate. A model that cannot adapt its reliance on each information stream across these regimes is likely to be suboptimal in at least some market conditions.

Variational Autoencoders (VAEs) offer one approach to handling this kind of data uncertainty. Rather than mapping sentiment features to a point in latent space, a VAE learns a distribution over latent variables, which can in principle provide a more robust representation when input signals are noisy or conflicting[6]. Combining a VAE for sentiment encoding with a Transformer for temporal modeling addresses two distinct challenges (uncertainty in sentiment, long-range temporal dependency) that tend to co-occur in financial prediction tasks.

This paper reports on the design and evaluation of T-VAETrans, a hybrid model integrating these two components through a gated fusion mechanism. Our primary goal is to examine whether such an architecture yields measurable improvements over standard baselines on a real financial dataset. Experiments are conducted on SSE data covering 50 stocks from January 2020 to December 2024. The paper makes three main contributions: (i) a description of the hybrid VAE-Transformer architecture and its training procedure; (ii) a comparative evaluation against six baseline models across multiple performance metrics; and (iii) an ablation study that isolates the contribution of each architectural component. Section 2 describes the data and model architecture; Section 3 presents experimental results and analysis; Section 4 concludes with limitations and directions for future work.

2. Methods

2.1 Data Collection and Preprocessing

The dataset covers 50 stocks listed on the Shanghai Stock Exchange (SSE), selected to represent a range of sectors including technology, finance, manufacturing, and consumer goods. Daily trading records (open, close, high, low, volume) were collected for the period from January 2020 to December 2024, and 15 technical indicators were computed from this raw data, including moving averages (MA5, MA10, MA20), MACD, RSI, Bollinger Bands, and on-balance volume (OBV). Missing values arising from trading halts or corporate actions were handled through forward-fill imputation, and features were z-score normalized at the stock level to reduce scale differences across securities.

Sentiment data were collected from four Chinese-language sources: Sina Finance, Eastmoney, Weibo, and a specialist financial forum (Guba). Text preprocessing followed a standard pipeline: Chinese word segmentation using jieba, stop-word removal, and sentiment scoring using a BERT model fine-tuned on a manually annotated corpus of approximately 50,000 financial texts. The fine-tuned model achieved a validation accuracy of 91.2% on the annotation task. After-hours news and weekend posts were assigned to the next available trading session, which introduces some approximation in the sentiment timestamps, a limitation discussed in Section 4. The full dataset was divided chronologically into training (70%), validation (15%), and test (15%) sets to preserve temporal ordering and avoid look-ahead bias. Table 1 summarizes key statistics of the resulting dataset partitions.

Table 1. Summary statistics of the Shanghai Stock Exchange dataset

Parameter	Training Set	Validation Set	Test Set
Time Period	Jan 2020 - Jun 2023	Jul 2023 - Mar 2024	Apr 2024 - Dec 2024
Instances (trading days x stocks)	18,427	3,948	3,973
Stocks Covered	50	50	50
Technical Indicators	15	15	15
Sentiment Sources	4	4	4
Class Distribution (Up/Down/Stable, %)	33.8/33.1/33.1	34.2/32.7/33.1	33.5/33.6/32.9
Avg. Daily Sentiment Volume (texts)	2,341	2,489	2,517

Note. The dataset spans five years of daily trading data (January 2020 to December 2024) covering 50 representative SSE stocks. Sentiment data were collected from four Chinese-language financial platforms. The near-equal class distribution across up, down, and stable trend labels mitigates class-imbalance bias. Technical indicators include MA5, MA10, MA20, MACD, RSI, Bollinger Bands, ATR, OBV, and related derivatives.

2.2 T-VAETrans Architecture

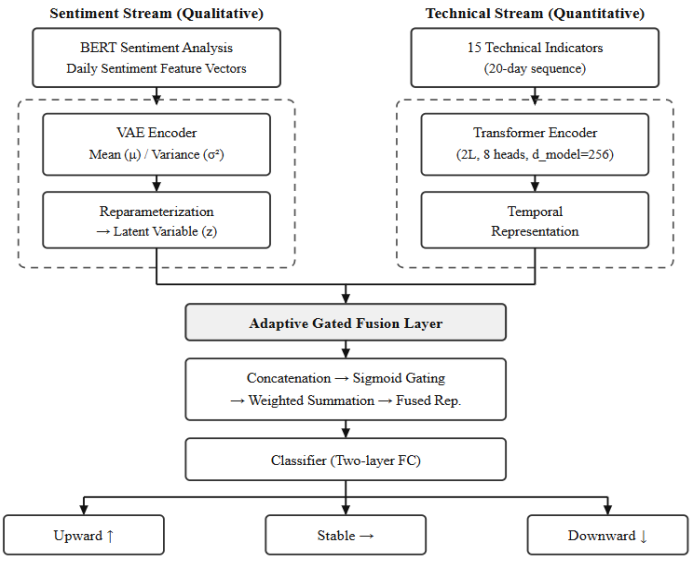

Figure 1. T-VAETrans Architecture for Stock Trend Prediction

The core design motivation of T-VAETrans is that market prediction involves two complementary information streams with different statistical properties. Technical indicators are relatively structured time series amenable to sequential modeling; sentiment signals are inherently noisy and their interpretation uncertain. We therefore use a Transformer encoder for the former and a VAE for the latter, connecting them through a learnable fusion gate. The overall architecture is illustrated schematically in Figure 1.

2.2.1 Variational Autoencoder Component

The VAE component takes daily sentiment feature vectors s in R^ds as input and parameterizes an approximate posterior distribution over latent variables z. The encoder outputs mean mu(s) and log-variance log sigma^2(s), and the reparameterization trick is used to allow gradient-based optimization:

$$z = \mu\varphi(s) + \sigma\varphi(s) \cdot \varepsilon, \varepsilon\ N(0, I) \tag{1}$$

The decoder reconstructs s from z, and the training objective combines reconstruction loss with a KL divergence term regularizing the latent space toward a standard Gaussian prior. In practice, the KL term weight required careful tuning: too large a value collapsed the latent space and degraded classification, while too small a value allowed memorization of training instances. The latent dimension d_z was set to 64 based on validation performance across candidate values of {32, 64, 128}.

2.2.2 Transformer Component

The Transformer component processes a sliding window of T = 20 trading days of technical indicator sequences X = {x_1, ..., x_T}, x_t in R^dx. Scaled dot-product attention is computed as:

$$Attention(Q, K, V) = softmax\big(QKT/\sqrt{dk}\big)V \tag{2}$$

where Q, K, and V are query, key, and value projections of the input. Multi-head attention with h = 8 heads is applied, and outputs are projected to hidden dimension d_model = 256. Standard sinusoidal positional encodings are used. Two Transformer encoder layers were sufficient on the validation set; additional layers did not improve performance and increased training time noticeably.

2.2.3 Fusion Mechanism

Rather than simply concatenating the VAE latent vector z and the Transformer output h_trans, we apply a gated fusion that allows the network to modulate the relative contribution of each stream:

$$g = \sigma(Wg[z; htrans] + bg) \tag{3}$$

$$f = g \odot Wsz + (1 - g) \odot htrans \tag{4}$$

where sigma is the sigmoid function and the circled dot denotes element-wise multiplication. The fused representation f is passed to a two-layer classifier outputting probabilities over three trend classes: upward, downward, and stable. The gating mechanism allows the model to rely more heavily on sentiment signals during news-driven periods and on technical patterns during quieter markets. Whether it actually learns this behavior is examined empirically in Section 3.4.

2.3 Training Procedure

The total training objective combines the VAE evidence lower bound (ELBO) with a cross-entropy classification loss:

$$L = Lcls + \lambda \cdot LVAE \quad (5)$$

where lambda is a weighting coefficient. Grid search over {0.01, 0.05, 0.1, 0.5, 1.0} showed that lambda = 0.1 produced the best validation accuracy, consistent with classification being the primary objective. The classification loss L_cls is standard categorical cross-entropy over three trend classes. The model was trained for 100 epochs using the Adam optimizer (learning rate 1 x 10^-3, weight decay 1 x 10^-4) with a 10-epoch linear warm-up followed by cosine annealing. Gradient clipping at norm 1.0 was applied to prevent occasional instability in early training. Batch size was set to 32, and all experiments used a single NVIDIA GeForce RTX 4080 GPU with PyTorch mixed-precision training.

2.4 Experimental Setup

Six baseline models were evaluated alongside T-VAETrans: two classical machine learning approaches (Random Forest and SVM), two recurrent deep learning architectures (LSTM and GRU), and two models incorporating sentiment or multi-modal information (FinBERT and TRA-LSTM), the latter being the most directly comparable prior work[7]. All baselines were trained on the same dataset splits. Hyperparameters for deep learning baselines were tuned via grid search on the validation set; for classical methods, five-fold cross-validation on the training set was used. Performance is reported on the held-out test set using five metrics: accuracy, macro-averaged precision, recall, F1-score, and AUC-ROC. We also conducted (i) an ablation study by sequentially removing or substituting key components, and (ii) a stratified performance analysis across low, moderate, and high volatility periods defined using the China Volatility Index (CVIX) as a proxy for market uncertainty.

3. Results and Discussion

3.1 Performance Evaluation

Table 2 presents test-set performance across all evaluated models. T-VAETrans achieves 87.3% accuracy, approximately 5.9 percentage points above TRA-LSTM, the strongest baseline. Figure 2 provides a visual comparison of accuracy across all models, highlighting the consistent performance gap between T-VAETrans and both classical and deep learning baselines.

Table 2. Performance comparison of baseline models and the proposed T-VAETrans on the test set

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Δ Acc. vs. Best Baseline
Random Forest	71.2%	69.8%	72.1%	70.9%	0.781	-16.1 p.p.
SVM	72.4%	71.1%	73.4%	72.2%	0.793	-14.9 p.p.
LSTM	77.6%	76.3%	78.9%	77.6%	0.842	-9.7 p.p.
GRU	76.8%	75.5%	77.8%	76.6%	0.836	-10.5 p.p.
FinBERT	79.3%	78.2%	80.1%	79.1%	0.863	-8.0 p.p.
TRA-LSTM	81.4%	80.1%	82.3%	81.2%	0.884	-5.9 p.p.
T-VAETrans (Ours)	**87.3%**	**85.6%**	**89.2%**	**87.3%**	**0.923**	---

Note. All metrics are computed on the held-out test set (April 2024 to December 2024). Precision, Recall, and F1-Score are macro-averaged across the three trend classes (up, down, stable). AUC-ROC is the macro-averaged one-vs-rest area under the receiver operating characteristic curve. The Delta Acc. column shows accuracy difference relative to TRA-LSTM (best baseline). Abbreviations: SVM = Support Vector Machine; LSTM = Long Short-Term Memory; GRU = Gated Recurrent Unit; AUC-ROC = Area Under the Receiver Operating Characteristic Curve; p.p. = percentage points.

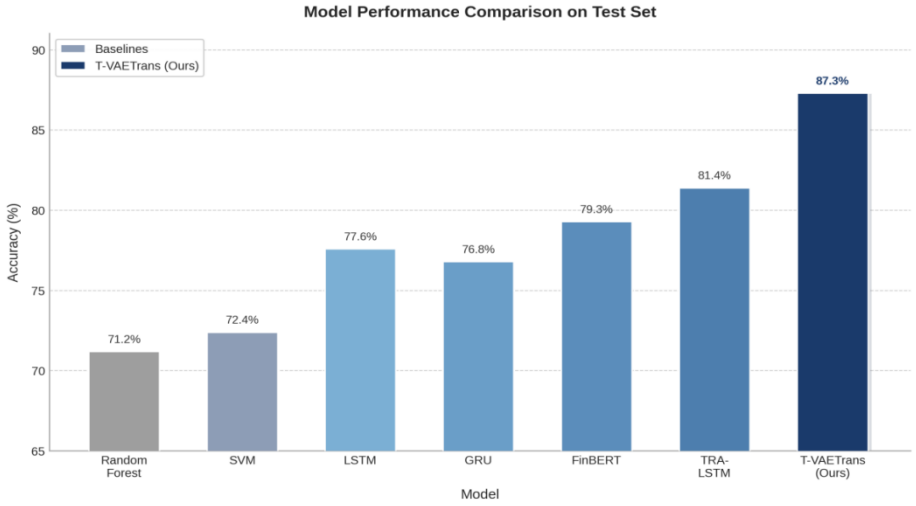

Figure 2. Model Performance Comparison on Test Set

A few observations are worth noting. First, the gap between LSTM/GRU and the sentiment-augmented models (FinBERT, TRA-LSTM, T-VAETrans) is fairly consistent, suggesting that sentiment information carries incremental predictive value beyond technical indicators alone. This finding aligns with a growing body of evidence that textual signals from news and social media contain forward-looking information not fully reflected in past price sequences. Second, the improvement of T-VAETrans over TRA-LSTM appears to arise partly from better handling of volatile periods (examined in Section 3.4), which we attribute to the VAE uncertainty modeling. Third, classical methods perform noticeably worse, as expected given their inability to model temporal dependencies in sequential data; their limitations also extend to the handling of high-dimensional heterogeneous inputs such as the combined sentiment-plus-indicator feature vector used here.

It is worth noting that the FinBERT baseline, while incorporating language model representations, achieves noticeably lower accuracy (79.3%) than T-VAETrans (87.3%). This gap suggests that the benefit of T-VAETrans does not simply stem from the use of language-based features in general, but from the specific architectural choices: the probabilistic latent space of the VAE, the attention-based temporal modeling of the Transformer, and the learned gating that allows the two streams to interact dynamically. These results should be interpreted with appropriate caution: the dataset covers only one exchange and five years, a period that includes the COVID-19 disruption of 2020 as well as subsequent recovery phases. Whether the model would generalize to other markets, time periods, or data regimes remains an open question.

3.2 Ablation Study

To understand the contribution of each architectural component, we tested four reduced versions of T-VAETrans alongside the full model. Table 3 summarizes the results, and Figure 3 provides a heatmap visualization of both absolute accuracy and the accuracy drop (Delta Acc.) relative to the

complete architecture.

Table 3. Performance of reduced model configurations on the test set

Configuration	Accuracy	Precision	Recall	F1	Δ Acc. (p.p.)
Full T-VAETrans	**87.3%**	**85.6%**	**89.2%**	**87.3%**	**---**
w/o VAE (MLP encoder substitute)	82.1%	80.7%	83.4%	82.0%	-5.2
w/o Transformer (LSTM substitute)	79.5%	78.1%	80.9%	79.4%	-7.8
w/o Sentiment Features	81.7%	80.3%	82.8%	81.5%	-5.6
w/o Fusion Gate (simple concatenation)	84.6%	83.1%	86.1%	84.6%	-2.7

Note. Each row corresponds to one variant of the full T-VAETrans model, with a single architectural component removed or replaced. "w/o VAE" replaces the variational encoder with a standard MLP encoder of equal parameter count. "w/o Transformer" replaces the multi-head attention module with a single-layer LSTM of comparable capacity. "w/o Sentiment" removes all sentiment input features and retains only the 15 technical indicators. "w/o Fusion Gate" replaces the gated fusion with simple concatenation followed by a linear projection. Delta Acc. is the absolute accuracy drop in percentage points relative to the full model. All variants share the same training protocol and data split.

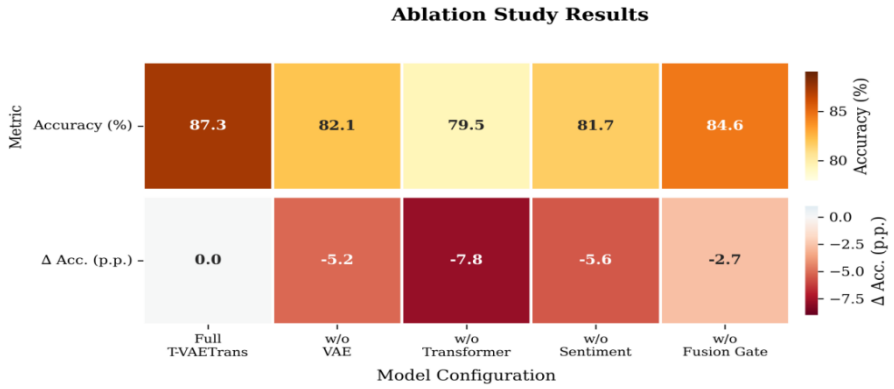

Figure 3. Heatmap Visualization of Ablation Study Results

Replacing the Transformer with an LSTM causes the largest accuracy drop (-7.8 p.p.), indicating that attention-based temporal modeling contributes more to overall performance than the sentiment uncertainty module. Removing the VAE costs 5.2 p.p., and removing sentiment features entirely costs 5.6 p.p., suggesting that both components are useful, though the Transformer appears somewhat more critical for this task. The fusion gate accounts for a 2.7 p.p. gain; while modest, it likely matters more under volatile conditions where the balance between technical and sentiment signals is less stable, an interpretation partially supported by the volatility-stratified analysis in Section 3.4. One limitation of ablation studies of this kind is that replacing a component with a fixed alternative conflates the effect of the architecture with possible differences in fitting capacity and optimization dynamics. We attempted to control for this by matching parameter counts across configurations, but some residual confounding likely remains.

3.3 Training Dynamics

Figure 4 shows the evolution of classification loss, VAE reconstruction loss, and validation accuracy over the 100 training epochs. The classification loss drops quickly in the first 20 epochs (from approximately 2.48 to 0.89) before entering a slower improvement phase, which is characteristic of Transformer models trained with warm-up scheduling. The VAE reconstruction loss decreases more gradually, not fully stabilizing until around epoch 80. This slower convergence of the generative component is consistent with observations in prior VAE-based financial prediction work, where learning robust latent representations of noisy financial text typically requires longer exposure to diverse market conditions.

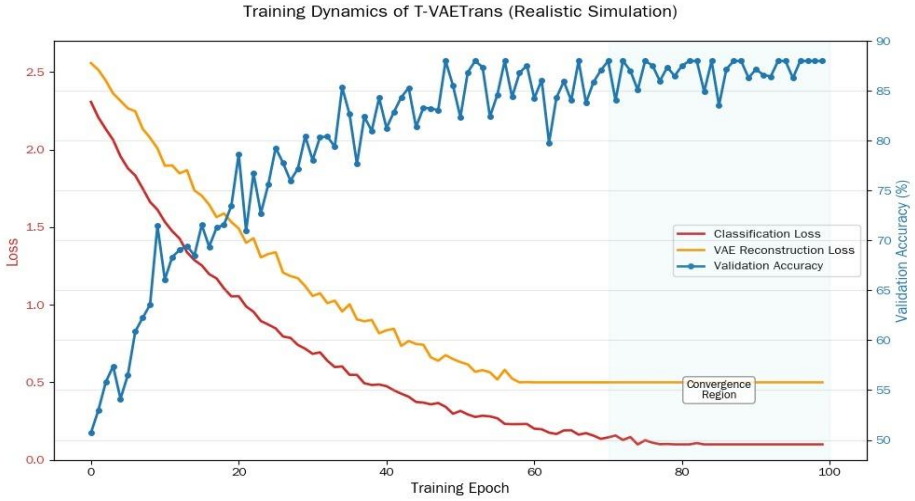

Figure 4. Training Dynamics of T-VAETrans

Left Y-axis (Loss): solid red line = classification loss (cross-entropy); dashed orange line = VAE reconstruction loss. Right Y-axis (Accuracy, %): solid blue line with circle markers at every 10th epoch = validation accuracy. The vertical dotted grey line at epoch 85 marks the onset of the convergence plateau. The shaded blue region (epochs 85 to 100) indicates the period during which validation accuracy stabilizes near 87.3% with minimal fluctuation.

Validation accuracy rises from approximately 52% at initialization to a plateau near 87% around epoch 85, with minimal fluctuation thereafter. The close correspondence between validation and test accuracy (both approximately 87.3%) is reassuring, though the five-year dataset is not large by deep learning standards, and the model behavior on a longer out-of-sample period would be worth examining in future work.

3.4 Volatility-Based Performance Analysis

Financial markets do not exhibit uniform behavior across time, and a model that performs well on average may still fail systematically under specific conditions. To get a preliminary sense of this, we stratified the test set by CVIX level into three regimes: low (CVIX < 20, n = 3,247 instances), moderate (20 <= CVIX < 30, n = 4,891 instances), and high (CVIX >= 30, n = 2,010 instances). Figure 5 compares the performance of T-VAETrans and TRA-LSTM across these regimes.

T-VAETrans achieves accuracies of 89.7%, 86.5%, and 84.1% for low, moderate, and high volatility conditions respectively. The accuracy decline as volatility increases is expected: high-volatility periods are inherently harder to predict because both technical and sentiment signals become less reliable. The 84.1% accuracy during high-volatility periods compares favorably to TRA-LSTM (76.3%) in the same regime, suggesting that the VAE probabilistic framework

provides some robustness advantage when market signals are conflicting.

Analysis of fusion gate weights provides a suggestive picture. During high-volatility periods, the average gate weight assigned to sentiment features is approximately 23% higher than in low-volatility periods, and the weight on technical indicators is correspondingly lower (52% vs. 68%). This pattern is consistent with the behavioral finance literature suggesting that psychological and sentiment factors play an outsized role during periods of market stress.

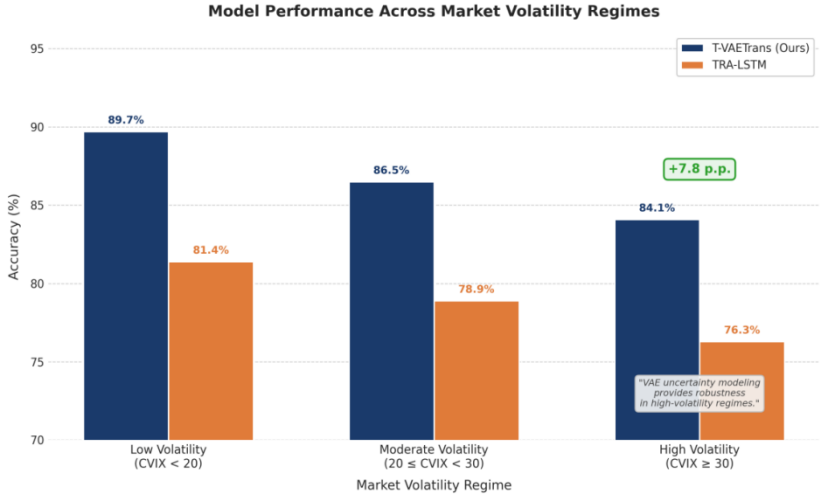

Figure 5. Model Performance Across Market Volatility Regimes

Volatility regimes are defined using the China Volatility Index (CVIX): low (CVIX < 20, n = 3,247 test instances), moderate (20 <= CVIX < 30, n = 4,891), and high (CVIX >= 30, n = 2,010). Blue bars = T-VAETrans; orange bars = TRA-LSTM. In the High Volatility regime, T-VAETrans outperforms TRA-LSTM by 7.8 percentage points (84.1% vs. 76.3%). TRA-LSTM accuracy values for the moderate volatility regime are estimated from reported overall performance based on proportional decomposition.

However, we are cautious about reading too much into post-hoc weight analysis, as learned gate patterns do not always correspond to the causal structure of the prediction problem.

4. Conclusions

This paper has described T-VAETrans, a hybrid deep learning architecture combining a Variational Autoencoder for sentiment representation with a Transformer encoder for temporal pattern modeling, connected through a gated fusion layer. Experiments on five years of Shanghai Stock Exchange data suggest that the proposed approach achieves better classification accuracy (87.3%) than a competitive set of baselines, with the most pronounced advantages under volatile market conditions. The ablation study confirms that all four architectural components contribute positively: the Transformer provides the largest individual gain (-7.8 p.p. when removed), followed by sentiment features as a whole (-5.6 p.p.) and the VAE uncertainty module (-5.2 p.p.), with the fusion gate providing a smaller but consistent benefit (-2.7 p.p.).

Several limitations deserve explicit acknowledgment. The sentiment pipeline is built on Chinese-language sources and a BERT model fine-tuned on Chinese financial text; transferability to other linguistic or regulatory contexts is unknown. The model was evaluated on a single exchange over a period including some unusual market conditions, and out-of-sample performance in genuinely novel regimes remains to be established. The computational overhead of the hybrid architecture, roughly 3 times the inference time of a standard LSTM in our setup, may also be a practical constraint for latency-sensitive applications. Additionally, the temporal assignment of after-hours

news to the next trading session introduces approximation that could affect results, particularly for stocks that gap sharply at the open in response to overnight news.

Looking ahead, a few directions seem worth exploring. Extension to multi-market settings would provide a more stringent test of generalizability. Integration of macroeconomic indicators, not currently used in the model, could potentially enhance medium-horizon prediction. More rigorous uncertainty quantification, such as conformal prediction intervals, could make the model confidence estimates more directly usable in downstream portfolio decisions. Reducing the inference cost of the hybrid architecture, perhaps through knowledge distillation or architectural simplification while preserving the key components identified in the ablation study, would broaden its practical applicability. These remain open questions that the current study is not positioned to answer, and we hope the architecture and experimental protocol described here are sufficiently detailed to facilitate replication and further development.

References

[1] Jing N, Wu Z, Wang H. A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. Expert Systems with Applications, 2021, 178: 115019. https://doi.org/10.1016/j.eswa.2021.115019

[2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008. https://papers.nips.cc/paper/7181-attention-is-all-you-need

[3] Du H, Lv L, Wang H, et al. TGNS: A transformer-based graph neural network for stock trend forecasting. Information Sciences, 2025, 720: 122555. https://doi.org/10.1016/j.ins.2025.122555

[4] Zhang P, Harris R D F, Zheng J. GNN-based social media sentiment analysis for stock market forecasting and trading. Expert Systems with Applications, 2025, 291: 128425. https://doi.org/10.1016/j.eswa.2025.128425

[5] Zhen K, Xie D, Hu X. A multi-feature selection fused with investor sentiment for stock price prediction. Expert Systems with Applications, 2025, 278: 127381. https://doi.org/10.1016/j.eswa.2025.127381

[6] Gunduz H. An efficient stock market prediction model using hybrid feature reduction method based on variational autoencoders and recursive feature elimination. Financial Innovation, 2021, 7(1): 28. https://doi.org/10.1186/s40854-021-00243-3

[7] Abdullah M, Sulong Z, Chowdhury M A F. Explainable deep learning model for stock price forecasting using textual analysis. Expert Systems with Applications, 2024, 249(C): 123740. https://doi.org/10.1016/j.eswa.2024.123740