Jurnal Masyarakat Informatika, 17(2), 2026
DOI:10.14710/jmasif.17.2.82051
ISSN: 2777-0648

Comparative Evaluation of FinBERT and IndoBERT for LSTM-Based Stock Price Prediction in Indonesia

Heri Santosa*1), Kusrini1), Rodrigo Martínez-Béjar2)

1)*Department of Computer Science, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia*
2)*Department of Computer Science, Universidad de Murcia, Murcia, Spain*
* *Corresponding author: heri.sant@students.amikom.ac.id*

Abstract

Stock price prediction remains challenging due to the nonlinear and dynamic behavior of financial markets, where price movements are influenced not only by historical market data but also by investor sentiment reflected in financial news. This study investigates the effectiveness of sentiment integration for stock price prediction by comparing FinBERT, a financial domain-specific language model, and IndoBERT, a language-specific Indonesian language model, within an LSTM-based forecasting framework. Experiments were conducted using daily stock price data and financial news collected for three Indonesian stocks (BBCA, BBRI, and BSDE) over the period 2019–2025. Sentiment scores extracted from financial news were aggregated on a daily basis and incorporated as additional features alongside historical price variables. Model performance was evaluated using MAE, RMSE, MAPE, R², and Directional Accuracy. The results indicate that sentiment-enhanced models do not consistently improve numerical forecasting accuracy compared with the baseline LSTM model. However, sentiment integration does not consistently improve directional prediction, although limited stock-specific differences are observed under certain stock-specific conditions. Comparative analysis further shows that neither FinBERT nor IndoBERT consistently outperforms the other across all datasets and metrics, suggesting that sentiment effectiveness is highly dependent on linguistic context and market characteristics. These findings highlight that sentiment information should be incorporated selectively rather than assumed to universally improve stock forecasting performance in emerging markets.

Keywords: *Stock price prediction, sentiment analysis, LSTM, FinBERT, IndoBERT.*

1 Introduction

Stock price prediction remains a challenging task because financial markets are dynamic, nonlinear, and highly sensitive to both historical market behavior and external information. Price movements are influenced not only by technical variables such as opening price, closing price, trading volume, and volatility, but also by investor expectations, financial news, market uncertainty, and broader economic conditions. Although traditional statistical forecasting models such as ARIMA have been widely used in financial time-series prediction, recent studies show that hybrid and deep learning approaches are often more suitable for capturing nonlinear and non-stationary stock market behavior [1].

Received: 29 January 2026; Revised: 05 June 2026; Accepted: 19 July 2026; Published: 08 August 2026

Among deep learning approaches, Long Short-Term Memory (LSTM) networks remain widely used for stock price prediction because they are designed to model sequential data and capture long-term temporal dependencies. The LSTM architecture introduced by Hochreiter and Schmidhuber (1997) addresses the vanishing-gradient problem commonly found in conventional recurrent neural networks. Recent studies continue to use LSTM and its variants for financial forecasting, including attention-based LSTM, stacked LSTM, LSTM-DNN, CNN-BiLSTM-Attention, and hybrid LSTM models [2-6]. These studies indicate that LSTM remains relevant as a forecasting baseline and comparative architecture in stock market prediction.

However, stock prediction models that rely only on historical price data may fail to capture the influence of textual information, particularly sentiment reflected in financial news and public discussions. Investor sentiment can influence market expectations and trading decisions, especially during periods of uncertainty. Previous studies have demonstrated that sentiment extracted from textual data can support stock movement prediction [7]. More recent studies also show that integrating sentiment analysis with deep learning models may improve stock forecasting performance under certain market conditions, although the improvement is not always consistent across datasets, sectors, or time periods [6-9].

The development of transformer-based language models has further improved the quality of sentiment extraction from financial texts. FinBERT is a financial domain-specific language model designed to capture financial terminology and contextual meaning in financial communication. Several studies have shown that FinBERT can be used to extract financial sentiment and integrate it with forecasting models for stock price or stock movement prediction [9-13]. These findings suggest that domain-specific sentiment models may provide useful information when financial news contains market-relevant signals.

In the Indonesian context, language-specific sentiment representation is also important because local financial news contains Indonesian linguistic structures, expressions, and contextual meanings that may not be fully captured by English-based financial sentiment models. IndoBERT, introduced through the IndoNLU benchmark, provides a strong pre-trained language model for Indonesian natural language understanding tasks [14]. Indonesian BERT-based models have also been shown to perform effectively in Indonesian sentiment and text classification tasks, including IndoBERTweet and fine-tuned IndoBERT variants [15-18]. Therefore, IndoBERT may offer advantages when extracting sentiment from Indonesian-language financial news.

Despite these developments, several research gaps remain. First, many sentiment-enhanced stock prediction studies focus on a single sentiment model, making it difficult to determine whether performance differences are influenced by financial-domain specificity, language specificity, or the forecasting architecture itself. Second, comparative studies between FinBERT as a financial domain-specific model and IndoBERT as an Indonesian language-specific model remain limited, particularly in the Indonesian stock market. Third, many prior studies emphasize numerical forecasting accuracy while providing limited discussion of directional accuracy, statistical significance, and the practical meaning of model improvements.

To address these gaps, this study compares FinBERT-enhanced LSTM, IndoBERT-enhanced LSTM, and baseline LSTM models for stock price prediction in Indonesia. The experiments are conducted on three Indonesian stocks, namely BBCA, BBRI, and BSDE, using historical stock price data and financial news collected from CNBC Indonesia during the 2019–2025 period. Model performance is evaluated using RMSE, MAE, MAPE, R^2, Directional Accuracy, and statistical significance testing. This study aims to determine whether sentiment integration improves stock price prediction performance and whether FinBERT or IndoBERT provides more effective sentiment representation for Indonesian stock forecasting.

The contributions of this study are threefold. First, this study provides a comparative evaluation of financial domain-specific and Indonesian language-specific sentiment models within the same LSTM-based stock prediction framework. Second, this study evaluates model performance not only using numerical error metrics, but also using Directional Accuracy and statistical significance testing. Third, this study provides empirical evidence that the effectiveness of sentiment integration in stock prediction is stock-dependent and should not be assumed to universally improve forecasting performance.

2 Literature Review

2.1 Stock Price Prediction Using Deep Learning

Stock price prediction has been widely studied using statistical, machine learning, and deep learning approaches. Traditional time-series models such as ARIMA are commonly used for financial forecasting, but their effectiveness is often limited when dealing with nonlinear, non-stationary, and volatile stock market data. Recent studies therefore increasingly adopt hybrid and deep learning models to improve forecasting performance, including wavelet-ARIMA-LSTM, ARIMA-LSTM comparison, attention-based LSTM, LSTM-DNN, and CNN-BiLSTM-Attention architectures [1-3, 5, 19].

Long Short-Term Memory (LSTM) networks are widely used in stock price prediction because they are designed to learn sequential dependencies and overcome the vanishing-gradient problem in conventional recurrent neural networks. The model was originally introduced by Hochreiter and Schmidhuber (1997) and has since become a common foundation for financial time-series forecasting. Earlier and recent studies show that LSTM-based models can capture temporal dependencies in stock market data and remain useful either as standalone forecasting models or as comparative baselines for more complex architectures [3, 4, 20-22].

Recent developments also show that LSTM is frequently extended with additional mechanisms to improve predictive performance. Attention mechanisms are used to emphasize important temporal patterns, while convolutional layers can capture local price movement structures before sequential learning is performed. Hybrid models that combine LSTM with other neural architectures have been reported to improve forecasting accuracy in several financial datasets [2, 5, 19]. These findings indicate that LSTM remains relevant in stock prediction research, especially as a baseline architecture for evaluating the added contribution of sentiment features.

2.2 Sentiment Analysis for Financial Market Prediction

Financial markets are influenced not only by historical price movements but also by investor sentiment and external information. News articles, public disclosures, and social media discussions can shape investor expectations and affect trading behavior. Sentiment analysis therefore provides a way to convert textual information into numerical signals that can be integrated into prediction models.

Earlier research has shown that social media sentiment and topic-based sentiment features can support stock movement prediction [7], [23].

Recent studies continue to explore the integration of sentiment analysis with deep learning models. Sentiment-enhanced forecasting has been applied to stocks, banking-sector assets, and cryptocurrency markets using LSTM, GRU, and transformer-based sentiment representations [6, 8, 9, 24]. These studies suggest that sentiment information can provide complementary signals when textual data reflects market-relevant expectations or uncertainty.

However, the benefit of sentiment integration is not always consistent. Sentiment features may improve prediction performance when news coverage is relevant and timely, but they may also introduce noise when the sentiment signal is weak, delayed, or unrelated to the target stock. Under uncertain market conditions, sentiment information can behave differently across sectors and time periods [6], [8]. Therefore, sentiment-enhanced forecasting models should be evaluated not only using numerical error metrics, but also using Directional Accuracy and statistical significance testing.

2.3 FinBERT for Financial Sentiment Analysis

Transformer-based language models have improved sentiment extraction because they can capture contextual meaning more effectively than traditional lexicon-based or conventional machine learning methods. FinBERT is one of the most widely used financial domain-specific language models for financial sentiment analysis. It is designed to capture finance-related terminology, contextual polarity, and domain-specific expressions in financial communication [10-12].

FinBERT has also been applied in prediction-oriented financial research. Several recent studies integrate FinBERT-generated sentiment scores with sequence forecasting models such as LSTM and GRU to support stock or cryptocurrency prediction [9, 13, 24]. These findings suggest that financial domain-specific sentiment representation may provide useful predictive information, especially when the analyzed texts contain explicit financial terminology or market-relevant news signals.

Nevertheless, FinBERT is primarily developed for financial-domain texts that are largely English-based. This creates a potential limitation when the model is applied to Indonesian financial news. Although FinBERT may capture finance-specific terms, it may not fully represent Indonesian linguistic nuances, local expressions, and contextual meanings. Therefore, its effectiveness in Indonesian stock prediction needs to be compared with language-specific Indonesian models.

2.4 IndoBERT and Indonesian Language-Specific Sentiment Model

Financial markets are influenced not only by historical price movements but also by investor sentiment and external information. News articles, public disclosures, and social media discussions can shape investor expectations and affect trading behavior. Sentiment analysis therefore provides a way to convert textual information into numerical signals that can be integrated into prediction models. Earlier research has shown that social media sentiment and topic-based sentiment features can support stock movement prediction [7], [23].

Language-specific models are important for Indonesian sentiment analysis because Indonesian texts contain linguistic structures, local entities, informal expressions, and contextual meanings that may not be fully captured by English-based or multilingual models. IndoNLU provides benchmark resources for Indonesian natural language understanding and includes IndoBERT as one of the important pre-trained models for Indonesian NLP tasks [14]. In addition, IndoBERTweet demonstrates the value of Indonesian domain-specific vocabulary initialization for social media text [15].

Recent studies on Indonesian text classification show that IndoBERT-based models remain competitive for various sentiment and multilabel classification tasks. Comparative evaluations involving IndoBERT, IndoBERTweet, and mBERT indicate that Indonesian BERT-based representations are effective for capturing local linguistic context [16]. Fine-tuned IndoBERT has also been applied successfully in aspect-based sentiment analysis tasks, such as hospital service reviews and other Indonesian review datasets [17], [18].

In Indonesian stock prediction, IndoBERT may provide advantages when the analyzed news articles are written in Indonesian and contain local expressions or context-specific sentiment. However, IndoBERT is not specifically trained on financial corpora, so it may be less effective than FinBERT when the text contains dense financial terminology. This creates an important empirical question: whether domain specificity from FinBERT or language specificity from IndoBERT provides better sentiment representation for Indonesian stock price prediction.

2.5 Research Gap

Based on the reviewed literature, several research gaps remain. First, although LSTM-based models have been widely used for stock price prediction, many studies still focus mainly on historical market data and technical indicators. The incremental value of sentiment information remains inconsistent and requires further empirical evaluation. Second, many sentiment-enhanced stock prediction studies use only one sentiment model, making it difficult to determine whether performance improvements are caused by domain-specific sentiment representation, language-specific representation, or the forecasting model itself.

Third, comparative studies between FinBERT as a financial domain-specific model and IndoBERT as an Indonesian language-specific model remain limited, particularly in the Indonesian stock market. This gap is important because Indonesian financial news may contain both financial terminology and local linguistic context. Fourth, many previous studies emphasize numerical prediction accuracy, while fewer studies discuss Directional Accuracy, statistical significance, and the practical interpretation of small performance differences.

To address these gaps, this study compares baseline LSTM, FinBERT-enhanced LSTM, and IndoBERT-enhanced LSTM models for Indonesian stock price prediction. The evaluation uses RMSE, MAE, MAPE, R^2, Directional Accuracy, and Wilcoxon Signed-Rank testing to assess whether sentiment integration produces meaningful improvements. By comparing three Indonesian stocks with different characteristics, this study provides empirical evidence on when sentiment integration helps or fails to improve LSTM-based stock prediction.

3 Research Methods

3.1 Research Framework

This study adopts an experimental quantitative approach to evaluate the contribution of sentiment information to stock price prediction performance. Experimental research is appropriate for comparing alternative predictive models under controlled conditions because it enables systematic assessment of the influence of individual variables while maintaining consistent model configurations [25].

The proposed framework consists of six stages: data collection, text preprocessing, sentiment extraction, feature engineering, model development, and model evaluation. Three experimental scenarios are constructed to enable direct comparison. The first scenario employs a baseline Long Short-Term Memory (LSTM) model using only historical stock market variables. The second scenario integrates sentiment features generated by FinBERT into the baseline model. The third scenario incorporates sentiment features generated by IndoBERT into the same LSTM architecture.

To ensure a fair comparison, all experimental scenarios use identical network architecture lookback windows, training procedures, optimization strategies, and evaluation metrics. Consequently, the experimental design was intended to isolate the contribution of sentiment features while minimizing differences arising from model architecture and training configuration [2], [3].

The experiment is conducted using three Indonesian stocks: Bank Central Asia (BBCA), Bank Rakyat Indonesia (BBRI), and Bumi Serpong Damai (BSDE). These companies were selected because they represent different industry sectors and volatility characteristics, allowing evaluation of whether sentiment integration produces consistent effects across varying market conditions.

The overall research workflow is illustrated in Figure 1. The framework shows how historical price data and financial news are processed through separate preprocessing stages, merged using a Feature-level fusion strategy, and then evaluated under baseline, FinBERT-enhanced, and IndoBERT-enhanced LSTM scenarios.

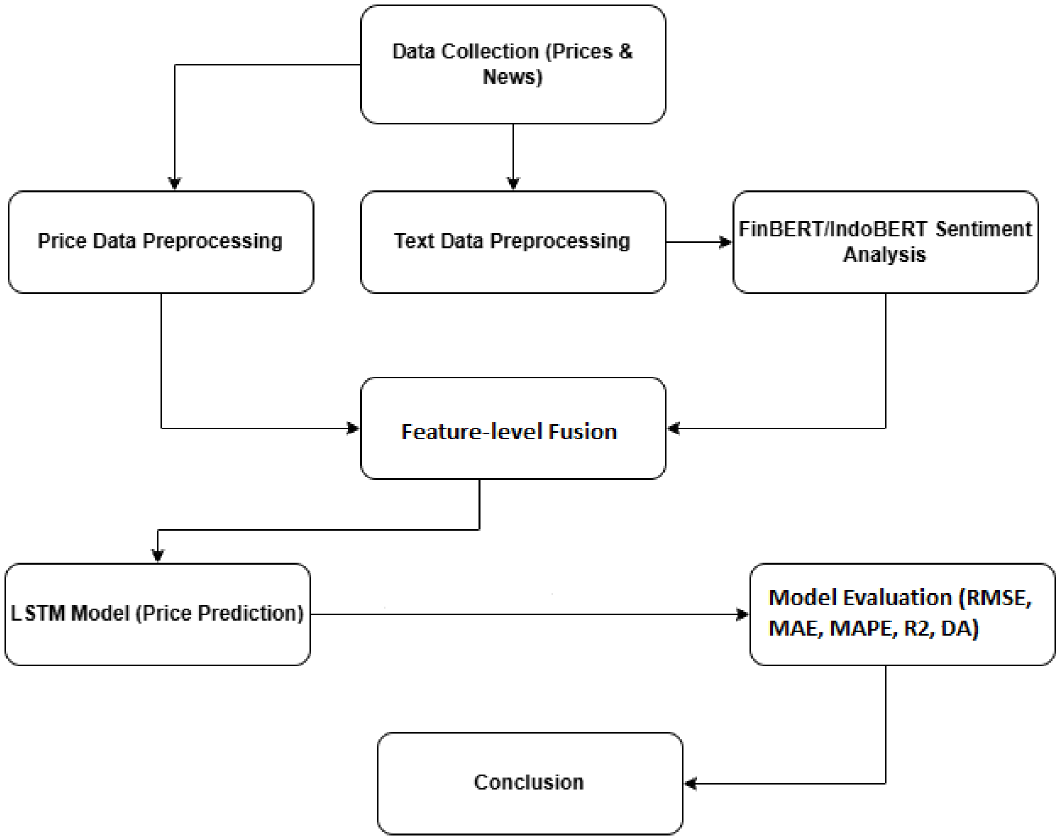

Figure 1 Overall Research Framework for Sentiment-Enhanced LSTM Stock Price Prediction

3.2 Data Collection and Dataset Description

The dataset consists of two primary sources: historical stock market data and financial news data. Historical stock data were collected from January 2019 to December 2025 and include daily opening price, highest price, lowest price, closing price, and trading volume. The closing price is used as the prediction target because it represents the most widely adopted benchmark variable in stock forecasting research [3], [26].

Financial news articles were collected from CNBC Indonesia using stock-specific keywords related to BBCA, BBRI, and BSDE. For each target trading date, the retrieval procedure first searched for a relevant article published on the same date. When no unique article was found, the search was extended backward for up to nine preceding calendar days. The first relevant article that had not previously been selected was retained, and its actual publication date was recorded separately from the target trading date. Consequently, the number of retrieved records should not be interpreted as the number of trading days on which a new article was necessarily published. Each news record contains publication date, article title, and source information. News articles were subsequently aligned with trading dates to enable sentiment extraction and integration into the forecasting model.

The use of financial news as an external information source is motivated by previous studies demonstrating that investor sentiment significantly influences stock market behaviour and price movements [7], [27]. The main characteristics of the financial news dataset are presented in Table 1, providing an overview of the dataset size, source distribution, and labeling structure.

Table 1 Financial News Dataset Summary

Stock	**Source**	**Period**	**Target Trading Dates**	**Retrieved Unique Articles**	**Target Dates Without a Retrieved Article**
BBCA	CNBC Indonesia	2019–2025	1,706	1,706	0
BBRI	CNBC Indonesia	2019–2025	1,706	1,706	0
BSDE	*CNBC Indonesia*	*2019–2025*	*1,705*	*555*	*1,150*

Note: A retrieved article may have been published on the target trading date or within the preceding nine calendar days. Therefore, "Retrieved Unique Articles" does not necessarily represent same-day news coverage.

3.3 Text Data Processing

Before sentiment analysis, the collected news data underwent several preprocessing procedures to improve consistency and reduce noise. These procedures included duplicate removal, missing-value handling, date normalization, and alignment of publication dates with corresponding trading days. Articles that could not be associated with the target company were excluded from the analysis.

Unlike traditional machine learning approaches that rely heavily on stemming and handcrafted linguistic features, transformer-based language models utilize contextual tokenization mechanisms capable of preserving semantic information within the original text [28]. Therefore, the textual content was processed directly using the tokenizer associated with each pre-trained language model. This preprocessing strategy preserves contextual relationships among words and enables the transformer architecture to generate contextual representations that better capture sentiment information in financial texts [28], [29].

3.4 Sentiment Feature Extraction and Aggregation

Sentiment extraction was performed using two transformer-based language models: FinBERT and IndoBERT. Because the original CNBC Indonesia articles were written in Indonesian, the texts used in the FinBERT pipeline were first translated into English using an automated translation procedure. Long articles were divided into segments of up to approximately 4,500 characters during translation and subsequently recombined before tokenization. The translated text was then processed using the yiyanghkust/finbert-tone tokenizer and classification model, with truncation applied at a maximum sequence length of 512 tokens. In contrast, the Indonesian texts were processed directly, without translation, using the taufiqdp/indonesian-sentiment Indonesian BERT-based sentiment classification model with a maximum sequence length of 256 tokens. This design enables comparison between a financial domain-specific representation applied to translated English text and an Indonesian language-specific representation applied directly to the source text.

FinBERT is a domain-specific language model trained on financial corpora and specifically designed for financial sentiment classification tasks [10], [12]. Because FinBERT was developed using financial documents and financial terminology, it is expected to capture sentiment patterns relevant to investment decisions more effectively than general-purpose language models. IndoBERT is a pre-trained Indonesian language model designed to capture contextual and semantic characteristics unique to Indonesian texts [14]. Recent evaluations also demonstrate that IndoBERT-based architectures perform competitively across various Indonesian natural language processing tasks, including sentiment classification [16].

Each news article was processed independently to obtain sentiment polarity predictions. For each article, the model logits were transformed into class probabilities using the softmax function. A continuous sentiment score was then calculated as P(positive) − P(negative), producing a value that approaches +1 for strongly positive texts and −1 for strongly negative texts, while values near zero indicate neutral or weakly polarized content. The resulting scores were stored as Sentiment_Fin and Sentiment_Indo for the FinBERT and Indonesian-language pipelines, respectively. For descriptive sentiment labels, the class with the highest probability was selected only when its confidence was at least 0.60; otherwise, the article was labeled neutral. The continuous probability-difference score, rather than the categorical label, was used as the numerical input feature for the LSTM model.

Because stock prediction is performed at the daily level, article-level sentiment scores were aggregated into daily sentiment indicators. When multiple articles were published on the same trading day, sentiment scores were averaged to represent the overall sentiment environment available to investors. For trading days without relevant news coverage, a neutral sentiment value of zero was assigned. This strategy preserves temporal continuity and avoids introducing missing observations into the time-series dataset.

The use of two sentiment models enables direct comparison between a financial domain- specific language model and a language-specific Indonesian language model, thereby providing empirical evidence regarding the most suitable sentiment representation for Indonesian stock prediction. The sentiment-label distribution presented in Table 2 indicates whether the dataset is balanced across positive, negative, and neutral classes.

Table 2 Distribution of Sentiment Labels

Stock	Model	Positive	Neutral	Negative
BBCA	FinBERT	95	489	1,122
BBCA	IndoBERT	461	1,012	233
BBRI	FinBERT	45	630	1,031
BBRI	IndoBERT	828	770	108
BSDE	FinBERT	25	199	331
BSDE	IndoBERT	149	365	41

3.5 Feature Engineering

Feature engineering was conducted separately for each experimental scenario. The baseline model used five historical market variables: opening price, highest price, lowest price, closing price, and trading volume. In the FinBERT-enhanced scenario, the Sentiment_Fin variable was added to the baseline feature set. Similarly, the IndoBERT-enhanced scenario incorporated Sentiment_Indo as an additional feature. All numerical variables were normalized using Min-Max scaling to transform values into a comparable numerical range. Feature scaling is commonly employed in deep learning applications because it improves numerical stability and accelerates convergence during optimization [29], [30].

Subsequently, a sliding-window mechanism with a lookback period of 20 trading days was applied to transform the sequential data into supervised learning samples. Each input sequence contains observations from the previous twenty trading days, while the target corresponds to the closing price of the subsequent trading day. Sliding-window sequence generation is widely used in recurrent neural network forecasting models because it allows the network to learn temporal dependencies from historical observations [3], [31]. The dataset was divided chronologically into training and testing subsets using an 80:20 ratio. Random shuffling was intentionally avoided because temporal ordering must be preserved when modelling financial time-series data.

3.6 LSTM Model Development

The prediction model employed in this study is a stacked Long Short-Term Memory (LSTM) neural network. LSTM was selected because it was specifically developed to address the vanishing-gradient problem commonly encountered in conventional recurrent neural networks while maintaining the ability to learn long-term temporal dependencies [31]. These characteristics make LSTM particularly suitable for modelling sequential financial data that exhibit complex temporal relationships.

Previous studies consistently report that LSTM-based architectures achieve strong performance in stock market forecasting because they can capture nonlinear patterns and long-range dependencies within financial time-series data [2, 3, 26]. The proposed architecture consists of two stacked LSTM layers containing 64 and 32 hidden units, respectively. A dropout rate of 0.3 was applied after each LSTM layer to reduce overfitting. The output layer consists of a single neuron with linear activation representing the predicted closing price.

Model training was performed using the Adam optimizer and Mean Squared Error (MSE) as the loss function. The maximum number of training epochs was set to 100, with a batch size of 32.

EarlyStopping and ReduceLROnPlateau callbacks were employed to improve model generalization and stabilize convergence during training [29], [30]. As shown in Table 3, the LSTM model was configured using the selected architectural and training parameters applied in the experiment.

Table 3 LSTM Model Configuration

Parameter	Value
Model type	Stacked LSTM
Input features baseline	Open, High, Low, Close, Volume
Input features FinBERT scenario	Open, High, Low, Close, Volume, Sentiment_Fin
Input features IndoBERT scenario	Open, High, Low, Close, Volume, Sentiment_Indo
Lookback window	20 trading days
Train-test split	80:20 chronological split
First LSTM layer	64 units
Second LSTM layer	32 units
Dropout rate	0.3
Output layer	Dense (1 neuron)
Activation output	Linear
Optimizer	Adam
Loss function	MSE
Maximum epochs	100
Batch size	32
Regularization	EarlyStopping, ReduceLROnPlateau

3.7 Model Evaluation

Model performance was evaluated using five complementary metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), coefficient of determination (R^2), and Directional Accuracy. RMSE and MAE quantify prediction error magnitude, while MAPE expresses prediction error as a percentage relative to actual values. R^2 evaluates the proportion of variance explained by the prediction model. These metrics are widely adopted in stock forecasting research because they provide a comprehensive evaluation of regression performance [2], [3].

In addition to numerical prediction accuracy, Directional Accuracy was used to evaluate the ability of the model to correctly predict upward and downward price movements. This metric is particularly relevant in investment decision-making because market participants frequently prioritize prediction direction over exact price values [32]. The combination of multiple evaluation metrics enables assessment of sentiment integration from both statistical and practical perspectives, thereby providing a more comprehensive understanding of its contribution to stock prediction performance.

3.8 Statistical Significance Testing

The evaluation metrics presented in this study provide information regarding predictive accuracy; however, differences in metric values alone do not necessarily indicate statistically meaningful performance improvements. Therefore, statistical significance testing was conducted to determine whether the observed differences between the baseline LSTM model and sentiment-enhanced models were statistically significant.

This study employs the Wilcoxon Signed-Rank Test, a non-parametric statistical procedure commonly used to compare paired predictive models without requiring assumptions of normality [33]. The test evaluates whether the prediction errors generated by two competing models originate from the same distribution.

For each stock, absolute prediction errors generated by the baseline LSTM model were compared against those produced by the FinBERT-enhanced and IndoBERT-enhanced models. The null hypothesis states that there is no statistically significant difference between the prediction errors of the compared models. A significance level of $\alpha = 0.05$ was adopted. Therefore, p-values lower than 0.05 indicate statistically significant differences, whereas p-values greater than or equal to 0.05 indicate that observed performance differences may be attributed to random variation.

The inclusion of statistical significance testing provides additional evidence regarding whether sentiment integration contributes meaningful predictive improvements beyond conventional evaluation metrics such as RMSE, MAE, MAPE, R^2, and Directional Accuracy. The Wilcoxon Signed-Rank Test was selected because it does not require normality assumptions and is widely applied in machine learning performance comparison studies. Although forecast-specific procedures such as the Diebold-Mariano test are also available, Wilcoxon testing provides a robust and computationally efficient framework for assessing predictive performance differences in this study.

4 Results and Discussion

4.1 Overall Prediction Performance

Three prediction scenarios were evaluated in this study: the baseline LSTM model, FinBERT-enhanced LSTM, and IndoBERT-enhanced LSTM. The experiments were conducted on three Indonesian stocks, namely BBCA, BBRI, and BSDE. The results demonstrate that sentiment integration does not produce uniform improvements across all stocks. Instead, the impact of sentiment information appears to be stock-dependent and varies according to the sentiment model employed.

For BBCA, both sentiment-enhanced models produced higher prediction errors than the baseline model. The baseline LSTM achieved the lowest average prediction error (184.19), while FinBERT and IndoBERT generated average errors of 192.19 and 195.02, respectively. This finding suggests that sentiment information did not provide additional predictive value for BBCA during the observation period.

In contrast, the BBRI dataset showed positive effects from FinBERT integration. The average prediction error decreased from 118.86 in the baseline model to 112.97 after incorporating FinBERT sentiment features. However, IndoBERT produced a higher average error of 123.60, indicating that language-specific sentiment representations were less effective than financial-domain sentiment representations for this stock. For BSDE, FinBERT produced results similar to the baseline model, while IndoBERT achieved the lowest average prediction error (24.61), outperforming both the baseline model (25.92) and FinBERT (26.11). This result indicates that language-specific sentiment information may be more beneficial for stocks that receive relatively limited financial news coverage.

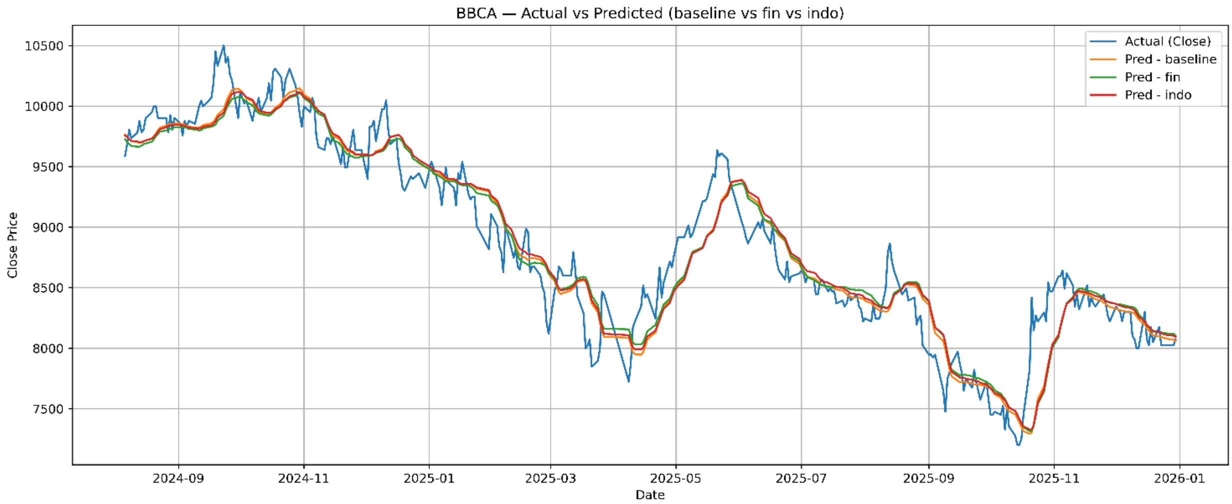

Figure 2 Actual and predicted closing prices for BBCA

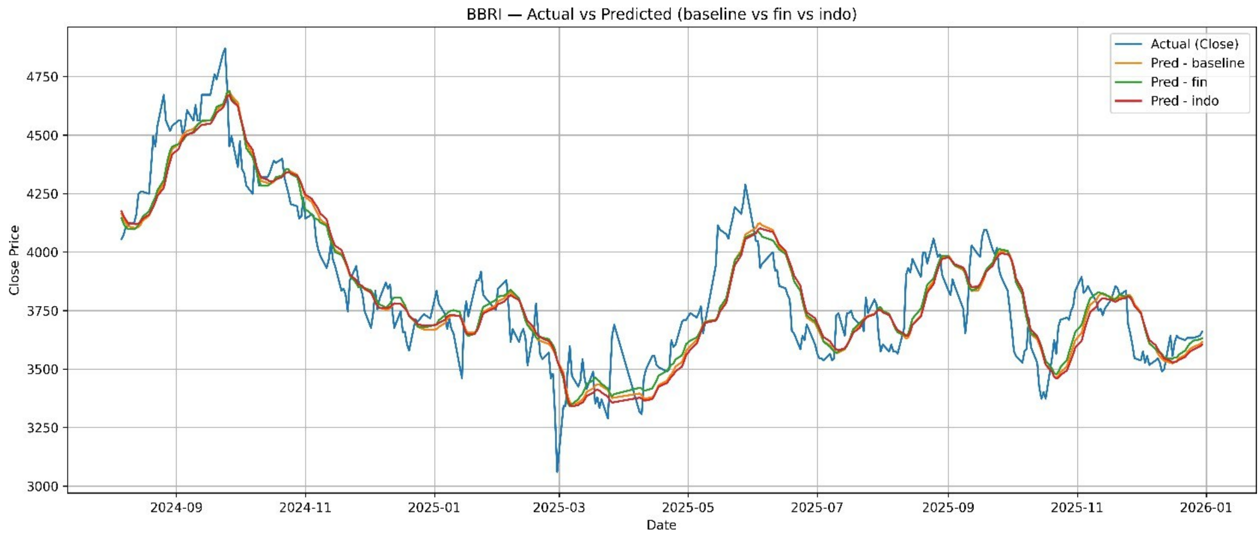

Figure 3 Actual and predicted closing prices for BBRI

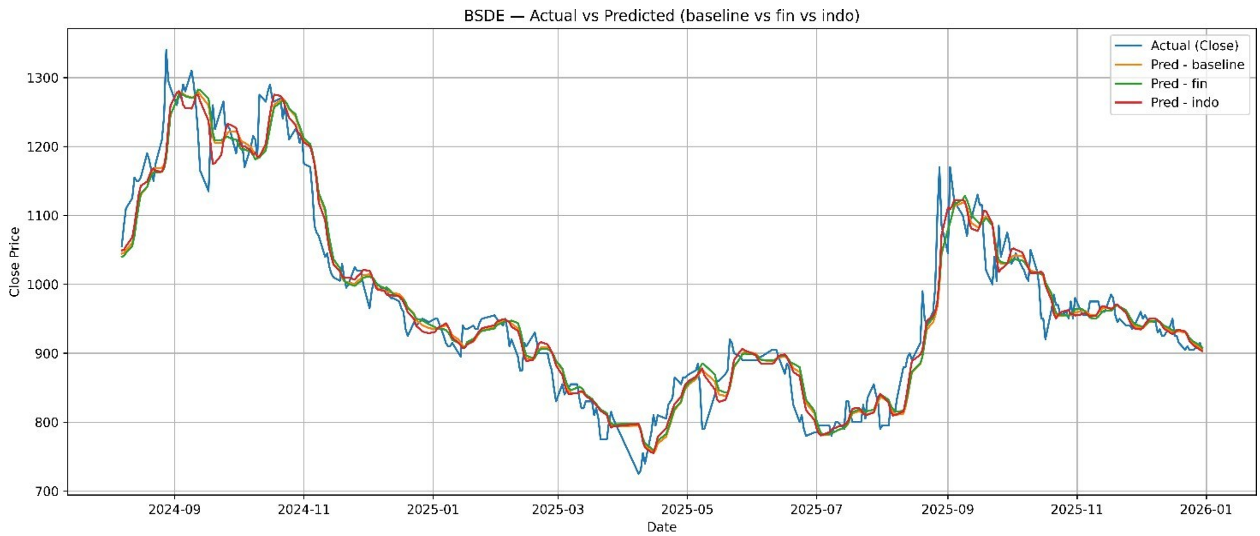

Figure 4 Actual and predicted closing prices for BSDE

Overall, the results indicate that sentiment integration can improve prediction accuracy under specific conditions, but the effectiveness of sentiment features depends on both stock characteristics and the sentiment extraction model. To provide a clearer visual comparison of forecasting behavior, Figures 2-4 present the actual closing prices and predicted values generated by the baseline, FinBERT-enhanced, and IndoBERT-enhanced models for each stock in the testing period. These figures illustrate that the major price trends, but the predictions are smoother than the actual series and tend to lag during sharp upward or downward movements.

4.2 Directional Accuracy Analysis

In addition to numerical prediction accuracy, this study evaluated the ability of the models to correctly predict price movement direction through Directional Accuracy (DA). The results show that Directional Accuracy remained close to the random-walk benchmark of 50% for most experimental scenarios. For BBCA, Directional Accuracy ranged between 47.48% and 48.37%. Similarly, BBRI achieved values between 50.15% and 50.74%. These results indicate that although the models were capable of approximating future prices, they were less effective in consistently predicting whether prices would increase or decrease on the following trading day. To further evaluate whether the observed Directional Accuracy values differed significantly from random guessing, a binomial test was performed using 50% as the benchmark probability. The resulting p-values for BBCA and BBRI were greater than 0.05 across all model configurations, indicating that the observed Directional Accuracy was statistically indistinguishable from random prediction.

For BSDE, Directional Accuracy values ranged from 40.18% to 43.15%, which are substantially below the 50% benchmark. The corresponding binomial test results were statistically significant ($p < 0.05$), suggesting that directional prediction performance was significantly different from random guessing; however, the difference reflects poorer-than-random directional forecasting performance. These findings support the reviewer's observation that improvements in numerical forecasting metrics do not necessarily translate into improvements in directional forecasting capability. Therefore, although sentiment-enhanced models may reduce prediction error in certain cases, they do not consistently improve market direction prediction.

Table 4 Directional Accuracy and Binomial Test Results

Stock	Model	Directional Accuracy (%)	Binomial p-value
BBCA	Baseline	48.37	0.586
BBCA	FinBERT	47.48	0.383
BBCA	IndoBERT	48.37	0.586
BBRI	Baseline	50.74	0.828
BBRI	FinBERT	50.74	0.828
BBRI	IndoBERT	50.15	1.000
BSDE	Baseline	41.67	0.0026
BSDE	FinBERT	43.15	0.0140
BSDE	IndoBERT	40.18	0.0004

4.3 Error Distribution Analysis

To further examine the distribution and variability of daily prediction errors beyond average error metrics, the distribution of absolute prediction errors was analyzed using boxplots. Figure 5 presents the absolute error distributions of the Baseline LSTM, FinBERT-enhanced LSTM, and IndoBERT-enhanced LSTM models for BBCA, BBRI, and BSDE. Lower median values and narrower interquartile ranges (IQR) indicate a more concentrated distribution of prediction errors, while a larger spread and more outliers suggest greater variability in model performance.

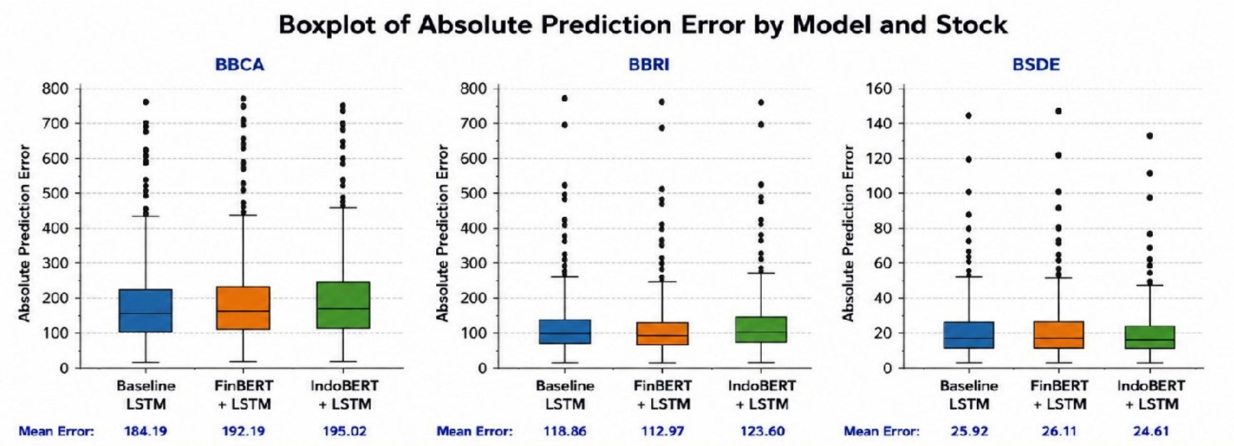

Figure 5 Distribution of absolute prediction errors across models and stocks

Figure 5 shows that the distribution of prediction errors varies across stocks and sentiment configurations. For BBCA, the baseline LSTM exhibits the lowest median error and a slightly narrower interquartile range compared with the sentiment-enhanced models, indicating a more concentrated distribution of daily prediction errors. In contrast, FinBERT achieves the lowest median error for BBRI, supporting the improvement observed in the MAE and RMSE results. For BSDE, IndoBERT produces the lowest median error and the most compact error distribution among the evaluated models. All models have outliers, but error distributions align with evaluation and Wilcoxon test results, supporting that sentiment integration's effectiveness depends on the stock, not universally beneficial.

4.4 Best-Performing Model Analysis

To facilitate a comprehensive comparison, the best-performing model for each stock was identified by considering the combined results of MAE, RMSE, MAPE, R^2, Directional Accuracy, and statistical significance testing. Rather than relying on a single evaluation metric, this approach provides a more balanced assessment of overall model effectiveness. The results are summarized in Table 5.

Table 5 Best-Performing Model by Stock

Stock	Best Model	MAE	RMSE	R^2	DA (%)	Statistical Result
BBCA	Baseline LSTM	184.19	252.82	0.9000	48.37	Better than sentiment models
BBRI	FinBERT + LSTM	112.97	143.43	0.8151	50.74	Significant improvement
BSDE	IndoBERT + LSTM	24.61	34.80	0.9409	40.18	Significant improvement

As shown in Table 5, no single model consistently outperformed the others across all stocks. The baseline LSTM achieved the best overall performance for BBCA, FinBERT-enhanced LSTM performed best for BBRI, and IndoBERT-enhanced LSTM achieved the highest performance for BSDE. These findings further support the conclusion that the effectiveness of sentiment integration is highly dependent on stock characteristics and the suitability of the sentiment representation model.

4.5 Statistical Significance Analysis

To determine whether the observed performance differences between models were statistically meaningful, the Wilcoxon Signed-Rank Test was performed on the prediction errors generated by each model. For BBCA, both FinBERT and IndoBERT produced statistically significant differences relative to the baseline model ($p < 0.001$). However, the average prediction errors of both sentiment-enhanced models were higher than those of the baseline model. Therefore, the significant differences observed for BBCA indicate performance degradation rather than improvement.

For BBRI, FinBERT significantly reduced prediction error compared with the baseline model ($p = 1.69 \times 10^{-9}$). This result suggests that financial-domain sentiment information contributed meaningful predictive value for BBRI. In contrast, IndoBERT produced significantly larger errors than the baseline model despite also yielding statistically significant differences.

For BSDE, FinBERT did not produce statistically significant changes relative to the baseline model ($p = 0.738$), indicating that the observed performance differences could be attributed to random variation. Conversely, IndoBERT significantly reduced prediction error ($p = 0.00015$), suggesting that Indonesian-language sentiment information contributed positively to prediction performance .

These significance results should be interpreted together with the magnitude and direction of the observed error differences. The Wilcoxon test evaluates whether paired daily absolute errors differ systematically within the available prediction run, but statistical significance does not by itself imply a large or economically meaningful improvement. For example, the MAE reductions observed for BBRI and BSDE are modest relative to the corresponding price scales. Accordingly, the findings support consistent within-run error differences rather than definitive evidence of practical gains.

The statistical analysis demonstrates that sentiment integration does not universally improve stock price prediction performance. Instead, its effectiveness depends on the interaction between stock characteristics, market conditions, and the sentiment representation model. The Wilcoxon Signed-Rank Test results are summarized in Table 6.

Table 6 Wilcoxon Signed-Rank Test Results

Stock	Comparison	p-value	Significant	Interpretation
BBCA	Baseline vs FinBERT	6.88×10^{-6}	Yes	FinBERT significantly worsened performance
BBCA	Baseline vs IndoBERT	6.11×10^{-17}	Yes	IndoBERT significantly worsened performance
BBRI	Baseline vs FinBERT	1.69×10^{-9}	Yes	FinBERT significantly improved performance
BBRI	Baseline vs IndoBERT	2.82×10^{-10}	Yes	IndoBERT significantly worsened performance
BSDE	Baseline vs FinBERT	0.738	No	No significant difference
BSDE	Baseline vs IndoBERT	0.00015	Yes	IndoBERT significantly improved performance

4.6 Discussion

The findings suggest that the usefulness of sentiment information depends on the characteristics of the underlying stock and the nature of the sentiment model. For BBRI, FinBERT produced lower prediction errors than both the baseline model and IndoBERT in the reported experiment. Because FinBERT was trained on financial corpora, it is capable of capturing finance-specific terminology and contextual relationships that may not be adequately represented by general-language models [10], [12]. Consequently, financial-domain sentiment information appears to provide meaningful signals for predicting price movements in highly visible banking stocks.

For BSDE, however, IndoBERT achieved the best overall performance. One possible explanation is that the Indonesian-language representation captured information that was not equally represented by the FinBERT pipeline. However, this interpretation remains tentative because the study did not directly quantify linguistic characteristics, financial terminology density, or translation-related semantic shifts in the articles. Accordingly, the observed improvement cannot be attributed exclusively to language specificity and may also be influenced by news sparsity, translation effects, model calibration, and zero imputation.

The interpretation of the BSDE result should nevertheless consider the substantially lower density of retrieved news. Only 555 unique article records were available for 1,705 target trading dates, whereas the remaining dates received a neutral value of zero. This sparse coverage creates a potential confounding factor because the apparent benefit of IndoBERT may reflect not only Indonesian linguistic representation, but also the interaction between infrequent sentiment signals, zero imputation, and the temporal dynamics of BSDE. Therefore, the present results do not establish that language specificity alone caused IndoBERT's improvement. Future research should conduct a sensitivity analysis restricted to dates with retrieved news and compare alternative approaches for representing dates without relevant articles.

A further limitation concerns daily sentiment aggregation. When several articles shared the same publication date, their continuous sentiment scores were averaged into a single daily value. Although this procedure provides a consistent daily feature, it can attenuate strong article-level signals; for example, an intensely negative article may be offset by a mildly positive article published on the same date. The model also did not include the number of articles per day as a separate feature. Future studies may examine weighted aggregation based on confidence, relevance, recency, or article volume.

The results also reveal an important limitation of sentiment-enhanced stock prediction. Although sentiment information occasionally improves numerical forecasting accuracy, it does not consistently improve directional forecasting performance. Directional Accuracy values remained close to the random-walk benchmark for BBCA and BBRI and below 50% for BSDE. This finding indicates that predicting exact price values and predicting market direction are distinct challenges that may require different modelling strategies.

Overall, the study demonstrates that sentiment integration can provide statistically significant improvements under specific circumstances. However, the results do not support the assumption that sentiment-enhanced models universally outperform traditional LSTM models. Instead, the effectiveness of sentiment integration appears to be conditional upon stock-specific characteristics and the suitability of the sentiment extraction model.

The testing period also includes a pronounced downward market movement from late 2024 to early 2025, which may reflect broader market uncertainty and macro-political transition effects in Indonesia. During such periods, stock prices may be influenced by factors beyond firm-specific news sentiment, including policy expectations, investor risk perception, and market-wide capital flows. This condition may partly explain why sentiment-enhanced models were unable to consistently improve directional prediction, despite showing lower numerical errors for certain stocks.

The model-level stability of these findings is also limited by the stochastic nature of LSTM training. Each stock–model configuration was trained once without a controlled and repeatedly evaluated random-seed protocol. Therefore, the Wilcoxon Signed-Rank results characterize paired daily error differences within the reported runs, but they do not quantify variability arising from random weight initialization, dropout, or optimization order. Future experiments should repeat each configuration across multiple fixed seeds and report the mean and standard deviation of all evaluation metrics.

5 Conclusion

This study conducted a comparative evaluation of sentiment-enhanced LSTM models for stock price prediction by integrating sentiment features derived from FinBERT and IndoBERT. Experiments were performed on three Indonesian stocks, namely BBCA, BBRI, and BSDE, using daily historical stock price data and sentiment information extracted from financial news. Model performance was assessed using multiple evaluation metrics, including RMSE, MAE, MAPE, R^2, Directional Accuracy, and statistical significance testing.

The experimental results indicate that incorporating sentiment features does not consistently improve numerical prediction accuracy compared with the baseline LSTM model that relies solely on historical market data. The effect of sentiment integration is stock-specific. For BBCA, the baseline LSTM produced better performance than both sentiment-enhanced models, indicating that sentiment features did not provide additional predictive value for this highly liquid banking stock. For BBRI, the FinBERT-enhanced model achieved lower prediction error than the baseline, and the Wilcoxon Signed-Rank Test confirmed that this difference was statistically significant. In contrast, IndoBERT did not improve BBRI prediction performance. For BSDE, the IndoBERT-enhanced model produced the lowest prediction error and showed statistically significant improvement over the baseline, while FinBERT did not provide a significant improvement.

The Directional Accuracy analysis further shows that sentiment integration did not consistently improve the ability of the models to predict upward or downward price movements. For BBCA and BBRI, Directional Accuracy remained close to the 50% random-walk benchmark, while for BSDE it was below 50%. These findings suggest that reducing numerical prediction error does not necessarily translate into better directional forecasting performance. Therefore, sentiment-enhanced LSTM models should not be assumed to provide universal benefits in practical stock movement prediction.

The comparative analysis demonstrates that neither FinBERT nor IndoBERT consistently outperforms the other across all stocks and evaluation metrics. FinBERT appears more useful when financial-domain sentiment provides relevant predictive signals, as observed in BBRI. Meanwhile, IndoBERT appears more effective in cases where Indonesian-language contextual sentiment is more informative, as observed in BSDE. These results indicate that the effectiveness of sentiment models is influenced by stock characteristics, linguistic context, news coverage, and market conditions.

This study contributes empirical evidence that sentiment integration in stock price prediction should be applied selectively rather than treated as a universally superior approach. As limitations, this study examined only three Indonesian stocks and relied solely on news-based sentiment features from publicly available financial news. Future research may extend this work by incorporating additional stocks, longer observation periods, social media sentiment, macroeconomic indicators, and more advanced forecasting architectures such as attention-based LSTM, GRU, or transformer-based time-series models. Further studies may also apply additional forecast comparison tests, such as the Diebold-Mariano test, to strengthen the evaluation of predictive performance differences.

Declaration of the Use of Generative AI

During the preparation of this manuscript, the authors used ChatGPT by OpenAI to support language refinement, grammatical review, and editorial clarity. The tool was not used to generate research data, conduct the experiments, calculate the reported results, fabricate references, or replace the authors' scientific interpretation. All AI-assisted text was critically reviewed and revised by the authors, who remain fully responsible for the accuracy, integrity, originality, and final content of the manuscript.

Bibliography

[1] S. A. Panchal, L. Ferdouse, and A. Sultana, "Comparative Analysis of ARIMA and LSTM Models for Stock Price Prediction," *in 2024 IEEE/ACIS 27th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, Beijing, China: IEEE, Jul. 2024, pp. 240–244. doi: 10.1109/SNPD61259.2024.10673919 .

[2] S. Sang and L. Li, "A Novel Variant of LSTM Stock Prediction Method Incorporating Attention Mechanism," *Mathematics*, vol. 12, no. 7, p. 945, Mar. 2024, doi: 10.3390/math12070945.

[3] K. Alam, M. H. Bhuiyan, I. U. Haque, M. F. Monir, and T. Ahmed, "Enhancing Stock Market Prediction: A Robust LSTM-DNN Model Analysis on 26 Real-Life Datasets," *IEEE Access*, vol. 12, pp. 122757–122768, 2024, doi: 10.1109/ACCESS.2024.3434524.

[4] M. Diqi, E. Utami, Kusrini, and F. W. Wibowo, "A Study on Methods, Challenges, and Future Directions of Generative Adversarial Networks in Stock Market Prediction," *in 2024 IEEE International Conference on Technology, Informatics, Management, Engineering and Environment (TIME-E)*, Bali, Indonesia: IEEE, Aug. 2024, pp. 148–153. doi: 10.1109/TIME-E62724.2024.10919813.

[5] B. H. C. and I. J. Jacob, "An Efficient Hybrid Model for Stock Market Price Prediction Using CNN-BiLSTM with Attention Mechanism and Sentimental Analysis*," Eng. Technol. Appl. Sci. Res*., vol. 15, no. 4, pp. 25565–25571, Aug. 2025, doi: 10.48084/etasr.11886.

[6] D. Vallarino, "An AI-Enhanced Forecasting Framework: Integrating LSTM and Transformer-Based Sentiment for Stock Price Prediction," i, vol. 4, no. 3, pp. 1–15, Sep. 2025, doi: 10.58567/jea04030001.

[7] T. H. Nguyen, K. Shirai, and J. Velcin, "Sentiment analysis on social media for stock movement prediction," *Expert Systems with Applications*, vol. 42, no. 24, pp. 9603–9611,

Dec. 2015, doi: 10.1016/j.eswa.2015.07.052.

[8] L. Bacco, L. Petrosino, D. Arganese, L. Vollero, M. Papi, and M. Merone, "Investigating Stock Prediction Using LSTM Networks and Sentiment Analysis of Tweets Under High Uncertainty: A Case Study of North American and European Banks," *IEEE Access*, vol. 12, pp. 122239–122248, 2024, doi: 10.1109/ACCESS.2024.3450311.

[9] W. Jun Gu et al., "Predicting Stock Prices with FinBERT-LSTM: Integrating News Sentiment Analysis," *in Proceedings of the 2024 8th International Conference on Cloud and Big Data Computing*, Oxford United Kingdom: ACM, Aug. 2024, pp. 67–72. doi: 10.1145/3694860.3694870.

[10] D. Araci, "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," *arXiv*, Aug. 2019, doi: 10.48550/arXiv.1908.10063.

[11] Z. Liu, D. Huang, K. Huang, Z. Li, and J. Zhao, "FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining," *in Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, Yokohama, Japan: International Joint Conferences on Artificial Intelligence Organization, Jul. 2020, pp. 4513–4519. doi: 10.24963/ijcai.2020/622.

[12] Y. Yang, M. C. S. UY, and A. Huang, "FinBERT: A Pretrained Language Model for Financial Communications," *arXiv*, Jul. 2020, doi: 10.48550/arXiv.2006.08097.

[13] T. Jiang and Q. Zeng, "Financial sentiment analysis using FinBERT with application in predicting stock movement," *arXiv*, Jun. 2025, doi: 10.48550/arXiv.2306.02136.

[14] B. Wilie et al., "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," *in Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Suzhou, China: Association for Computational Linguistics, 2020, pp. 843–857. doi: 10.18653/v1/2020.aacl-main.85.

[15] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," *in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 10660–10668. doi: 10.18653/v1/2021.emnlp-main.833.

[16] F. Indriani, R. A. Nugroho, M. R. Faisal, and D. Kartini, "Comparative Evaluation of IndoBERT, IndoBERTweet, and mBERT for Multilabel Student Feedback Classification," *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 8, no. 6, pp. 748–757, Dec. 2024, doi: 10.29207/resti.v8i6.6100.

[17] A. Maretta and A. Meiriza, "Aspect-Based Sentiment Analysis of Hospital Service Reviews Using Fine-Tuned IndoBERT," *JAIC*, vol. 9, no. 5, pp. 2541–2551, Oct. 2025, doi: 10.30871/jaic.v9i5.10765.

[18] T. Aprilah, D. R. I. M. Setiadi, and W. Herowati, "Enhancing Aspect-Based Sentiment Analysis via Hugging Face Fine-Tuned IndoBERT," *JAIC*, vol. 9, no. 6, pp. 3821–3830, Dec. 2025, doi: 10.30871/jaic.v9i6.11409.

[19] J. Zhang, H. Liu, W. Bai, and X. Li, "A hybrid approach of wavelet transform, ARIMA and LSTM model for the share price index futures forecasting," *The North American Journal of Economics and Finance*, vol. 69, p. 102022, Jan. 2024, doi: 10.1016/j.najef.2023.102022.

[20] D. M. Q. Nelson, A. C. M. Pereira, and R. A. De Oliveira, "Stock market's price movement prediction with LSTM neural networks," *in 2017 International Joint Conference on Neural Networks (IJCNN)*, Anchorage, AK, USA: IEEE, May 2017, pp. 1419–1426. doi: 10.1109/IJCNN.2017.7966019.

[21] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," *European Journal of Operational Research*, vol. 270, no. 2, pp. 654–669, Oct. 2018, doi: 10.1016/j.ejor.2017.11.054.

[22] M. S. Zakka and E. Emigawaty, "Stock Price Prediction Using Deep Learning (LSTM) with a Recursive Approach," *JAIC*, vol. 9, no. 5, pp. 2468–2477, Oct. 2025, doi: 10.30871/jaic.v9i5.10514.

[23] T. H. Nguyen and K. Shirai, "Topic Modeling based Sentiment Analysis on Social Media for Stock Market Prediction," *in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Beijing, China: Association for Computational Linguistics, 2015, pp. 1354–1364. doi: 10.3115/v1/P15-1131.

[24] A. S. Girsang and Stanley, "Hybrid LSTM and GRU for Cryptocurrency Price Forecasting Based on Social Network Sentiment Analysis Using FinBERT," *IEEE Access*, vol. 11, pp. 120530–120540, 2023, doi: 10.1109/ACCESS.2023.3324535.

[25] J. W. Creswell, "Research design: qualitative, quantitative, and mixed methods approaches", 4. ed. Los Angeles, Calif.: SAGE, 2014.

[26] J. Zhang, L. Ye, and Y. Lai, "Stock Price Prediction Using CNN-BiLSTM-Attention Model," *Mathematics*, vol. 11, no. 9, p. 1985, Apr. 2023, doi: 10.3390/math11091985.

[27] B. Liu, "Synthesis Lectures on Human Language Technologies: Sentiment Analysis and Opinion Mining. in Synthesis Lectures on Human Language Technologies Series," San Rafael: Morgan & Claypool Publishers, 2012.

[28] L. Tunstall, L. von Werra, and T. Wolf, "Natural language processing with transformers: building language applications with Hugging Face, First edition," Sebastopol, CA: O'Reilly Media, 2022.

[29] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning. in Adaptive computation and machine learning," Cambridge, Massachusetts: The MIT Press, 2016.

[30] S. Raschka and V. Mirjalili, "Python machine learning: machine learning and deep learning with Python, scikit-learn, and TensorFlow 2, 3rd ed," Birmingham: Packt Publishing, 2020.

[31] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[32] F. Provost and T. Fawcett, "Data science for business: what you need to know about data mining and data-analytic thinking, First edition," Beijing Cambridge Farnkam Köln Sebastopol Tokyo: O'Reilly, 2013.

[33] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, vol. 1, no. 6, p. 80, Dec. 1945, doi: 10.2307/3001968.