The Fusion Research of Financial News Sentiment Quantification and Gold Price Time-Series Modeling

Xiawei Wang[1,a,*], Shengfeng Guo[1,b], Zhangting Wei[1,c]

[1]Wuhan University of Science and Technology, Wuhan, China

[a]15072286700@163.com, [b]sunwxn@wust.edu.cn, [c]m13581989902@163.com

*Corresponding author

Keywords: Gold Price Forecasting, News Sentiment Quantification, BERT, GRU, Multi-Modal Fusion

Abstract: Gold price forecasting is a critical topic in financial quantitative research, with its volatility driven by both traditional supply-demand dynamics and market sentiment. Addressing the issue of insufficient accuracy in existing forecasting models due to the neglect of sentiment factors, this study innovatively proposes a multi-modal analysis framework integrating BERT sentiment quantification and GRU time-series modeling. The framework enhances prediction performance through the synergistic modeling of unstructured text (financial news) and structured data (gold prices). The research adopts a "Fine-tuning-Quantification-Fusion" framework. Firstly, 8115 manually labeled (Positive: 1; Neutral: 0; Negative: -1) gold-related news headlines are used as a corpus to fine-tune pre-trained large language models. Comparative analysis reveals BERT achieves optimal performance (accuracy: 86%). Secondly, the fine-tuned BERT model classifies sentiment for the remaining unlabeled headlines. The label value serves as both the classification output and sentiment score. The daily sentiment index is quantified by averaging the labels of multiple news items per day. Finally, this sentiment index is incorporated into gold price time-series forecasting. Comparing models with and without the sentiment index demonstrates that the GRU model incorporating the sentiment index delivers superior prediction performance (test set MAE=3.96, R^2=0.99). Diebold-Mariano (DM) tests further validate that models with the sentiment index significantly outperform baseline models ($P<0.01$).

1. Introduction

Gold price forecasting is a central topic in financial research, with its volatility driven by both supply-demand relationships and market sentiment. In recent years, events such as geopolitical conflicts and policy adjustments have frequently triggered sharp fluctuations in gold prices. Traditional forecasting models, however, often fail to capture such irrational changes due to their neglect of sentiment factors. Investor behavior is significantly influenced by news sentiment; for instance, sudden financial news often causes abrupt shifts in market sentiment, leading to short-term sharp volatility in gold prices. Therefore, integrating structured data with unstructured text sentiment to construct a multi-modal analysis framework has become a key breakthrough direction for improving prediction accuracy.

Existing research has made significant progress in areas such as the linkage mechanism between sentiment and financial markets, methods for measuring text sentiment, and time-series forecasting models. Firstly, the predictive value of sentiment for financial markets has been widely validated [1, 2]. Gold prices are also significantly affected by investor sentiment [3, 4]. Secondly, text sentiment quantification methods have evolved: traditional dictionary methods rely on static lexicons and suffer from drawbacks such as poor transferability and inability to recognize negation contexts [5,6].In contrast, deep learning methods based on large language models like BERT significantly improve the accuracy of financial text classification through dynamic semantic capture [7,8]. Finally, time-series models (e.g., GRU, LSTM) have demonstrated excellent performance in gold price forecasting [9, 10]. However, most studies have not systematically integrated sentiment

  DOI: 10.25236/iiicec.2025.018

variables with multi-source heterogeneous data. Overall, existing methods still exhibit significant limitations in sentiment measurement granularity, multi-modal data fusion, and time-series-sentiment collaborative modeling.

Addressing the above issues, this study systematically integrates unstructured text and structured time-series data according to a Fine-tuning-Quantification-Fusion framework, as shown in Figure 1. Specifically, based on 8115 manually annotated news headlines, the ERNIE/BERT models are fine-tuned to generate a daily sentiment index. This index is then incorporated as an additional input feature into LSTM/GRU time-series models. Empirical results show that the fine-tuned BERT model achieves the best classification performance; the GRU model incorporating the sentiment index exhibits the best prediction performance; and DM tests confirm that models incorporating the sentiment index significantly outperform those without it. The innovations of this study are reflected in three aspects: (1) Theoretical Innovation: Establishing an event-driven gold price prediction model, elucidating the linkage mechanism between financial news sentiment indices and gold price fluctuations to reveal the impact pathway of market sentiment on investment decisions; (2) Methodological Innovation: Proposing the "manual annotation - large model fine-tuning - dynamic classification" sentiment quantification paradigm; (3) Application Innovation: Constructing a hybrid prediction model based on multi-modal data fusion, integrating time-series price data and textual sentiment features to build a more comprehensive forecasting system.

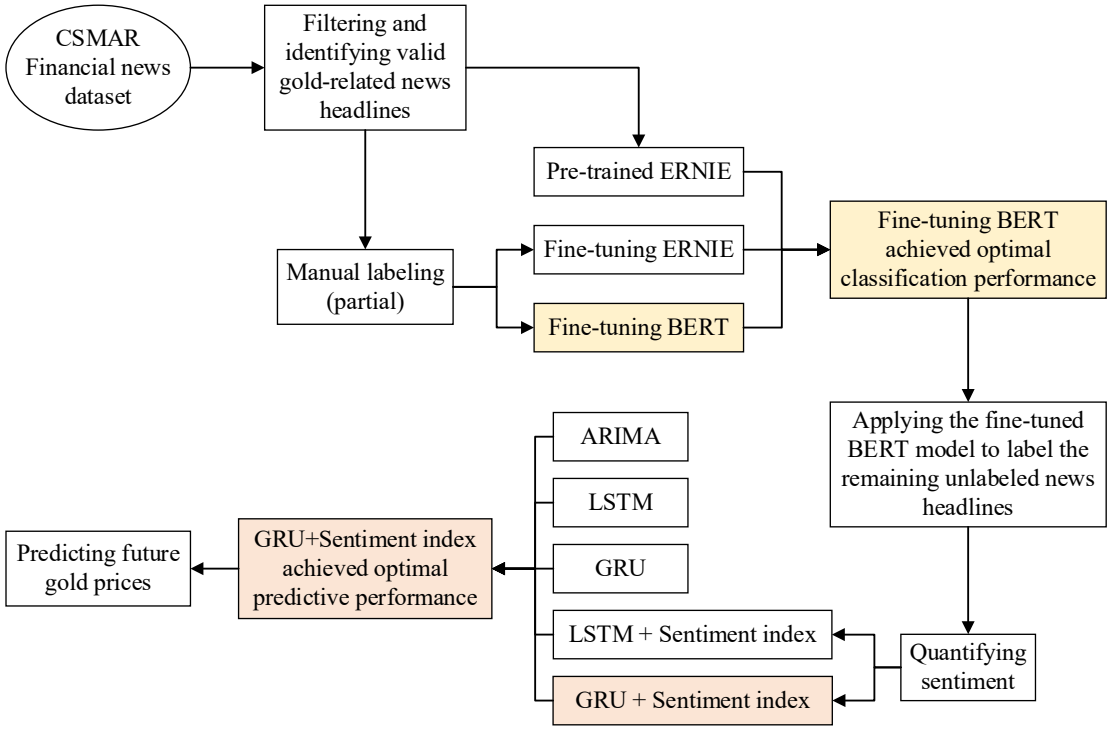

Figure 1 Research Framework Diagram.

2. Data Sources and Processing

2.1. News Text Data

Financial news headlines from January 1, 2015, to July 7, 2024, were obtained through the CSMAR WinGo Financial Text Data Platform, collecting over 5.9 million raw data entries.

The collected raw data were further filtered to include only news headlines containing keywords such as "gold", "gold price". Duplicate news and non-financial related reports (e.g., "golden week for tourism", "golden period of development" used metaphorically) were removed from the filtered headlines.

Ultimately, over 30,000 valid news headlines were retained, averaging 10 news items per day, covering all trading days of the London Bullion Market Association (LBMA), London Metal Exchange (LME), and Shanghai Gold Exchange (SGE). Sample data of some financial news

headlines are shown in Table 1.

Table 1 Examples of gold-related news headlines.

Date	News headline	Source
2023/1/3	US Dollar Index Remains Weak; Gold Price Edges Higher Within Narrow Range	Everbright Futures
2023/1/3	Spot Gold Hits New High Since Late June at $1,846.15 per Ounce	Cailian Press
2023/5/3	Fed Rate Decision Imminent! Hong Kong's Three Major Indices Under Pressure; Gold Stocks Rally Against Market Trend	Securities Times e Company
2023/12/26	Behind Gold's Record Highs in 2023: "Buy Gold in Times of Chaos" Rule Proves True Once Again	21st Century Business Herald

2.2. Gold Price Data

Gold price data from December 29, 1978, to March 14, 2025, was downloaded from the World Gold Council website. The original data represented prices in various currencies per troy ounce. To avoid interference, the raw data was processed: weight units were converted strictly according to World Gold Council conversion rules, resulting in gold price data in "Chinese Yuan per Gram" for the period January 1, 2015, to December 31, 2024. Sample gold price data is shown in Table 2.

Table 2 Example of gold price dataset.

Date	Gold price (CNY/g)
2024/12/5	616.15
2024/12/6	616.56
2024/12/9	623.54
2024/12/10	627.04
2024/12/11	631.73
2024/12/12	627.37

It is important to note that, as evident from the sample data, gold prices exhibit date discontinuity, with missing dates corresponding to weekends when exchanges are closed. Therefore, in subsequent processing, it is assumed by default that weekend data does not exist, ensuring date continuity.

3. Model Establishment

3.1. Large Language Model Sentiment Classification

This study employs pre-trained Transformer-based language models for financial text sentiment classification, primarily comparing the fine-tuning effects of the ERNIE and BERT models.

BERT (Bidirectional Encoder Representations from Transformers) is a bidirectional pre-trained language model based on the Transformer architecture. Its core innovations are Masked Language Modeling (MLM), where 15% of input tokens are randomly masked and predicted based on context, and Next Sentence Prediction (NSP), determining if two sentences are consecutive. The study uses the BERT-base-Chinese version, with input representation:

$$E = [CLS] + \sum_{i=1}^{n} (Token_i + Position_i + Segment_i) \tag{1}$$

Where [*CLS*] is the output token used for classification tasks.

For comparison, the ERNIE (Enhanced Representation through kNowledge IntEgration) model is also used. Its characteristic is enhancing semantic representation through knowledge masking strategies, such as entity-level masking (e.g., masking entire entities like "Federal Reserve") and phrase-level masking (e.g., identifying domain-specific phrases like "interest rate hike expectations" as masking units). The study employs the ERNIE 3.0-mini-zh version, primarily considering its

hierarchical masking strategy's suitability for understanding professional terminology in financial texts.

Supervised Fine-tuning refers to the technical paradigm of adjusting the parameters of a pre-trained language model (BERT, ERNIE) using labeled datasets to adapt it to specific downstream tasks. Unlike unsupervised learning relying on intrinsic data distributions, this method optimizes model representations through explicit supervision signals (e.g., classification labels, sequence labeling), representing a typical discriminative learning approach.

In this study, the explicit supervision signal is a manually labeled dataset of gold-related news headlines. The effective labeled dataset comprises 8115 entries, spanning the period January 1, 2023, to July 7, 2024. Examples are shown in Table 3, where the "Label" column: 1 represents Positive; -1 represents Negative; 0 represents Neutral. ERNIE and BERT are separately supervised trained using this labeled dataset, enabling the models to learn the mapping relationship between "news headlines" and "labels," thus adapting them to the specific task.

Table 3 Example of manually labeled dataset.

Date	News headline	Label
2023/1/4	Treasury Yields Fall, Pushing Up New York Gold Prices on Jan 3rd	1
2023/1/17	Dollar Index Edges Higher, International Gold Price Fluctuated and Closed Lower	-1
2023/1/18	Shanghai Gold Exchange Issues Notice on Continuing Efforts to Control Market Risks	0

It is important to note that manual labeling inevitably introduces subjective bias. To minimize labeling errors and bias, the team conducted cognitive bias testing and periodic calibration training for annotators before labeling to unify standards. Additionally, multiple independent annotations and cross-validation were performed during the labeling process.

Considering the computational power and time required to train large language models, news headlines from January 1, 2024, to July 7, 2024 (total 4880 items) were first used as the training corpus. The dataset was split into training and test sets (ratio 8:2), followed by model training.

3.2. Time-Series Model Prediction

The Gated Recurrent Unit (GRU) is a variant of Recurrent Neural Networks (RNNs) for sequence modeling, mitigating the vanishing gradient problem through gating mechanisms. Its core components are the update gate and reset gate, controlling the flow of historical information and the integration of current input. Let the input sequence be $\{x_t\}$ and the hidden state be h_t, the GRU computation is as follows:

$$\begin{cases} z_t = \sigma(W_z[h_{t-1}, x_t]), & (update\ gate) \\ r_t = \sigma(W_r[h_{t-1}, x_t]), & (reset\ gate) \\ \tilde{h}_t = \tanh(W_h[r_t \odot h_{t-1}, x_t]), & (candidate\ state) \\ h_t = (1-z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t, & (final\ state) \end{cases} \tag{2}$$

Where σ is the Sigmoid function, $\odot$ denotes element-wise multiplication, and W_* are trainable parameters. By simplifying the LSTM structure (merging the memory cell and hidden state), GRU reduces the number of parameters while maintaining the ability to model long-term dependencies, making it suitable for efficient time-series data modelling.

This study validates the impact of incorporating sentiment variables on time-series model predictions by integrating sentiment variables with time-series models. To this end, Long Short-Term Memory (LSTM) and GRU models are compared, evaluating their prediction performance both with and without the sentiment index. When inputting the sentiment index, strict alignment with gold price dates is essential to avoid misalignment. The dataset was split into training and test sets (ratio 8:2), followed by model training.

4. Indicator Selection

4.1. Model Input Indicators

The model input indicator system comprises two categories: basic time-series indicators and textual sentiment indicators. The basic time-series indicator is the gold price time-series data within the selected time range, denoted as P_t. The textual sentiment indicator is the news sentiment quantification score generated by the large language model, denoted as S_t, expressed mathematically as:

$$S_t = \frac{1}{N}\sum_{i=1}^{N} SentimentSocre(H_i) \tag{3}$$

Where N represents the total number of financial news headlines on a specific date, H_i represents the i-th news headline on that day, and $SentimentScore(H_i)$ represents the mapping rule for sentiment scores (as shown in Table 4). The mean of the quantified sentiment indices of all financial news headlines on a given day is taken as the daily sentiment index.

Table 4 Sentiment score mapping rules.

Label	Quantified index	News headline
Positive	1	Surge in Gold Demand Drives Price Increase
Neutral	0	Federal Reserve Keeps Interest Rates Unchanged
Negative	-1	Easing Geopolitical Tensions Reduce Gold's Safe-Haven Demand

The model input indicators are summarized, with example data shown in Table 5.

Table 5 Model input example.

Date	Gold price (CNY/g)	Daily sentiment index
2024-06-27	543.03	0.05
2024-06-28	544.62	0.56
2024-07-01	544.28	0.14
2024-07-02	545.12	-0.13
2024-07-03	551.96	0.27
2024-07-04	551.17	0.81
2024-07-05	555.94	0.78

4.2. Model Evaluation Indicators

4.2.1. Classification Performance Metrics

The performance of this classification model is primarily evaluated using four metrics: accuracy, precision, recall, and F_1–score. These metrics are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F_1 - score = \frac{2\,precision \cdot recall}{precision + recall} \tag{7}$$

Where: TP (True Positive) denotes the number of samples that are actual positive samples and are classified as positive. FP (False Positive) denotes the number of samples that are actual negative

samples but are classified as positive.TN (True Negative) denotes the number of samples that are actual negative samples and are classified as negative. FN (False Negative) denotes the number of samples that are actual positive samples but are classified as negative.

4.2.2. Predictive Performance Metrics

To evaluate the quality of predictive performance, we primarily select Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). Additionally, the Diebold-Mariano (DM) test is employed to determine whether the difference in predictive accuracy between two models—one incorporating daily sentiment index and one without—is statistically significant. The calculation formulas are as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{8}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\% \tag{9}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{10}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{11}$$

Where y_i is the actual value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the mean of the actual values, and n is the sample size.

The Diebold-Mariano (DM) test is used to assess whether the difference in predictive accuracy between two forecasting models is statistically significant. The null hypothesis states that there is no significant difference in the forecast accuracy between the two models. The DM test statistic is given by:

$$DM = \frac{\bar{d}}{\sqrt{\frac{V(d_t)}{T}}} \tag{12}$$

Where $\bar{d}$ is the mean loss differential, $V(d_t)$ is its estimated variance, and T is the sample size.

Given a significance level $\alpha = 0.05$, if the p-value associated with the DM statistic is less than α, the null hypothesis is rejected. This indicates a statistically significant difference in predictive accuracy between the two models, implying that one model is significantly superior to the other in terms of forecasting performance.

5. Model Solution Results

5.1. Classification Model

Python was used to implement sentiment classification of news headlines using various models, outputting classification evaluation metrics. The performance comparison of all classification models is summarized in Table 6.

According to the results in Table 6, the fine-tuned BERT model achieves an accuracy of 87% on the test set, significantly outperforming the fine-tuned ERNIE (83%) and the pre-trained ERNIE model (48%). Additionally, examining the recall rates for each label shows that fine-tuned ERNIE has the highest recall for the Positive label (94%), while fine-tuned BERT has the highest recall for

the Neutral and Negative labels (67% and 89%, respectively). This indicates that fine-tuned BERT exhibits the best overall classification performance, while fine-tuned ERNIE is more adept at identifying positive sentiment. Considering other metrics comprehensively, it is concluded that in the current news text classification task, fine-tuned BERT may be better at capturing fine-grained semantic features in financial news texts.

Table 6 Performance comparison of classification models.

Model	Accuracy	Positive Recall	Neutral Recall	Negative Recall
Pre-trained ERNIE	0.48	0.46	0.15	0.83
Fine-tuned ERNIE	0.83	0.94	0.52	0.83
Fine-tuned BERT	0.87	0.93	0.67	0.89

Therefore, based on the above results, the superior-performing BERT model is selected for subsequent sentiment classification tasks. Specifically, the BERT model is first fine-tuned using the entire dataset of 8115 manually labeled data points. Subsequently, the fine-tuned BERT model is applied to the remaining unlabeled news data to classify the sentiment of news headlines. Finally, the daily sentiment index is calculated based on Formula (3). The classification performance evaluation using the full labeled dataset is shown in Table 7, with high overall accuracy (Accuracy=0.86).

Table 7 Classification report for fine-tuned BERT with multiple sample sizes.

	Precision	Recall	F1-score	Support
Negative	0.91	0.8	0.85	327
Neutral	0.75	0.68	0.71	357
Positive	0.88	0.95	0.91	939
Accuracy			0.86	1623
Macro avg	0.85	0.81	0.82	1623
Weighted avg	0.86	0.86	0.86	1623

To visualize the correlation between the daily sentiment index and gold prices, the time series of gold prices and corresponding daily sentiment index from January 1, 2024, to July 7, 2024, are plotted in Figure 2. The results show that fluctuations in the sentiment index generally correlate with increases or decreases in gold prices, with negative sentiment index having a more pronounced impact.

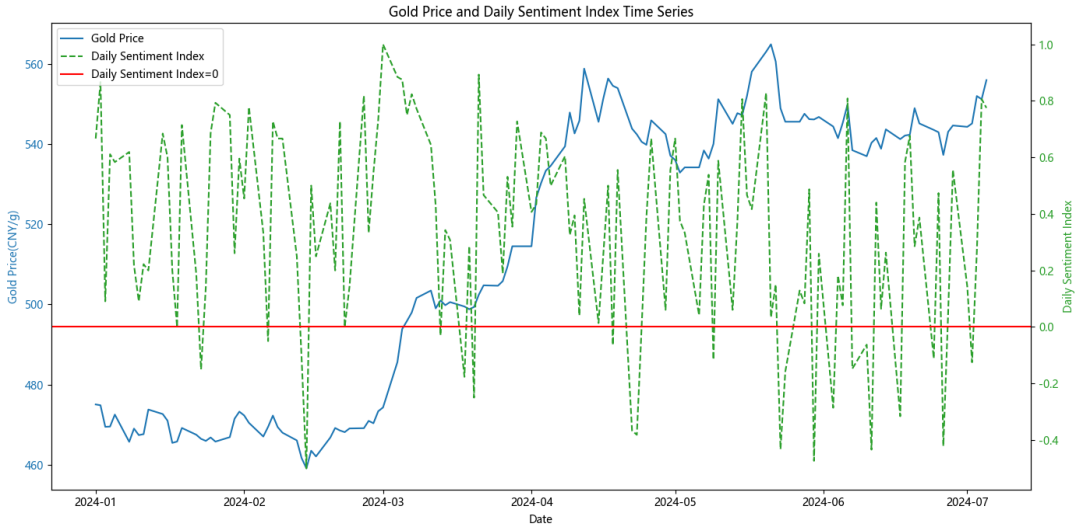

Figure 2 Gold price and daily sentiment index time series.

5.2. Prediction Model

Python was used to implement gold price prediction using various models, outputting prediction evaluation metrics. The performance comparison of all prediction models is summarized in Table 8.

Table 8 Performance comparison of prediction models.

	Training set				Test set			
Prediction Model	MAE	MAPE	RMSE	R^2	MAE	MAPE	RMSE	R^2
LSTM	3.08	0.97%	4.31	0.99	8.46	1.84%	10.57	0.95
GRU	2.8	0.89%	3.84	1	4.84	1.07%	5.94	0.98
LSTM + Daily Sentiment Index	2.76	0.89%	3.84	1	7.34	1.55%	9.34	0.96
GRU + Daily Sentiment Index	2.05	0.66%	2.86	1	3.96	0.85%	5.12	0.99

According to the results in Table 8, among the models compared, the "GRU + Daily Sentiment Index" group exhibits the best performance across all metrics. This indicates that GRU, when fused with market sentiment data, can more effectively capture the time-series characteristics and sentiment-driven patterns of gold prices, providing the optimal modeling solution for gold price forecasting.

Figure 3 and Figure 4 show the data fitting curve and the data learning loss curve, respectively, for the GRU model incorporating the daily sentiment index.

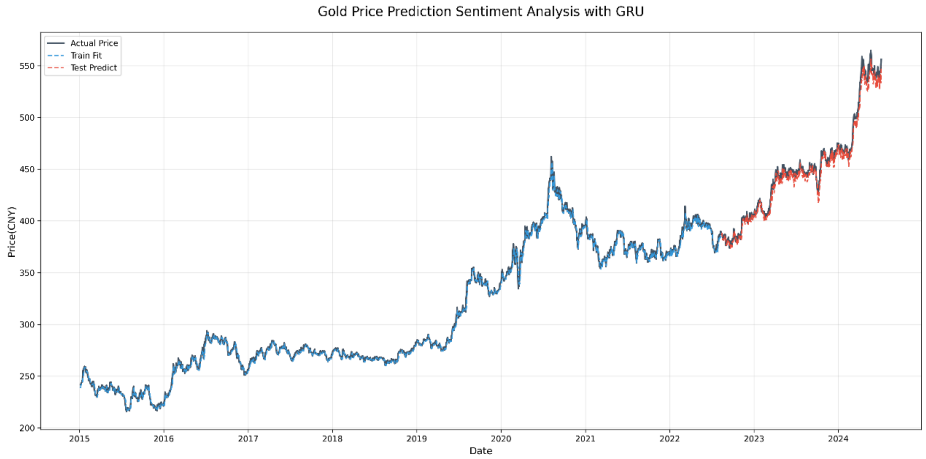

Figure 3 Fitting curve of GRU model with sentiment index.

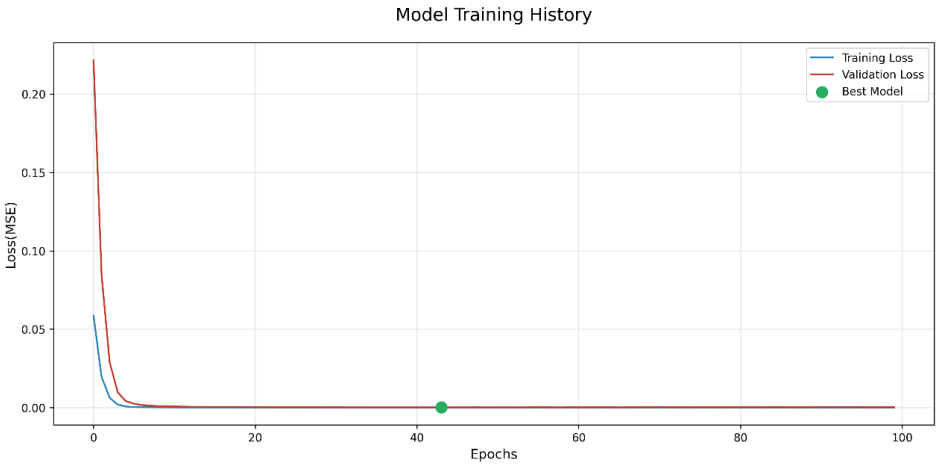

Figure 4 Learning loss curve of GRU model with sentiment index.

To further validate the effectiveness of the sentiment index for time-series models, the DM test was used to statistically compare the model incorporating this index with the baseline model. In this experiment, DM tests were conducted separately for the LSTM group and the GRU group, using Mean Squared Error (MSE) as the loss function. The null hypothesis (H_0) and alternative hypothesis (H_1) for the DM test are as follows:

H_0: The model incorporating the sentiment index and the model without it perform equally in

prediction.

H_1: The model incorporating the sentiment index performs better in prediction than the model without it.

The DM test results are shown in Table 9.

Table 9 DM test results.

Group	DM statistic	p-value
LSTM Group	-9.21	0.00
GRU Group	-16.05	0.00

From Table 9, the DM test results for both the LSTM group and the GRU group reject the null hypothesis ($P<0.01$) and accept the alternative hypothesis. At a 99% confidence level, time-series models incorporating the sentiment index perform significantly better in prediction than those without it.

6. Conclusion

This study effectively enhances the accuracy of gold price forecasting by constructing a multi-modal analysis framework that integrates BERT sentiment quantification and GRU time-series modeling. In sentiment classification, the fine-tuned BERT model achieved the highest accuracy (86%) in financial sentiment classification, outperforming ERNIE (83%). This demonstrates the superior capability of BERT in comprehending complex semantic structures. For predictive modeling, the GRU model integrating textual sentiment features yielded optimal forecasting performance. This validates that multi-modal data integration enhances prediction quality, where sentiment indicators mitigate limitations inherent in purely numerical models. The sentiment index (especially negative sentiment) significantly drives gold price movements, confirming the "sentiment contagion effect" in behavioral finance. Caution is warranted regarding the potential failure of sentiment factors in extreme market conditions (e.g., liquidity crises). Future research will focus on further optimizing sentiment quantification models to reduce biases from subjective labeling and exploring the integration of more types of unstructured data (such as social media data) with structured data to build more accurate and robust gold price forecasting models, providing stronger decision support tools for investors and financial institutions.

References

[1] Loughran T, McDonald B. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks[J]. The Journal of finance, 2011, 66(1): 35-65.

[2] Huang D, Jiang F, Tu J, et al. Investor sentiment aligned: A powerful predictor of stock returns[J]. The review of financial studies, 2015, 28(3): 791-837.

[3] ZHENG M G, PENG Q T, TAO S M, et al. Economic policy uncertainty, investor sentiment and gold price volatility[J]. Gold Science and Technology, 2022, 30(06): 891-900.

[4] LIU J E, GAO J H. Research on the dynamic relationship between investor sentiment and gold futures prices[J]. Price: Theory & Practice, 2017, (09): 80-83.

[5] Tetlock P C. Giving content to investor sentiment: The role of media in the stock market[J]. The Journal of finance, 2007, 62(3): 1139-1168.

[6] TANG G H, JIANG F W, ZHANG D S. Research progress on financial market text sentiment[J]. Economic Perspectives, 2016, (11): 137-147.

[7] Huang A H, Wang H, Yang Y. FinBERT: A large language model for extracting information from financial text[J]. Contemporary Accounting Research, 2023, 40(2): 806-841.

[8] Shi Y, An Y, Zhu X, et al. Better to hear all parties: understanding the impact of homophily in online financial discussion[J]. Electronic Commerce Research and Applications, 2022, 54: 101159.

[9] LIANG L Y, HUANG Y. Short-term forecasting method for gold futures prices -- Based on CEEMDAN and LSTM model analysis of COMEX gold futures price data[J]. Price: Theory & Practice, 2023, (09): 164-168.

[10] LI Y H, SUN J Y. Interval prediction of Chinese gold futures prices based on deep learning[J/OL]. Journal of Chongqing Technology and Business University (Natural Science Edition), (2024) 1-13. [Note: Provide DOI/URL if available]