JURNAL INFORMATIKA, Vol.12 No.1 April 2025, Page 1-9
ISSN (Print) 2355-6579 | ISSN (Online) 2528-2247
DOI: https://doi.org/10.31294/inf.v12i1.22299

Predicting Stock Price Movements with Technical, Fundamental, and Sentiment Analysis Using the LSTM Model

Muhammad Ighfar Saputra[1], Erna Nurmawati[2*], Rayhan Abyasa[3]

[1,2,3] Politeknik Statistika STIS
Otto Iskandardinata Street Number 64C, Jakarta 13330, Indonesia

Corresponding e-mail: erna.nurmawati@stis.ac.id

Submission:	Revision:	Acceptance:	Available Online:
07-06-2024	06-12-2024	06-02-2025	04-03-2025

Abstract

The challenge of minimizing risk and maximizing profit is what traders in the stock market have been endeavoring to solve for years. Stock prices typically exhibit the characteristic of volatility, influenced by various factors and necessitate a substantial amount of data to identify patterns in price movements. Considering the significant data requirements and the rapid advancement of big data and artificial intelligence, the LSTM (Long-Short Term Memory) model stands as a suitable approach for utilization in Deep Learning. The independent variables employed encompass technical indicator variables, currency exchange rates, interest rates, the Jakarta Composite Index (IHSG), and sentiment data extracted from Twitter tweets. The results indicate that sentiment analysis using the IndoBERT model achieved an accuracy of 0.69, while LSTM analysis produced the model with the smallest error for the fourth (4th) combination of variables, comprising closing price, technical indicators, IHSG, exchange rate, and Twitter sentiment, as well as the twelfth (12th) combination of variables, encompassing closing price, technical indicators, and IHSG. These combinations yielded average RMSE errors of 1.765E-04 and 1.978E-04, respectively. Hyperparameter optimization is done to six hyperparameter, number of unit layer, dropout rate, learning rate, batch size, optimizer, and timestamps. Following hyperparameter optimization, the best-identified model was the fourth (4th) combination of variables, yielding a minimal error of 7.580E-05 and an RMSE of 332.66 in the evaluation of test data.

Keywords: Stock, LTSM, Sentiment analysis

1. Introduction

The challenge of minimizing risk and maximizing profits by predicting future price movements has long been the pursuit of traders in the stock market (Ardyanta & Sari, 2021). Stock price movements are recognized for their volatility and non-linearity, making precise stock price forecasting a formidable task due to its susceptibility to various influencing factors. Hence, the development of a model capable of predicting stock price movements presents an intriguing challenge.

In analyzing stock price movements, investors commonly employ two techniques. The first technique is technical analysis, which involves predicting stock price movements using historical data such as opening and closing prices, transaction volume, average stock price, and more. The second technique, known as fundamental analysis, utilizes qualitative and quantitative data derived from company profiles, financial conditions, market conditions, politics, business, and economic conditions (Huang et al., 2021). However, with the advancement of information technology and social media, a third technique has also emerged, known as sentiment analysis (Derakhshan & Beigy, 2019). Sentiment is defined as an individual's perspective or opinion regarding information (Aggarwal, 2022).

Given the numerous factors that influence stock price movements, it is important to combine analytical techniques, such as technical analysis, with both sentiment and fundamental analysis when predicting stock price movements (Ardyanta & Sari, 2021). Research (Ardyanta & Sari, 2021) indicates that incorporating fundamental variables, such as exchange rates and foreign stock price indexes, can enhance prediction accuracy. Conversely, studies (Wu et al., 2022) and (Mndawe et al., 2022) demonstrate that employing sentiment variables can increase accuracy and underscore the significant role of sentiment features in predicting stock closing prices. Additionally, research (Wahyuddin, 2022) provides evidence that utilizing sentiment variables can reduce the Root Mean Square Error (RMSE) by 39.55%, highlighting the significance of sentiment features in stock price analysis.

This study aims to integrate three analytical techniques, necessitating the utilization of indicators for each analysis method. Technical analysis will employ commonly used technical indicators, as demonstrated in studies (Kumbure et al., 2022) and (Peng et al., 2021). However, identifying the indicators that exert the most influence on the model remains uncertain, necessitating the implementation of feature selection techniques. For fundamental analysis, the study will incorporate fundamental indicator variables explored in research (Ardyanta & Sari, 2021) and (Bhandari et al., 2022), encompassing stock price indexes, interest rates, and exchange rates. In contrast, sentiment analysis will encompass sentiment variables derived from Twitter sentiment analysis, related to the methodology employed in research (Mndawe et al., 2022).

Various types of research endeavor to model stock price movements using a spectrum of approaches, spanning from statistical-based models like Autoregressive Integrated Moving Average (ARIMA) to the incorporation of machine learning and deep learning techniques. Given the swift advancement of big data and text mining, which has recently gained prominence in predicting stock prices (Wu et al., 2022), coupled with the substantial data requirements, the application of deep learning proves to be highly fitting. In research (Kumbure et al., 2022), an examination of approximately 150 pieces of literature concerning machine learning applications in predicting stock movements, spanning from 2000 to 2021, reveals that only 15 of these studies incorporate sentiment analysis and text mining originating from sources such as news, emails, and social media. Notably, these studies were published within the timeframe of 2016 to 2021.

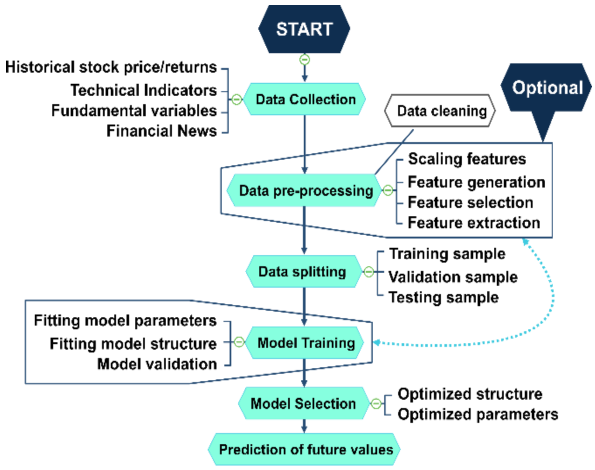
The model employed in this study is a deep learning model known as the Long-Short Term Memory (LSTM). LSTM is an advancement of the Recurrent Neural Network (RNN) designed to analyze sequential data and address issues of exploding and vanishing gradients inherent in traditional(Qu et al., 2019). LSTM stands out as superior to all other models, albeit the margin of difference might not be substantial due to the limited dataset. However, with an expansion of data to around 2000 days, LSTM would likely emerge as the clear winner by a significant margin (Mesquita et al., 2020). LSTM and Hybrid LSTM models demonstrate superiority in predicting future stock prices. Meanwhile, CNN and Hybrid models are particularly effective in excelling when used to predict stock trends (Shah et al., 2022).

Moreover, the process of model development entails hyperparameter optimization, which involves refining the structure and parameters of a model (Kumbure et al., 2022). This crucial step is essential for enhancing model performance (Yang & Shami, 2020). Subsequently, the model will be leveraged for the prediction of stock prices within a maximum horizon of 5 days, following a methodology akin to that of research conducted in (Mndawe et al., 2022) and (Attanasio et al., 2019).

The focal point of this research will be the shares of one of Indonesia's largest banks, namely Bank Central Asia (BCA). Bank BCA made its initial public offering in May 2000, with a total of 123,275,050,000 outstanding shares (as of December 2021). Furthermore, it stands as Indonesia's bank with the highest market capitalization, amounting to 1,070.92 trillion (as of April 5, 2023) (Asia, 2022; Bhandari et al., 2022). Given its extensive operational history and significant market capitalization, BCA is anticipated to possess an extensive historical dataset and facilitate the ease of gathering tweets related to BCA shares for subsequent sentiment analysis.

2. Research Methods

The research methodology's flow is depicted in the following workflow.



Source: (Wahyuddin, 2022)

Figure 1. Stock Price Prediction Flow

2.1 Data Collection

The utilized data encompasses the historical daily stock price records of BCA bank and the Jakarta Composite Index, covering the interval from January 19, 2005, to August 30, 2022. These datasets were procured from the Yahoo Finance website (Yahoo, 2022), comprising a total of 4,361 rows. The research incorporates tweet data from the Twitter social media platform pertaining to BCA bank shares. Additionally, the research encompasses data concerning the rupiah exchange rate to the US dollar, sourced from Investing.com (Investing. com, 2022). Furthermore, data regarding Indonesian interest rates is included, obtained from the official website of Bank Indonesia (Indonesia, 2022).

2.2 Variable Creation

2.2.1 Technical Indicator

The data utilized for computing technical indicators is derived from historical price data, including opening price, closing price, highest price, lowest price, and transaction volume. These data points are subsequently processed using specific formulas, generating numerical values indicative of a stock's movement tendency (Ardyanta & Sari, 2021). In this study, the chosen set of technical indicators comprises 12 indicators that have seen widespread selection in prior literature (Kumbure et al., 2022; Peng et al., 2021), and have been employed in various related studies (Ardyanta & Sari, 2021; Sivapurapu, 2020; Yang et al., 2020). The formulas corresponding to these 12 technical indicators are detailed in Table 1 below.

Table 1. Technical Indicator Formulas

Technical Indicator	Formulas
SMA (Simple 10-day Moving Average)	$\frac{(C_t + C_{t-1} + C_{t-2} + \cdots + C_{t-10}}{10}$
WMA (Weighted 10-day Moving Average)	$\frac{(10C_t + 9C_{t-1} + 8C_{t-2} + \cdots + C_{t-10})}{10 + 9 + \cdots + 1}$
EMA (Exponential 10-day Moving Average)	$(C_t \times \alpha) + EMA_{t-1} \times (1 - \alpha)$
Momentum 10-day	$C_t - C_{t-10}$
StoK (Stochastic K%)	$\frac{C_t - LL_{t-14}}{HH_{t-14} - LL_{t-14}} \times 100$
StoD (Stochastic D%)	$\frac{\sum_{i=0}^{2} K_{t-i}\%}{3}$
RSI (Relative Strength Index)	$100 - \frac{100}{1 + \left(\frac{\sum_0^{13} Up_{t-i}}{14}\right) + \left(\frac{\sum_0^{13} Dw_{t-i}}{14}\right)}$
MACD (Moving Average Convergence Divergence)	$EMA_{12}(t) - EMA_{26}(t)$
ADO (A/D Oscillator)	$\frac{H_t - C_{t-1}}{H_t - L_t}$
CCI (Commodity. Channel Index)	$\frac{M_t - SM_t}{0.015D_t}$
LWillR (Larry William'sR%)	$\frac{HH_{t-14} - C_t}{HH_{t-14} - LL_{14}}$
MFI (Money Flow Index)	$100 - \frac{100}{1 + \frac{14 - Day\ Positive\ MF}{14 - Day\ Negative\ MF}}$

Source: (Kumbure et al., 2022; Peng et al., 2021)

2.2.2 Fundamental Indicator

Fundamental indicators that will be utilized to depict economic conditions encompass the rupiah exchange rate against the US Dollar, Indonesia's interest rate, and the Jakarta Composite Index (IHSG). Research (Ardyanta & Sari, 2021) establishes that incorporating exchange rates and stock indices can augment model accuracy.

In this study, sentiment analysis will be conducted following the methodology outlined in research (Mndawe et al., 2022). Twitter data was collected using the snscrape library (pypi, 2023), followed by data cleaning procedures. These processes encompassed eliminating duplicates, converting all text to lowercase, removing URLs, emojis, usernames, dates, spaces, and hashtags (#), and ultimately standardizing slang words or abbreviations.

Subsequently, the data will undergo manual labeling, where each datum will receive a label of "1" if the tweet is positive, "0" if neutral, and "-1" if negative. This labeling process will be executed on a randomly selected subset of 4000 data points.

The final phase involves training the sentiment model using the Bidirectional Encoder Representations from Transformers (BERT) pre-trained language model (Turchin et al., 2023), employing the labeled data. The trained model will subsequently be employed to classify the entirety of the tweet data.

Tweets originating from the same day will be quantified as a percentage of positive tweets in relation to the total number of both positive and negative tweets on that specific day. Neutral tweets are disregarded since they do not exert any influence on stock prices. Tweets falling on days when the stock market is closed will be incorporated into the subsequent day's data, following the approach outlined in research (Ardyanta & Sari, 2021).

To assess the outcomes of sentiment classification, a confusion matrix is employed to categorize three classes, wherein "tp" represents true positives, "fp" signifies false positives, "fn" represents false negatives, and "tn" denotes true negatives. Precision, recall, accuracy, and F1-score are commonly used metrics to evaluate prediction results, with the corresponding formulas as follows:

$$\begin{cases} precision = \frac{tp}{tp+fp}, \\ recall = \frac{tp}{tp+fn}, \\ accuracy = \frac{tp+tn}{tp+tn+fp+fn}, \\ F1 - Score = \frac{2*precision*recall}{precision+recall} \end{cases} \quad (1)$$

Tweets that are on the same day will be calculated as a percentage of the number of positive tweets against the total number of positive and negative tweets on that day. Neutral tweets are not used because they do not have any impact on the stock price. Tweets that are on the day the stock market is closed will be included the following day when the stock market reopens as applied to research (Ardyanta & Sari, 2021).

$$Sentimen\ skor\ harian = \frac{positif}{positif+negatif} \tag{2}$$

2.3 Feature Selection

The feature selection algorithm employed in this study is the Genetic Algorithm (GA). GA is a metaheuristic and stochastic optimization algorithm inspired by natural evolutionary processes (Katoch et al., 2021). The procedures and operators within GA imitate genetic and evolutionary principles found in nature, including crossover and mutation.

In this study, GA will amalgamate each technical indicator and assess their performance using a fitness function predicated on Root Mean Square Error (RMSE). GA will optimize the fitness values, ultimately providing the optimal ranking for the subset of technical indicators. Research (Gamarra A & Quintero M, 2013) demonstrates that models employing feature selection through GA exhibit superior performance compared to conventional models.

2.4 Sharing and Standardization of Data

The entire dataset is divided into two segments: training data and testing data. The training data will serve to train the model, encompassing data from January 19, 2005, to August 31, 2022. Afterwards, the testing data will be utilized to evaluate the model, incorporating data from September 1 to November 25, 2022 (note that the data range might alter over time). Data testing will only be used when modeling has been completed for estimating errors from model predictions and should not be used during the model training process to prevent Data Leak from occurring (Jiang et al., 2020).

Moreover, prior to commencing model training, all data will undergo standardization using the Min-Max Scaler. The Min-Max Scaler is a method for transforming data, which, in this study, will alter the data range for each variable to a scale of 0 to 1.

As highlighted in research (Ardyanta & Sari, 2021), the significance of amalgamating analysis techniques like technical analysis with both sentiment and fundamental analysis is underscored. Therefore, this study will incorporate a fusion of techniques such as technical and fundamental analysis, technical analysis with sentiment analysis, or even all three combined. Each approach will be represented through technical, fundamental, and sentiment variables. Consequently, a synthesis of variables from each analysis technique will be undertaken, with the stipulation that closing prices and technical indicators will consistently feature in the model.

Table 2. Looking for the best combination of variables.

Variable Combination	Closing Price	Technical Indicator	IHSG	Exchange Rate	Interest Rate	Twitter Sentiment
KomVar1	✓	✓	✓	✓	✓	✓
KomVar2	✓	✓	×	✓	✓	✓
KomVar3	✓	✓	✓	×	✓	✓
KomVar4	✓	✓	✓	✓	×	✓
KomVar5	✓	✓	✓	✓	✓	×
KomVar6	✓	✓	✓	✓	×	×
KomVar7	✓	✓	✓	×	✓	×
KomVar8	✓	✓	✓	×	×	✓
KomVar9	✓	✓	×	✓	✓	×
KomVar10	✓	✓	×	✓	×	✓
KomVar11	✓	✓	×	×	✓	✓
KomVar12	✓	✓	✓	×	×	×
KomVar13	✓	✓	×	✓	×	×
KomVar14	✓	✓	×	×	✓	×
KomVar15	✓	✓	×	×	×	✓
KomVar16	✓	✓	×	×	×	×

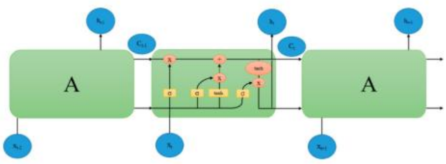
Source: Research Process

Table *2* explains the combination of variable for which the best combination will be sought. Closing price and technical indicator are used in every KomVar and other variable is used in turn for every KomVar. One KomVar will be selected as the best combination and will be use in the model.

2.5 Modelling

2.5.1 Long-Short Term Memory (LSTM) Model

Hochreiter and Schmidhuber introduced the LSTM model in 1997 (Yu et al., 2019). The LSTM model consists of three gate networks: the input gate, the forget gate, and the output gate (Qu et al., 2019). The input gate is responsible for deciding whether to introduce new information into the cell. The forget gate is responsible for determining which information to discard or retain to mitigate the vanishing gradient problem. Lastly, the output gate governs the quantity of information to be generated (Wu et al., 2022).



Source: Research Process

Figure 2. LSTM Model Flow

2.5.2 Validation

The validation approach employed in this study is Walk Forward Validation, a method frequently utilized in the analysis of time series data (Kumbure et al., 2022). This evaluation process involves consecutive, overlapping partitions in the form of training, validation, and testing (Baranochnikov & Ślepaczuk, 2022), and is employed to assess the model's performance. The chosen evaluation metric is an error-based measure, specifically the Root Mean Squared Error (RMSE), as extensively adopted in prior research (Kumbure et al., 2022).

2.5.3 Hyperparameter Optimization

In accordance with research (Yu et al., 2019), while the LSTM model is frequently employed as an effective approach in time series data analysis, there exist certain drawbacks associated with its utilization. Firstly, neural network models, including LSTM, tend to be challenging to interpret due to their intricate computational processes. Secondly, akin to other neural network models, the LSTM model encompasses numerous parameters that researchers must adjust, including layer count, neurons per layer, time lag, and more. However, due to time and computational constraints, identifying an optimal parameter combination that yields superior performance can be complex, thus necessitating the undertaking of hyperparameter optimization.

2.6 Evaluating Models

The optimized model is subsequently employed for predicting stock prices on the testing data in order to derive an estimated error value.

3. Results and Discussion

3.1 Formulation of Technical Indicators & Feature Selection

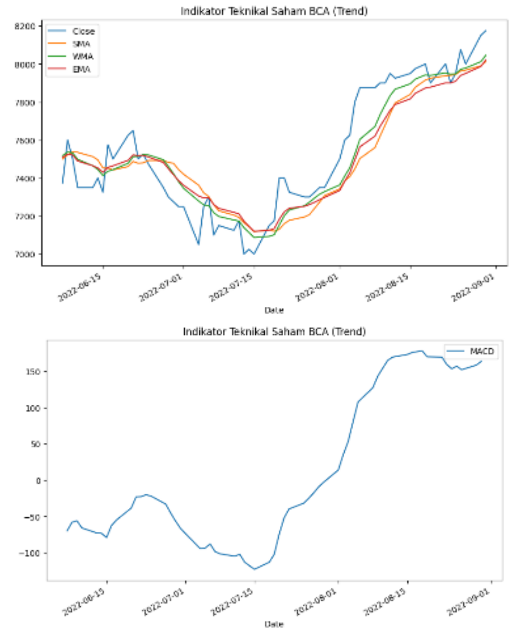
During the stage of formulating technical indicators, researchers utilized the pandas_ta library to calculate 12 technical indicators by employing stock price data, encompassing open, close, high, low, and volume prices.

Table 3. Descriptive Statistics of Stock Price History

Indicator	*Open*	*Close*	*High*	*Low*	*Volume*
Number of Data	4479	4479	4479	4479	4479
Average	2710.77	2710.34	2737.18	2682.20	1.09 e+08
Standard Deviation	2246.58	2235.31	2265.27	2225.59	1.33 e+08
Smallest Value	177.50	177.50	180.00	177.50	0
First Quantile (25%)	700.00	700.00	705.00	685.00	4.83 e+07
Median (50%)	2050.00	2055.00	2070.00	2030.00	7.28 e+07
Third Quantile (75%)	4460.00	4487.50	4530.00	4450.00	1.19 e+08
Biggest Value	8300.00	8200.00	8300.00	8075.00	1.95 e+09

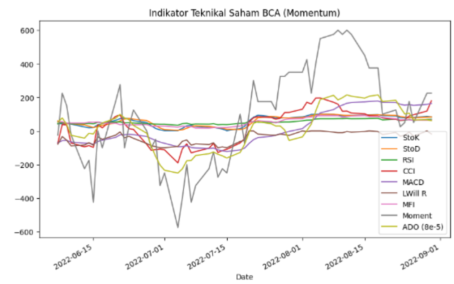
Source: Research Process

Broadly, these technical indicators can be categorized into two types: trend indicators, which signify the trend or directional movement of stocks, and momentum indicators, which highlight momentum or potential turning points within stock price movements.



Source: Research Process

Figure 3. Trending Technical Indicators



Source: Research Process

Figure 4. Momentum Technical Indicators

A feature selection process for technical variables is executed utilizing a fusion of the LSTM-GA algorithm to rank various combinations of technical variables, aiming to identify the combinations resulting in the lowest error and thereby determining the optimal set of technical variables. Here are the top 3 outcomes in terms of combinations of technical variables that exhibit the smallest errors:

According to the findings presented in Table 4, the combination of technical variables yielding the most optimal fitness value consists of SMA, StoD, RSI, MACD, ADO, CCI, and MFI variables. Furthermore, these technical variables are combined with variables originating from other analysis techniques, specifically fundamental and sentiment variables.

Table 4. Ranking Feature Selection Technical Variables

Ranking	Technical Variables	Fitness Value
1	SMA, StoD, RSI, MACD, ADO, CCI, MFI	0.0067
2	StoK, RSI, ADO, CCI	0.0070
3	EMA, Moment, StoK, RSI, LWill R, MFI	0.0072

Source: Research Process

3.2 Sentiment analysis

The acquisition of sentiment data from the Twitter social media platform involves utilizing the snscrape library for scraping. This scraping process entails collecting all Twitter tweets that include the phrases "BCA shares" and "BBCA" (the stock code for BCA bank) posted prior to August 31, 2022. The selection of the keywords "BCA shares" and "BBCA" was made due to their direct association with BCA bank shares, with the expectation that the acquired tweets would revolve around stock discussions. This procedure resulted in the collection of 70,225 Twitter tweets.

Table 5. Total Scraping Twitter

Keyword	Number of Tweets
Saham BCA	19,040
BBCA	51,185
Total	**70,225**

Source: Research Process

The preprocessing stage involves eliminating duplicates, refining sentences, and manually labeling data intended for model training input. The cleaning process entails converting all letters to lowercase, removing URLs, emojis, usernames, dates, spaces, and hashtags (#). This step is undertaken to eliminate irrelevant characters that could potentially impact the input data for the model.

The IndoBERT model was developed and trained using 4,000 manually labeled data samples. The dataset was initially divided into a 70% proportion for training data and a 30% proportion for test data. Following the training phase, the model underwent testing using the designated test data to perform tweet classification. This sentiment classification process yielded the following confusion matrix.

Table 6. Confusion Matrix Results of Sentiment Classification

		Negative	Neutral	Positive
True Sentiment	Negative	91	20	20
	Neutral	15	238	77
	Positive	32	88	223
		Predicted Sentiment		

Source: Research Process

Drawing from the confusion matrix, the sentiment analysis model's performance is discernible through the accuracy, recall, precision, and F1-Score values, which are displayed in the subsequent table.

Table 7. Precision, Recall, F1-Score, Support

	Precision	*Recall*	*F1-Score*	*Support*
Positive	0.66	0.69	0.68	131
Neutral	0.69	0.72	0.70	330
Negative	0.70	0.65	0.67	343
Accuracy		0.69		804

Source: Research Process

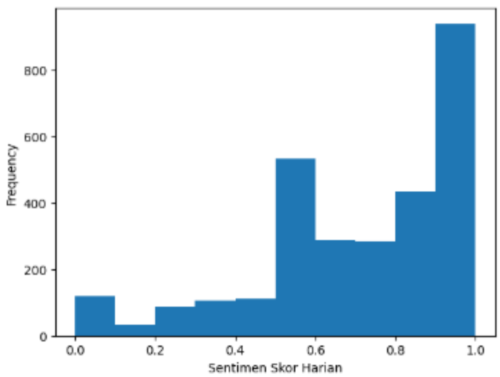
Next, the calculation of the daily sentiment score involves associating the tweet sentiments with the corresponding day. This entails determining the percentage of positive tweets in relation to the overall count of both positive and negative tweets that occurred on that specific day.

Table 8. Daily Sentiment Calculation Results

Date	Persentation
2022-08-24	0.785714
2022-08-25	1
2022-08-26	0.636364
2022-08-29	0.636364
2022-08-30	0.857143

Source: Research Process

The computation of this daily sentiment score involves utilizing formula (2), whereby a score approaching 1 signifies a considerable proportion of positive tweets in relation to the combined total of positive and negative tweets. Conversely, a score closer to 0 indicates a relatively smaller number of positive tweets compared to the overall count of both positive and negative tweets. A score of 0.5 is assigned when the quantity of positive and negative tweets is identical, or when no tweets were posted on that particular day. The subsequent histogram displays the distribution of daily sentiment scores:



Source: Research Process

Figure 5. Daily Score Sentiment Histogram

The histogram reveals that the majority of sentiment score data, accounting for 73.74% of the daily scores, indicates a positive trend (greater than 0.5).

3.3 Modeling and Combination of Analysis Techniques

Modeling is conducted for every combination of technical, fundamental, and sentiment variables. The LSTM model is configured with hyperparameters drawn from research (Peng et al., 2021), featuring 150 units in the hidden layer, a tanh activation function, Adam optimizer, and 5 units in the output layer. Each variable combination undergoes a 10-fold walk forward validation procedure.

Table 9. RMSE 10-Fold Walk Forward Validation (In E-04)

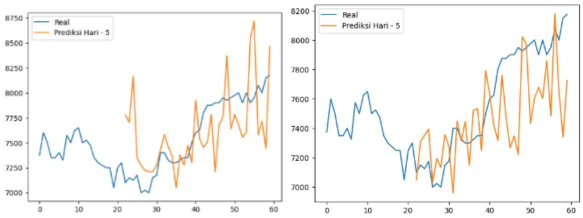
Variable Combination	Eror RMSE			
	Min	Mean	Max	Range
KomVar1	1.066	2.617	5.004	3.035 ± 1.969
KomVar2	0.608	2.545	6.796	3.702 ± 3.094
KomVar3	0.750	2.453	4.579	2.665 ± 1.915
KomVar4	**0.280**	1.978	3.739	**2.010** ± 1.729
KomVar5	0.423	2.147	5.310	2.867 ± 2.443
KomVar6	1.030	2.397	6.719	3.875 ± 2.844
KomVar7	0.535	2.454	6.809	3.672 ± 3.137
KomVar8	0.641	2.036	3.950	2.296 ± 1.654
KomVar9	0.783	2.805	7.238	4.011 ± 3.227
KomVar10	0.596	3.051	5.938	3.267 ± 2.671
KomVar11	0.597	2.106	4.348	2.473 ± 1.875
KomVar12	0.505	**1.765**	**3.519**	2.012 ± **1.507**
KomVar13	0.460	2.740	6.391	3.426 ± 2.965
KomVar14	0.910	2.342	6.192	3.551 ± 2.641
KomVar15	0.458	2.369	4.619	2.539 ± 2.080
KomVar16	0.453	1.864	5.456	2.955 ± 2.502

Source: Research Process

From Table 9 can be seen that from the 10-fold walk forward validation, there are two best combinations of variables namely KomVar4 and KomVar12. The combination of variables 4 produces a minimum RMSE and the smallest mean value compared to other variable combinations. On the other hand, the variable combination 12 has the smallest average RMSE error and the smallest range, so the variable combination 12 is stable.

After the model has been trained, the KomVar4 and KomVar12 models are used to predict stock prices for the next 5 days using test data, the results can be seen in the following figure.

From Table 9 can be seen that from the 10-fold walk



Source: Research Process

Figure 6. Stock Price Prediction Results for the Next 5 Days (KomVar4 (Left), KomVar12 (Right))

Drawing insights from the depicted graph above, the prediction outcomes for the fifth day from the KomVar4 model exhibit considerable volatility and deviate significantly from the actual value. To evaluate the model using the RMSE metric, the predictions are replicated 10 times to ascertain the average RMSE value. The resulting assessment yields an RMSE value of 390.85 for the KomVar4 model.

Similarly, considering the illustration above, the fifth-day predictions generated by the KomVar12 model exhibit fluctuations and a relatively diminished proximity to the actual value, although these deviations are not as volatile as those observed with KomVar4. Evaluating the model through the RMSE criterion, the prediction process is reiterated 10 times to determine the average RMSE value. This assessment results in an RMSE value of 366.53 for the KomVar12 model, which is smaller than that of KomVar4.

3.4 Hyperparameter Optimization

For the two optimal variable combinations, namely variable combination 4 and variable combination 12, a Hyperparameter Optimization process will be conducted using Grid Search to identify the most suitable hyperparameters. The hyperparameters under consideration are presented in the subsequent table:

Table 10. Hyperparameter To Be Combined

Hyperparameter	Interval
Jumlah unit layer	100, 150, 200, 300
Dropout Rate	0, 0.1, 0.3, 0.5
Learning Rate	0.1, 0.01, 0.0001
Ukuran batch	8, 16, 32, 64
Optimizer	Adam, RMSprop
Timestamps	10, 20, 30

Source: Research Process

Due to the presence of 1152 hyperparameter combinations and 2 variable combinations, which will be evaluated using a 5-fold walk forward validation, a total of 11,520 models will be trained through the grid search process. Presented below are the top 10 hyperparameters for each variable combination, identified based on their minimal RMSE values:

Table 11. Best Hyperparameter KomVar4

Rank	Hyperparameter						RMSE
	Number of units	Dropout Rate	Learning Rate	Batch size	Optimizer	Time stamps	
1	100	0.3	0.001	32	Adam	10	0.000758
2	100	0.1	0.001	32	Adam	10	0.000771
3	100	0.5	0.001	32	Adam	30	0.000840
4	150	0.5	0.001	32	Adam	10	0.000880
5	150	0.3	0.001	32	Adam	10	0.000924

Source: Research Process

Table 12. Best Hyperparameter KomVar12

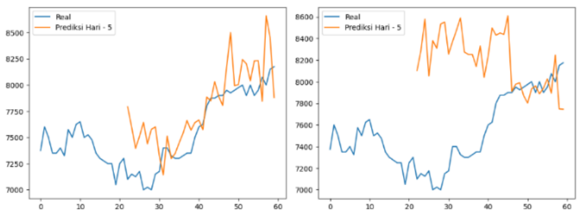
Rank	Hyperparameter						Rank
	Number of units	Dropout Rate	Learning Rate	Batch size	Optimizer	Time stamps	
1	150	0.1	0.001	32	Adam	10	0.000804
2	200	0	0.001	32	Adam	30	0.000863
3	200	0	0.001	32	Adam	10	0.000899
4	100	0	0.001	32	Adam	10	0.00092
5	300	0	0.001	32	Adam	10	0.000943

Source: Research Process

3.5 Model Evaluation

Once the optimal hyperparameter combinations for each model have been determined, the models are reutilized to forecast stock prices using the test data. The chosen hyperparameter combination represents the best configuration, signifying the sequence number one or, equivalently, the combination of variables yielding the smallest error within each respective model.

The KomVar4 model is subsequently reconstructed employing the hyperparameters: LSTM units of 100, dropout rate of 0.3, learning rate of 0.001, Adam optimizer, timestamps of 10, and a batch size of 32. Similarly, the KomVar12 model is reconstructed employing the hyperparameters: LSTM units of 150, dropout rate of 0.1, learning rate of 0.001, Adam optimizer, timestamps of 10, and a batch size of 32. Both models are then employed to predict stock prices for the ensuing 5 days using the test data. The outcomes are depicted in the following figure.



Source: Research Process

Figure 7. Stock Price Prediction Results for the Next 5 Days After Hyperparameter Optimization (KomVar4 (Left), KomVar12 (Right))

Analyzing the graph provided, a notable improvement is apparent in the predicted outcomes for the fifth day from the KomVar4 model after the implementation of hyperparameter optimization. The predictions are now more accurate and exhibit closer alignment with the actual values. Assessing the model's performance based on the RMSE metric, the prediction process is iterated 10 times to determine the average RMSE value, resulting in a value of 332.66. This outcome signifies significant enhancement compared to the pre-hyperparameter optimization state, which yielded an RMSE of 390.85.

In contrast, following the hyperparameter optimization process, the prediction outcomes for the KomVar12 model have diverged further from the actual values. Evaluating the model based on the RMSE metric, the prediction process is once again replicated 10 times to obtain an average RMSE value of 790.85. This result signifies a notable discrepancy when juxtaposed with the pre-hyperparameter optimization performance, where the average RMSE was 366.53, indicating both an improvement and a deterioration in different aspects.

4. Conclusion

Out of the 12 formulated technical indicators, following the feature selection process, only 7 indicators were retained for inclusion in the model. These indicators encompass SMA, StoD, RSI, MACD, ADO, CCI, and MFI. This particular combination was chosen due to its optimal fitness value, aligning with the model's performance requirements. The sentiment analysis endeavor has been successfully executed using the IndoBERT model, achieving an accuracy rate of 69%. This figure, however, falls short of being deemed satisfactory, considering the intricate nature of sentiment analysis. The computation of the daily sentiment score reveals that 73.74% of the daily scores tend toward a positive trend, surpassing the threshold of 0.5. Building upon the outcomes of the 10 walk forward validations, two LSTM models with the best variable combinations were obtained. Specifically, the fourth (4th) combination of variables, comprising closing price, technical indicators, IHSG, exchange rate, and Twitter sentiment yielded an average RMSE of 1.978, while the twelfth (12th) combination of variables, encompassing closing price, technical indicators, and IHSG recorded an average RMSE of 1.765. To enhance the model's performance, hyperparameter optimization was conducted for both types of variable combination models. As a result, the RMSE values for each model were obtained: 0.000758 for KomVar4 (the fourth combination of variables) and 0.000804 for KomVar12 (twelfth combination of variables). Subsequently, the models underwent evaluation using the test data, wherein the KomVar4 model showcased significant superiority over the KomVar12 model, demonstrating an RMSE of 332.66 compared to 790.85.

Reference

Aggarwal, D. (2022). Defining and measuring market sentiments: A review of the literature. Qualitative Research in Financial Markets, 14(2), 270–288.

Ardyanta, E. I., & Sari, H. (2021). A Prediction of Stock Price Movements Using Support Vector Machines in Indonesia.

Asia, B. C. (2022). Kepemilikan Saham BCA.

Attanasio, G., Cagliero, L., Garza, P., & Baralis, E. (2019). Combining news sentiment and technical analysis to predict stock trend reversal. 2019 International Conference on Data Mining Workshops (ICDMW), 514–521.

Baranochnikov, I., & Ślepaczuk, R. (2022). A comparison of LSTM and GRU architectures with novel walk-forward approach to algorithmic investment strategy (Vol. 21). University of Warsaw, Faculty of Economic Sciences.

Bhandari, H. N., Rimal, B., Pokhrel, N. R., Rimal, R., Dahal, K. R., & Khatri, R. K. C. (2022). Predicting stock market index using LSTM. Machine Learning with Applications, 9, 100320.

Derakhshan, A., & Beigy, H. (2019). Sentiment analysis on stock social media for stock price movement prediction. Engineering Applications of Artificial Intelligence, 85, 569–578.

Huang, Y., Capretz, L. F., & Ho, D. (2021). Machine learning for stock prediction based on fundamental analysis. 2021 IEEE Symposium Series on Computational Intelligence (SSCI), 1–10.

Indonesia, B. (2022). Data Suku Bunga.

Investing.com. (2022). USD IDR Historical Data.

Jiang, M., Liu, J., Zhang, L., & Liu, C. (2020). An improved Stacking framework for stock index prediction by leveraging tree-based ensemble models and deep learning algorithms. Physica A: Statistical Mechanics and Its Applications, 541, 122272. https://doi.org/10.1016/j.physa.2019.122272

Katoch, S., Chauhan, S. S., & Kumar, V. (2021). A review on genetic algorithm: past, present, and future. Multimedia Tools and Applications, 80, 8091–8126.

Kumbure, M. M., Lohrmann, C., Luukka, P., & Porras, J. (2022). Machine learning techniques and data for stock market forecasting: A literature review. Expert Systems with Applications, 197, 116659.

Mesquita, C. M., Oliveira, R. A. De, & Pereira, A. C. M. (2020). Combining an LSTM neural network with the Variance Ratio Test for time series prediction and operation on the Brazilian stock market. 2020 International Joint Conference on Neural Networks (IJCNN), 1–8.

Mndawe, S. T., Paul, B. S., & Doorsamy, W. (2022). Development of a stock price prediction framework for intelligent media and technical analysis. Applied Sciences, 12(2), 719.

Peng, Y., Albuquerque, P. H. M., Kimura, H., & Saavedra, C. A. P. B. (2021). Feature selection and deep neural networks for stock price direction forecasting using technical analysis indicators. Machine Learning with Applications, 5, 100060.

pypi. (2023). Snscrape python project.

Qu, Q., Li, Z., Tang, J., Wu, S., & Wang, R. (2019). A trend forecast of import and export trade total volume based on LSTM. IOP Conference Series: Materials Science and Engineering, 646(1), 012002.

Shah, J., Vaidya, D., & Shah, M. (2022). A comprehensive review on multiple hybrid deep learning approaches for stock prediction. Intelligent Systems with Applications, 200111.

Sivapurapu, S. A. (2020). Comparitive Study of Time Series and Deep Learning Algorithms for Stock Price Prediction. International Journal of Advanced Computer Science and Applications, 11(6).

Turchin, A., Masharsky, S., & Zitnik, M. (2023). Comparison of BERT implementations for natural language processing of narrative medical documents. Informatics in Medicine Unlocked, 36, 101139.

Wahyuddin, E. P. (2022). Pemodelan Prediksi Harga Saham Menggunakan Long-Short Term Memory: Studi Empiris pada Saham Bank Rakyat Indonesia.

Wu, S., Liu, Y., Zou, Z., & Weng, T.-H. (2022). S_I_LSTM: stock price prediction based on multiple data sources and sentiment analysis. Connection Science, 34(1), 44–62.

Yahoo. (2022). Yahoo Finance - Stock Market Live, Quotes, Business & Finance News.

Yang, C., Zhai, J., & Tao, G. (2020). Deep learning for price movement prediction using convolutional neural network and long short-term memory. Mathematical Problems in Engineering, 2020.

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing, 415, 295–316.

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. Neural Computation, 31(7), 1235–1270.