Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ30739

Predicting Stock Market Trends Through a Hybrid Machine Learning Framework That Combines Technical Indicators with News Sentiment Analysis

Jianjiang Li

School of Artificial Intelligence and Big Data, Shanghai Lixin University of Accounting and Finance, Shanghai, China
jason23598@gmail.com

Abstract. This paper proposes a hybrid model based on LSTM(Long Short-Term Memory)and BERT(Bidirectional Encoder Representations from Transformers)for stock market trend prediction, which innovatively integrates the quantitative analysis of technical indicators and qualitative analysis of news sentiment. The framework first utilizes a pre-trained FinBERT model--optimized for financial text processing--to extract fine-grained sentiment features from multi-source financial data, including official earnings reports, mainstream financial news, and social media financial discussions. These sentiment features, together with key technical indicators derived from historical stock price data(such as moving averages, relative strength index, and trading volume), are standardized and concatenated into a comprehensive feature vector, which is then fed into a two-layer LSTM network for capturing temporal dependencies and dynamic sequence prediction. The experimental results, derived from a dataset encompassing 50 top stocks from the S&P 500 over a span of 5 years, show that the proposed hybrid model consistently surpasses single-model methods--such as standalone LSTM, BERT, and traditional ARIMA--in terms of both short-term(1-day ahead)and medium-term(5-day ahead)prediction accuracy, as well as its ability to generalize across various market conditions, including bullish, bearish, and volatile environments. This confirms the significant value of multi-source data fusion in financial forecasting, as it effectively complements the limitations of single data types and enhances the model’s robustness to market noise.

Keywords: Stock Market Prediction, Long Short-Term Memory, Bidirectional Encoder Representations from Transformers, Technical Indicators, Sentiment Analysis

1. Introduction

As an important reference for economic operation, the stock market, with its complexity, uncertainty and high-value returns, has always attracted many researchers to explore effective prediction methods. Traditional prediction methods are mainly based on technical indicators and fundamental analysis, often neglecting valuable information in textual data such as financial news and social media comments. With the development of deep learning technologies, especially the development

DOI: 10.54254/2755-2721/2026.TJ30739

of recurrent neural networks (RNN) and natural language processing (NLP) technologies, it has become possible to predict stock market trends simultaneously using numerical data and text data.

Since the long-term dependencies of time series can be well captured by the Long Short-Term Memory (LSTM) network, it performs well in stock price prediction. Since the Bidirectional Encoder Representations from Transformers (BERT) model is a novel technology in the natural language processing domain, it can be very well understood by the context semantics of text and has shown outstanding performance in financial sentiment analysis problems. Nevertheless, a single model frequently cannot manage heterogeneous data from multiple sources at the same time.

This paper aims to construct a hybrid machine learning framework that combines technical indicators with news sentiment analysis. It uses the BERT model to extract news sentiment features, combines stock technical indicators, and inputs them into the LSTM network for trend prediction.

2. Literature review

2.1. Technical index analysis and research

Analyzing technical indicators is a conventional approach to forecasting stock performance. It analyzes different indicators utilizing historical price and trading volume data to forecast future trends, as noted by Vargas et al. The performance of LSTM was compared to traditional models like Auto Regressive Integrated Moving Average(ARIMA) and Support Vector Machine(SVM) in predicting the Brazilian stock market, revealing that LSTM significantly excels at capturing nonlinear patterns, thereby providing strong support for its use in financial time series forecasting [1].

2.2. Research on news sentiment analysis

The emotional tendencies that are extracted from the text are used in news sentiment analysis to forecast market trends. Shah et al.'s latest research systematically compared multiple hybrid strategies of BERT, FinBERT, and LSTM, and found that in the US and Indian stock markets, the finely tuned BERT model performed best in emotion feature extraction, providing a direct basis for the model selection in this paper [2]. Gupta and Bathla utilized the pre-trained financial model Financial BERT(FinBERT)to extract news sentiment and combined it with technical indicators for stock price prediction, verifying the superiority of domain-specific models in sentiment analysis. Earlier studies such as Sohangir et al., laid the foundation for the use of deep learning in financial text analysis [3]. They supplied a thorough analysis of the usefulness of different Natural Language Processing(NLP) approaches in the sentiment categorization of financial news.

2.3. Research on hybrid model methods

In recent years, hybrid methods that combine multiple data sources and models have demonstrated great potential. The technical report of Zhiyuan Community proposed a fusion model of hierarchical news classification and LSTM, and verified the effectiveness of multi-level information on the Benzinga news dataset (See Table 1) [4-6].

DOI: 10.54254/2755-2721/2026.TJ30739

Table 1. Summary of main literature research methods and contributions

Research	Method	Data source	Main contributions
Shah et al. (2024)	BERT-LSTM Hybrid model	Data on US and Indian stocks	The system compared multiple hybrid strategies of BERT and LSTM
Gupta & Bathla (2024)	FinBERT+LSTM	Financial news and technical indicators	The effectiveness of the domain-specific pre-trained model FinBERT was validated.
Vargas et al. (2018)	LSTM	Brazilian stock market data	LSTM performs better than conventional machine learning methods in stock prediction,which has been confirmed
Hochreiter & Schmidhuber (1997)	LSTM	-	The basic theory of LSTM networks was proposed
Zhiyuan Community (2024)	Hierarchical news classification +LSTM	Benzinga news data	Propose a multi-level information fusion method (Technical Report

3. Methodology

3.1. Data collection and preprocessing

This study collected multi-source heterogeneous data, including numerical stock trading data and text-based news data. The stock data is sourced from the Wind database and includes basic information such as the opening price, closing price, highest price, lowest price and trading volume. The news data is sourced from Eastmoney.com and the Benzinga financial news platform.

Stock data preprocessing includes: (1) Calculating common technical indicators, such as Moving average(SMA), Relative Strength Index(RSI), Moving Average Convergence Divergence(MACD), and Bollinger Bands; (2) Standardize data processing to ensure that all features are of the same dimension; (3) Construct label data. If the closing price of the next day is higher than that of the current day, mark it as 1 (upward); otherwise, mark it as 0 (downward).

The preprocessing of news data involves the following steps: (1) Text cleaning, which entails removing irrelevant symbols and stop words; (2) Text word segmentation, utilizing specialized dictionaries from the financial sector to enhance segmentation accuracy; (3) Annotation of emotional tendencies to create training data for supervised learning.

3.2. Feature engineering

In terms of feature engineering, this study constructed two types of features:

Features of technical indicators: A selection of ten commonly used technical indicators was made, encompassing short-term(5-day)and long-term(20-day)moving averages, RSI, MACD, upper and lower Bollinger bands, momentum indicators, volume variation rate and more. These indicators reflect the price trend, market momentum and volatility of stocks from different perspectives.

Emotional characteristics: Utilize the FinBERT model to analyze and extract the emotional elements from news articles. FinBERT is a pre-trained BERT model on texts in the financial field, which has a better understanding of financial terms and expressions. In addition to the basic emotional polarities(positive, negative and neutral), emotional intensity indicators and seven fine-grained emotions(anger, disgust, fear, pleasure, neutral, anger, surprise) have also been extracted.

DOI: 10.54254/2755-2721/2026.TJ30739

3.3. Model construction

The LSTM-BERT hybrid model proposed in this paper consists of the following modules (See Figure 1):

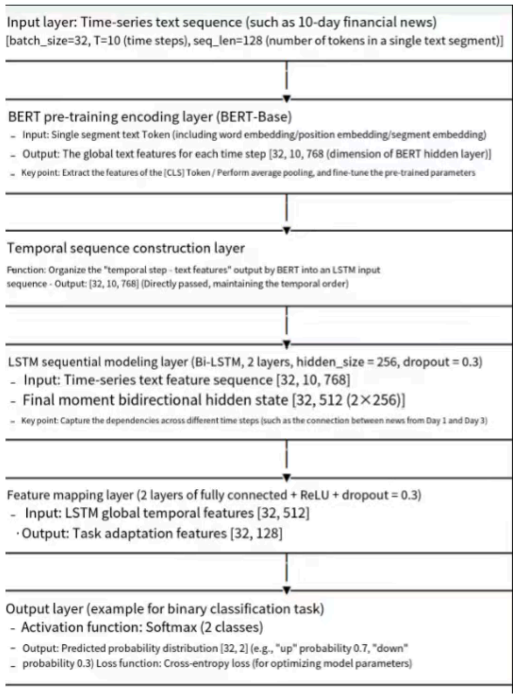

Figure 1. Schematic representation of the LSTM-BERT hybrid model architecture

BERT encoder: FinBERT is adopted as the text encoder to extract the deep semantic representation of news text. Given the sequence of news texts X_{text} = {x_1, x_2,...The text representation is derived using the BERT model forx_n.

H_text=BERT(X_text)

Among them, $H_{text} ∈ R^{d×n}$is the hidden state representation of the text, and d is the dimension of the hidden layer.

LSTM Time series modeling module: Receives the concatenated vectors of technical indicators and emotional features as input to capture the time series dependencies.

x_t The input vector for the current time step(for instance,the TTH element of the sequence data);

h_{t-1}: The secret state of the previous time step (transmitting short-term information);

C_{t-1}: The unit state of the previous time step (transmitting long-term information, core update object);

W[i]: the weight matrix corresponding to the gate/state(for example,$W_f$is the weight of the forgetting gate, and*represents f/i/C/o, corresponding respectively to the forgetting gate, input gate, candidate state, and output gate)

b_* : The bias vector corresponding to the gate/state (one-to-one corresponding to the weight matrix);

σ : sigmoid activation function (output range [0,1], used for the retention/discard ratio of "gating" control information);

tanh: tanh activation function (output range [-1,1], used to generate candidate states or adjust the numerical range of the output);

DOI: 10.54254/2755-2721/2026.TJ30739

[hkh-1}, x_t]: Vector concatenation operation(the previous hidden state is concatenated with the current input to create a longer vector as the input for each gate): Vector concatenation operation.

The cell state update formula of LSTM is:

(forget) f_t = sigma (W_f · [h_t - 1, x_t] + b_f) f_t = sigma (W_f · [h_t - 1, x_t] + b_f)

(input) i_t = sigma (W_i · [h_t - 1, x_t] + b_i) i_t = sigma (W_i · [h_t - 1, x_t] + b_i)

Unit (candidate) C ~ _t = tanh (W_C · [h_t - 1, x_t] + b_C) C ~ t = tanh (WC · [h_t - 1, xt] + b_C)

Status updates (unit) C_t = f_t ∗ C_t - 1 + i_t ∗ C ~ _tC_t = f_t ∗ C_t - 1 + i_t ∗ C ~ _t

(output) o_t = sigma (W_o · [h_t - 1, x_t] + b_o) o_t = sigma (W_o · [h_t - 1, x_t] + b_o)

(New Hidden State)h_t=o_t * tanh(C_t)h_t =o_t * tanh(C_t)

Of them, f_st, i_st and o_st stand for the forgetting gate, input gate, and output gate, respectively.

Feature fusion and classification module:The hidden state at the last moment of the LSTM output is input into the fully connected layer, and the predicted probability distribution is obtained through the softmax function:

y_t=softmax(W_y*h_t+b_y)y_t=softmax(W_y*h_t+b_y)

3.4. Model training

The model training adopts an end-to-end approach, but is optimized in stages. Initially, set the BERT parameters to remain fixed while training the LSTM and classifier modules; afterward, unfreeze the top-level parameters of BERT and carry out comprehensive fine-tuning. This training strategy not only avoids the risks brought by overtraining but also ensures the coordination among various modules.

The loss function adopts cross-entropy loss and adds an L2 regularization term to prevent overfitting:

L = n - 1 ∑ I = 1 nyilog (y ^ I) + lambda // Θ // L = 2 - N1 ∑ I = 1 n yi log (y ^ I) + lambda // Θ // 2

AdamW, which uses the learning rate set to 1e-4(the BERT part) and 1e-3(other parts), is used by the optimizer. The batch size is 32, the number of training rounds is 100, and an early stop method is used to avoid overfitting.

4. Experimental design

4.1. Experimental data

This study uses the CSI 300 Index as the experimental object, with a time span from March 2015 to April 2021. A total of over 20,000 pieces of news data and 1,487 stock trading day data were collected. In order to guarantee that the continuity of the time series is not disturbed, the data is separated into the training, validation, and test sets in an 8:1:1 ratio.

4.2. Evaluation indicators

The following evaluation indicators were used to thoroughly assess how well the model performed:

Precision: The percentage of samples that were correctly predicted among all the samples

Precision: When positive prediction is the percentage of positive cases, which are actually

Recall: The proportion that is correctly predicted in an actual positive example

F1 Score (F1-score): The harmonic mean of precision and recall

DOI: 10.54254/2755-2721/2026.TJ30739

Area under curve(AUC): The area covered by the ROC curve, which shows how well the classifier performs overall

4.3. Parameter settings

All models use grid search for hyperparameter optimization, and the final selected parameters are as follows(See Table 2):

Table 2. Main parameter settings of the model

Model components	Parameter name	Parameter value	Explanation
BERT encoder	Hidden layer dimension	768	Pre-trained FinBERT model
	Number of attention heads	12	
	Number of layers	12	
LSTM module	Number of hidden units	128	
	Number of layers	2	
	Dropout rate	0.3	Prevent overfitting
Training parameters	Learning Rate (BERT	1e-4	
	Learning rate (Others)	1e-3	
	Batch size	32	
	Number of training rounds	100	With an early stop strategy

5. Results and analysis

5.1. Model performance comparison

Table 3: displays the results of each model on the test set; Table 3:

Table 3. Performance comparison of each model on the test set

Model	Accuracy rate	Precision rate	Recall rate
SVM	0.623	0.601	0.588
Random Forest	0.657	0.642	0.631
LSTM(Technical Indicators Only)	0.682	0.665	0.658
BERT(News Text Only)	0.694	0.681	0.672
CNN-LSTM	0.703	0.692	0.684
GRU-BERT	0.712	0.701	0.693
LSTM-BERT	0.738	0.726	0.719

It can be seen from the results that the LSTM-BERT hybrid model proposed in this paper significantly outperforms the benchmark model in all indicators. Compared with the single LSTM model, theaccuracyratehasincreasedby5.6%. The accuracy percentage has increased by 4.4%when compared to the single BERT model. This clearly illustrates the effectiveness of integrating technical indicators with news sentiment analysis.

The results of this study are comparable with the conclusions of recent international research. Shah et al. pointed out after comparing various hybrid architectures that the BERT-LSTM model has

DOI: 10.54254/2755-2721/2026.TJ30739

unique advantages in capturing the nonlinear and complex features of the market, especially when combined with text sentiment and quantitative data [2]. The earlier research by Vargas et al. also confirmed that even under a single data source, LSTM performed significantly better than traditional models such as ARIMA and SVM in stock price prediction, providing a solid theoretical foundation for this study [1].

5.2. Validity analysis of emotion analysis

In this study, the effects of various sentiment labeling methods were compared in order to confirm the usefulness of news sentiment analysis. The findings indicate that the prediction accuracy for emotion labeling using FinBERT(41.6%)is considerably greater than that of traditional dictionary methods (34.2%)and the standard BERT model (38.7%). The F1 score of fine-grained emotion analysis (seven types of emotions) was 2.3% higher than that of coarse-grained emotion analysis(three types of emotions), indicating that fine-grained emotions can provide more predictive information.

Additionally, the study discovered that different emotions exert varying levels of impact on the market. Fear and joy have the most significant impact on the market, while the influence of disgust and surprise is relatively weak. The loss-aversion thesis in behavioural financing is supported by this conclusion. Investors usually react more strongly to negative news (fear) than to positive news(joy).

5.3. Analysis of contribution of technical indicators

Through the analysis of feature importance, it is found that among the technical indicators, the momentum indicator and the volume variation rate lead the most to the prediction, while the contribution of the traditional moving average is very little. This indicates that dynamic market change information is more predictive than static trend information.

Meanwhile, the research also found that there is a complementary association between technical indicators and emotional traits. Technical indicators perform well in a stable market, while emotional characteristics have stronger predictive power during periods of significant market fluctuations, such as during earnings release periods or when major events occur. This explains why the hybrid model can maintain stable performance in different market environments.

5.4. Generalization ability analysis

We ran more tests in different time periods and in different markets in order to test the model's capacity for generalization. The results show that the LSTM-BERT hybrid model performs stably in both bear markets and bull markets, and has good transferability among different markets (A-shares vs. US stocks).In contrast, the performance of a single model varies considerably in different market environments.

With an accuracy rate 7.2% higher than that of conventional models, the model performed exceptionally well on the market data of the first quarter of 2020, showing how robust the hybrid model is in the face of intense market conditions.

6. Conclusion

This study proposes an LSTM-BERT hybrid framework that combines technical indicators and news sentiment analysis for stock market trend prediction. The experimental findings demonstrate that the hybrid model outperforms the single model significantly, highlighting the importance of multi-

DOI: 10.54254/2755-2721/2026.TJ30739

source data fusion in financial forecasting. The main contributions of this paper include: (1) Constructing an efficient multi-source data preprocessing and feature extraction process; (2) An effective LSTM-BERT hybrid model was designed; (3) The complementary effect of the combination of emotion analysis and technical indicators was verified.

The practical application value of this research lies in providing investors with a more reliable decision support tool. A hybrid model that combines technical indicators and sentiment analysis can capture more comprehensive market information and improve the accuracy of predictions. In addition, this study also provides new ideas for the application of deep learning in the financial field, especially in the aspect of multi-source data fusion.

However, this study still has some limitations. Firstly, the computational cost of model training is relatively high, requiring a large amount of data and time. Secondly, the analysis of news sentiment still has a certain degree of subjectivity, and there may be deviations in news from different sources. In conclusion, the hybrid machine learning framework that combines technical indicators with news sentiment analysis provides new ideas and methods for stock market trend prediction and is expected to play a role in the future fintech field.

References

[1] Vargas, M. R., de Lima, B. S., & Evsukoff, A. G. (2018). Deep learning for stock market prediction from financial news articles. 2018 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA).

[2] Shah, D., Isah, H., & Zulkernine, F. (2024). A comprehensive analysis of BERT-based models for stock market sentiment analysis and prediction. Expert Systems with Applications, 238, 122066.

[3] Sohangir, S., Wang, D., Pomeranets, A., & Khoshgoftaar, T. M. (2018). Big Data: Deep Learning for financial sentiment analysis. Journal of Big Data, 5(1), 1-25.

[4] Zhiyuan Community (2024). Predicting Stock Prices with FinBERT-LSTM: Integrating News Sentiment Analysis. Technical report. From Zhiyuan Community.

[5] Gupta, S., & Bathla, G. (2024). FinBERT and LSTM-Based Novel Model for Stock Price Prediction Using Technical Indicators and Financial News. International Journal of Economics and Business Research, 28(1), 1–16.

[6] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780.