LSTM ALGORITHM ANALYSIS OF BANKING SECTOR STOCK PRICE PREDICTIONS

Rina Nopianti [1)] (0812 8320246)
Andreas Tri Panudju [2)]
Angrian Permana [3)]
Leni Triana [4)]

[1]Accounting Department, Bina Bangsa University,
[2]Industrial Engineering, Bina Bangsa University,
[3,4]Management Departement, Bina Bangsa University,
rinanopianti.binabangsa@gmail.com,
Panudju2002@gmail.com, mr.angrianpermana@gmail.com,
lenitriana.binabangsa@gmail.com

Abstract

Investing, buying or selling on the stock exchange demands data analytical expertise and skill. Because the stock market is so dynamic, it takes data modelling to predict stock prices accurately. Machine learning can currently process and forecast data with high accuracy. We proposed using the Long-Short Term Memory (LSTM) algorithm to model data to anticipate market prices. This study's primary goal is to assess the machine learning algorithm's accuracy in forecasting stock price data and the optimal model construction epochs. The RMSE value of the LSTM method and the data model obtained the variation of the epochs value.

Keywords: *LSTM Algorithm, Stock Price, Analysis Prediction, Machine Learning*

INTRODUCTION

A country's monetary policy has a significant impact on its economy. This financial instrument facilitates the purchase and sale of securities and serves as a source of funds for business and investment. Modal is a significant economic driver in some countries with a market economy, providing alternative funding for businesses. Besides that, the business now sells Bruto Produk (PDB). Thus, increasing PDB can help a developing country's economy (Nathaniel & Butar Butar, 2019).

Shares involve companies or business entities in the capital market (Rustam et al., 2018). The banking industry has solid and stable prospects year after year. In Forbes 2000 The World's Biggest Companies, three state-owned banks are listed: Bank Rakyat Indonesia, Bank Negara Indonesia, and Bank Mandiri. Due to daily fluctuations in stock prices, investors should pay attention and check historical banking data as a strategy for trading. This is very important because investors can assess the possible stock price of the company. In addition, investors can analyze the fundamentals of future investment banks. So, forecasting or forecasting methods are the best way to complete the analysis(Zhang et al., 2017).

There are many approaches to predicting stock prices. Methods for predicting stock prices include machine learning. Machine learning is a part of AI that aims to improve knowledge or performance (Cavalcante et al., 2016). Recurring Neural Network (RNN) is a machine learning algorithm. RRN is quite good at forecasting time series data (Kwang Gi Kim, 2019). The Short Term Memory Algorithm (LSTM) is based on the RNN algorithm and can extract information from data (Zhang et al., 2017).

Predicts models with LSTM algorithms for some variation of the times. Using 20 epochs, MSE is 0.00019 and RMSE is 0.014 (Fauzi, 2019) . Based on the variety of times, MSE and RMSE values have not changed much. (Chen et al., 2015) use three stages of prediction: preprocessing, data training, and testing. The RNN approach is not good at predicting stock prices. MSE values increase with various data used in SVM. LSTM generates the same MSE value even when using different data ranges (Dwiyanto et al., 2019). With small error numbers, LSTM can overcome long-term dependency and accurately predict stock prices.

As such, the research will include preprocessing, LSTM data modelling, data training, data testing, and data visualization. Initially, the share data of state-owned banks (BRI, BNI, BTN, and Mandiri) was obtained from yahoo finance. Stock data is 2016-2020. The second phase involves sharing training and testing data, with 80% (1006 data) going to training and 20% (251 data) to be tested. We tested Adam, RMSProp, and SGD. The trial also used epochs 25, 50, 75, and 100. The final stage displays the trial data as training, actual, and predictable.

RESEARCH METHODE

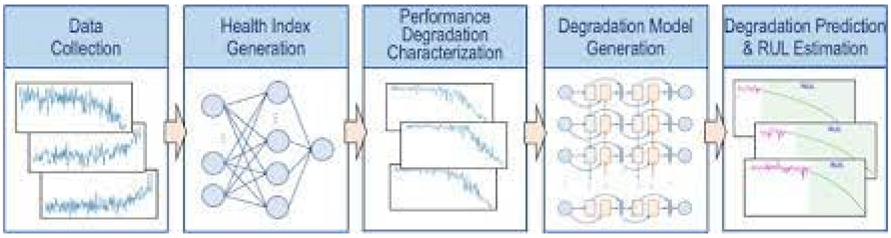

Figure 1. Data Processing System Design

Dates are sourced from Yahoo Finance and include BRI, BNI, BTN, and Mandiri from January 2016 to December 2020. Data includes date, price, volume, closing adj, and plit shares. Data will be used in preprocessing data segmentation. After that, the LSTM model is used, which is a type of Recurring Neural Network (RNN). After data modeling, data training and data testing were conducted by 80% and 20% respectively.

Data Collection

Tabel 1. BRI Stock Data

Date	Open	High	Low	Close*	Adj. Close**	Volume
27/12/202	4.380	4.440	4.380	4.430	4.263,68	56.465.900
26/12/2020	4.420	4.450	4.400	4.410	4.244,43	50.234.000
23/12/2020	4.450	4.470	4.360	4.450	4.282,93	123.276.700
⋮	⋮	⋮	⋮	⋮	⋮	⋮
06/01/2016	2.310	2.325	2.300	2.305	1.334,90	65.091.500
05/01/2016	2.320	2.340	2.315	2.320	1.343,59	49.569.000
02/01/2016	2.305	2.345	2.305	2.330	1.349,38	45.155.000

Yahoo Finance has banking stock statistics for BRI, BNI, BTN, and Mandiri. Date, open, close, high, low, close adj, volume column taken. When the stock market opens, the date is the date, month, and year for stock data. Open is the stock price data for the first day of trading. Daily price changes to sell or buy shares rationally. Close is the stock price at the end of the trade. Adj Close is the closing stock price that affects dividends and stock splits. The total number of shares or lots in a given period. The study used the daily closing stock price as a reference or data input for training and testing.

Data Preprosesing

The process of retrieving data in yahoo finance will be segmented or grouped in the data preprocessing step. Bank data is 1257 rows and 7 columns.

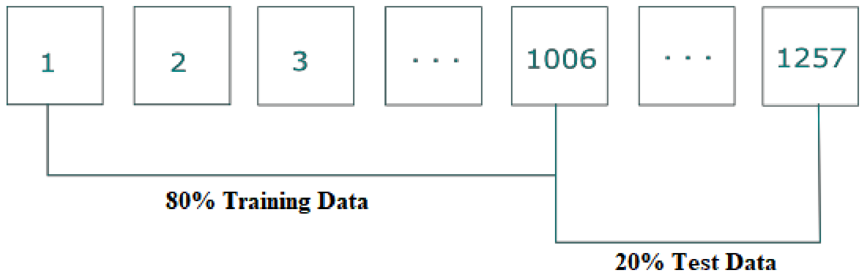

Figure 2. Segmentation of Training Data and Test Data

LSTM Model Design

Long Short Term Memory (LSTM) is a type of RNN used to solve hidden layer problems. The LSTM algorithm works by incorporating non-linear and dependent controls into an RNN cell (Gao et al, 2017), while maintaining a gradient of the destination function. LSTM overcomes vanish gradients or states where the gradient is 0 or close to 0. (Pulver &Lyu, 2017). That is, it takes into account past and present data. LSTM can record long-term reliance on efficiency.

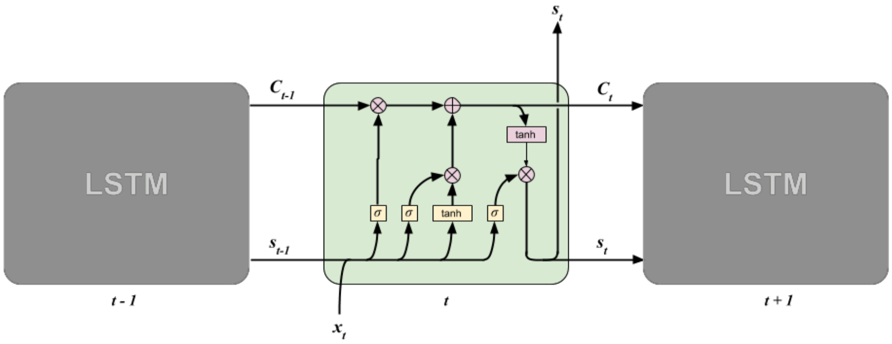

Figure 3. Long Short Term Memory Algorithm Design

In figure 3, it appears that LSTM has 3 gates including:

- Forget gate (ft) is used to determine the information to be eliminated from the cell using the sigmoid layer by activating the Relu function.
- Input gate (it) is used as a tool to forward information from the sigmoid layer that is done at the renewal and at the stage of the layer will be changed a vector to be updated.
- Output gate is used to display the contents of memory cells in the LSTM output process.

The initial stage in the LSTM method is to forget the information from C_{t-1}. This gate reads the values S_{t-1} and Xt and returns a value between 0 and 1 for each element in C_{t-1}. It is formulated, it is as follows:

$f_t = \sigma\,(W_f \,.\, [S_{t-1}, X_t] + b_f)$ ……………………(1)

It is possible to store temporary subject gender information in element C_{t-1} to use the correct pronouns. The previous subject in C_{t-1} can be omitted from Ct1.

The input gate determines which values to update. Create a new context vector candidate $\widetilde{C_t}$ for the layer. So it will be combined between the two for further context updates. So, in this example, the process is:

$$i_t = \sigma (W_i . [S_{t-1} , X_t] + b_i) \qquad (2)$$
$$\tilde{C}_t = tanh (W_c . [S_{t-1} , X_t] + b_c) \qquad (3)$$

When it will be done to update the old context C_{t-1} into the new context C_t. To eliminate things that have been decided then the process of forget gate (*ft*) in the equation (1) multiplied by the old context in the equation (2) and equation (3). Then, a new equation will be obtained as follows:

$$C_t = ft * C_{t-1} + i_t * \tilde{C}_t \qquad (4)$$

In the output gate process, updates will be performed on the cell and sigmoid layer to decide what parts of the context will be generated. So that the following equation will be obtained

$$o_t = \sigma (W_o . [S_{t-1} , X_t] + b_o) \qquad (5)$$
$$S_t = O_t * \tanh (C_t) \qquad (6)$$

Where for sigma (σ) is a sigmoid activation function with a range of values between -1 and 1 then *tanh* is a target activation function with a value (-1,1) while *Wt, Wi, Wc, Wo* is the matrix weight and for S_{t-1} is the previous hidden state and *bt, bi, bc, bo* is a vector can.

RESULTS AND DISCUSSIONS

Discussion of the study findings using the SOE banking dataset from 2015 to 2019, with 80% training data and 20% testing data focusing on stock data. The LSTM algorithm is separated into various processes using Adam, SGD, and RMSprop optimizers with variations of 25, 50, 75, and 100 epochs.

Banking Stock Dataset

The data used in the study was sorted from January 1, 2016 to December 30, 2020. The data analysis procedure that uses machine learning is focused on the price closing column, where this is the daily closing price for each stock data.

Table 2. Close Banking Stock Data

Date	BRI-(Close)	BNI-(Close)	BTN-(Close)	Mandiri-(Close)
27/12/202	4.430	7.925	2.150	7.750
26/12/2020	4.410	7.950	2.130	7.800
23/12/2020	4.450	7.925	2.120	7.725
⋮	⋮	⋮	⋮	⋮
06/01/2016	2.305	6.025	1.195	5.362
05/01/2016	2.320	6.025	1.220	5.400
02/01/2016	2.330	6.100	1.225	5.412

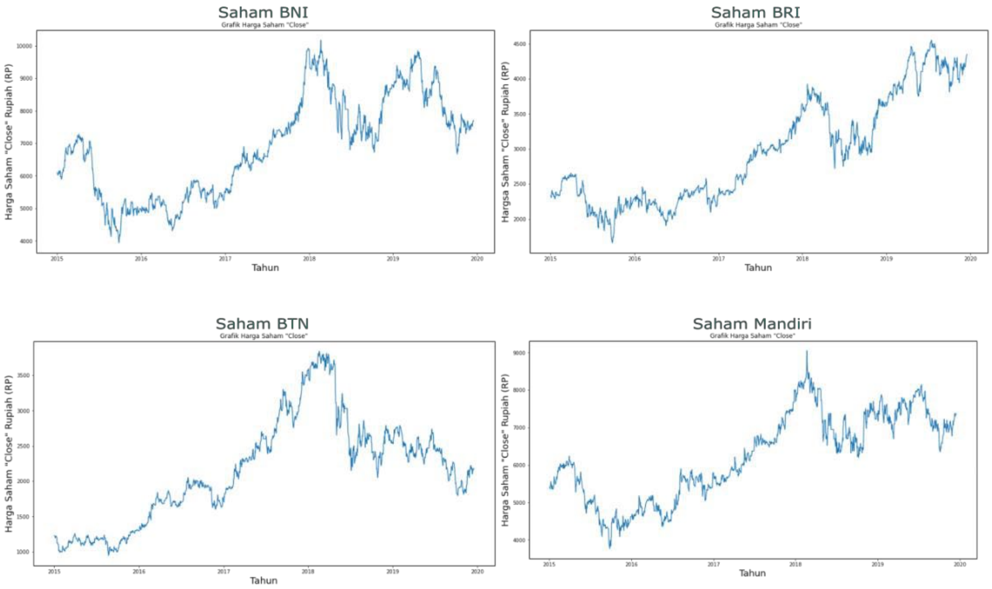

Figure 4. Actual Data of BNI, BRI, BTN and Mandiri Shares

The data in this study is displayed as a graph with horizontal lines representing close stock prices and vertical lines representing the year. The share prices closed the four lowest banks in 2016 and the highest in 2019. However, in contrast to the other four banks, BRI's share price continued to rise from 2016 to 2020, from about Rp. 1,300 to Rp. 4,500 per share. The other three banks also had their ups and downs, resulting in average prices or in the middle of the range.

Comparing Between Three Optimization Models

Table 3 examines optimization models, age variations, losses, and accuracy using stock data from four banks. The greater the value of the times, the lower the value of the resulting losses, which affects the accuracy of the optimization model. However, experimental data from four banks showed that each optimization model produced different accuracy statistics, requiring a separate investigation.

First, the SGD optimization model data shows it. In experiments, greater age values resulted in reduced loss rates and higher accuracy. Although a certain value displays an up and down value. This data shows that when epochs 100, there is a loss of 0.0013 in BRI bank data. The best accuracy of this optimization model is 61%. This is shown in three banks with a value of 100, namely BRI, BNI, and Mandiri, but not in BTN.

Second, Adam's optimization model shows it. The greater the value of the times, the lower the rate of loss. These second data losses are substantially smaller than the losses of the SGD optimization model. As can be seen from the four banks, the accuracy is always increasing, and practically everything is above 90%. The value of 25 gives BRI the lowest accuracy percentage of 89 percent, while the value of 100 gives BRI and Mandiri the highest accuracy rate of 95 percent.

Third, the RMSprop model shows it. By comparison, the trial losses of four small banks relative to the two optimization models. However, accuracy data varies by age. This optimization model can be inaccurate if the times variation is large enough.

Table 3. Stock Loss and Accuracy Data Testing

Optimization Model	Epochs	BRI		BNI		BTN		Mandiri	
		Loss	Accuracy	Loss	Accuracy	Loss	Accuracy	Loss	Accuracy
SGD	25	0,0014	57%	0,0023	57%	0,0021	49%	0,0021	55%
	50	0,0015	52%	0,0022	52%	0,0017	61%	0,002	56%
	75	0,0014	57%	0,0022	57%	0,0017	60%	0,0016	59%
	100	0,0013	61%	0,0019	61%	0,0017	58%	0,0017	61%
ADAM	25	4,28E-04	89%	6,27E-04	90%	5,40E-04	90%	6,43E-04	92%
	50	4,75E-04	92%	5,48E-04	91%	4,18E-04	91%	6,00E-04	92%
	75	3,76E-04	93%	5,89E-04	92%	3,95E-04	92%	5,09E-04	94%
	100	3,44E-04	95%	4,97E-04	92%	4,31E-04	92%	4,92E-04	95%
RMSprop	25	5,51E-04	95%	8,11E-04	92%	7,14E-04	84%	7,54E-04	86%
	50	4,67E+00	85%	6,54E-04	91%	5,90E-04	86%	6,85E-04	91%
	75	4,48E+00	53%	5,99E-04	79%	5,31E-04	90%	5,95E-04	88%
	100	4,44E+00	46%	5,52E-04	78%	4,91E-04	92%	5,63E-04	86%

Table 4 shows how the number of epochs, optimization models, computing times, and RMSEs affect data during training and testing. The greater the value of the times, the more computing time is generated. The RMSprop optimization model in Bank Mandiri's shares calculates in 103 seconds when epochs 25 years. The longest calculation time is 619 seconds in epochs 100 in BRI bank stocks.

However, when evaluated from RMSE data, values vary, with some showing up and down trends, and others showing monotonous declines. Monotonous declines indicate more accurate prediction data. First, in the SGD optimization model, RMSE data decreases, but not always, as stock data increases. The RMSprop optimization model is similar. However, because the RMSE data adam optimization model is more stable or monotonously lowered, the resulting prediction model is more accurate.

Table 4 shows how the number of epochs, optimization models, computing times, and RMSEs affect data during training and testing. The greater the value of the times, the more computing time is generated. The RMSprop optimization model in Bank Mandiri's shares calculates in 103 seconds when epochs 25 years. The longest calculation time is 619 seconds in epochs 100 in BRI bank stocks.

However, when evaluated from RMSE data, values vary, with some showing up and down trends, and others showing monotonous declines. Monotonous declines indicate more accurate prediction data. First, in the SGD optimization model, RMSE data decreases, but not always, as stock data increases. The RMSprop optimization model is similar. However, because the RMSE data adam optimization model is more stable or monotonously lowered, the resulting prediction model is more accurate.

Table 4. Computing Time Stock Data Testing and RMSE

Optimizer	Epochs	BRI		BNI		BTN		Mandiri	
		Computing Time	RMSE	**Computing Time**	RMSE	**Computing Time**	RMSE	**Computing Time**	RMSE
SGD	25	138	149,06	125	325,82	113	141,51	128	240,71
	50	274	162,73	249	316,28	241	108,26	253	236,15
	75	423	145,18	360	317,16	313	111,46	387	214,76
	100	619	134,99	501	291,62	493	116,94	490	219,97
ADAM	25	131	79,17	122	140,22	109	62,74	126	155,72
	50	253	75,71	220	167,81	225	64,16	262	126,21
	75	382	65,12	383	139,12	293	57,86	352	121,97
	100	524	57,31	505	140,24	433	48,32	447	116,64
RMSprop	25	125	57,39	114	136,04	113	105,49	103	146,41
	50	232	83,11	211	137,85	204	56,91	243	114,83
	75	394	179,51	376	194,68	311	51,44	329	127,91
	100	494	219,57	486	194,77	430	48,82	475	140,15

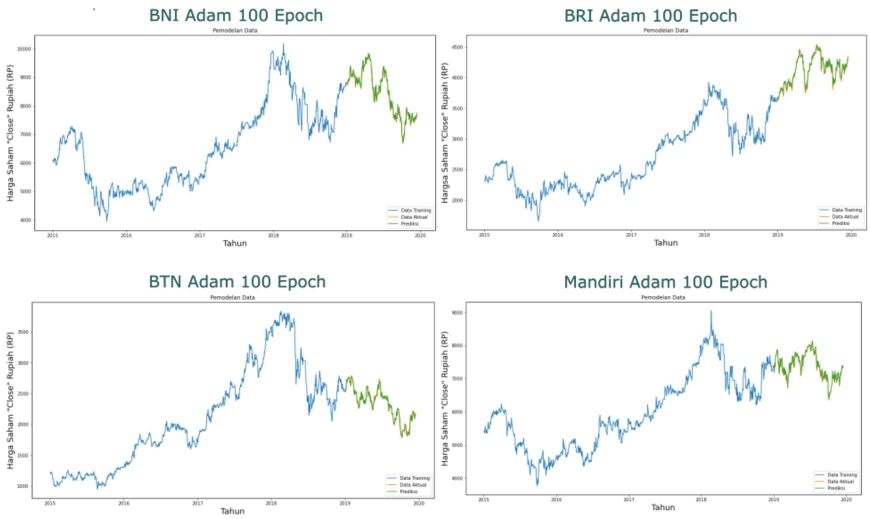

Figure 5. BNI, BRI, BTN and Mandiri Stock Prediction Data with Adam optimization model with 100 epoch

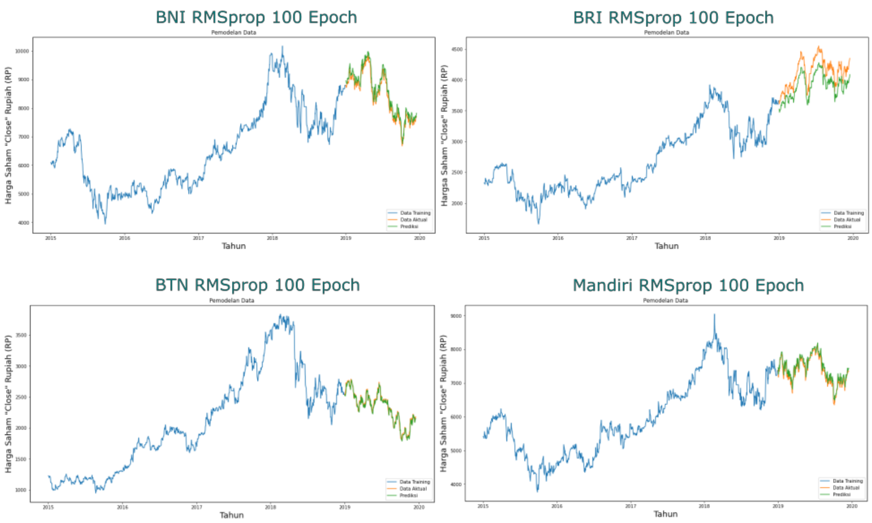

Figure 6. BNI, BRI, BTN and Mandiri Stock Prediction Data with RMSprop optimization model with 100 epoch

This is shown in Figure 6 using the RMSprop optimization model with epochs 100. The four charts show the difference between the actual and expected bri stock prices. Table 4 shows that bri bank's RMSE is 219.57, which is very high. As a result, the predicted and actual data lines are few. These estimates are less accurate in forecasting.

The experiment used an SGD optimization model with 100 epochs shown in Figure 7. Table 4 shows that the prediction model is less accurate between 100 and 300 RMSE. That is, the SGD optimization model is not accurate enough to check close stock data.

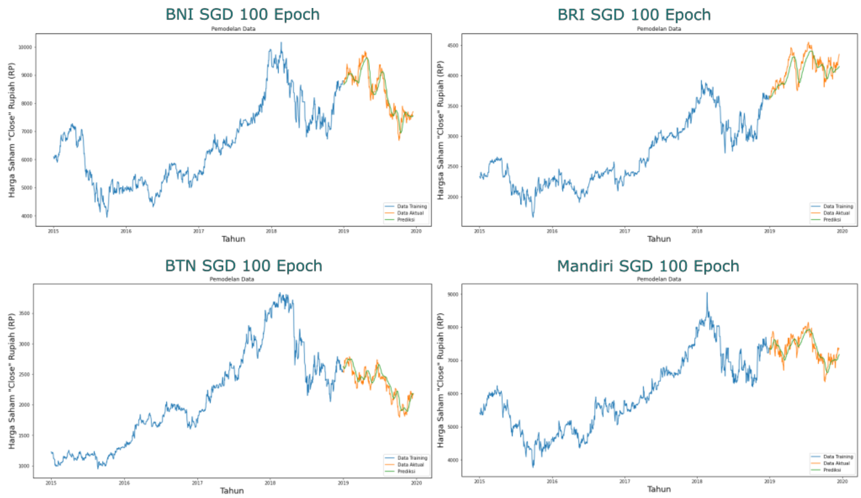

Figure 7. BNI, BRI, BTN and Mandiri Stock Prediction Data with SGD optimization model with 100 epochs

CONCLUSION

Optimization models, age variations, computing time, loss and accuracy values, and RMSE values were all examined in the study. The higher the value of the times, the longer it takes to complete the LSTM algorithm. Optimization models affect the effect of age variations on loss values and accuracy. According to Adam's optimization model, the higher the times, the lower the losses. The lower the value of the loss, the better the accuracy of the prediction of the stock data. RMSE is also substantially influenced by age value variations and optimization models, but Adam's model with age variations features the most stable RMSE decline from low to high age. Thus, utilizing Adam's optimization model on the LSTM algorithm results in a monotonously declining RMSE and high stock prediction value.

DAFTAR PUSTAKA

Cavalcante, R. C., Brasileiro, R. C., Souza, V. L. F., Nobrega, J. P., & Oliveira, A. L. I. (2016). Computational Intelligence and Financial Markets: A Survey and Future Directions. *Expert Systems with Applications*, *55*, 194–211. https://doi.org/10.1016/j.eswa.2016.02.006

Chen, K., Zhou, Y., & Dai, F. (2015). A LSTM-based method for stock returns prediction: A case study of China stock market. *2015 IEEE International Conference on Big Data (Big Data)*, *36*(3 PART 2), 2823–2824. https://doi.org/10.1109/BigData.2015.7364089

Dwiyanto, M. A., Djamal, C. E., & Maspupah, A. (2019). Prediksi Harga Saham menggunakan Metode Recurrent Neural Network. *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, 33–38.

Fauzi, A. (2019). Forecasting Saham Syariah Dengan Menggunakan LSTM. *Al-Masraf : Jurnal Lembaga Keuangan Dan Perbankan*, *4*(1), 65. https://doi.org/10.15548/al-masraf.v4i1.235

Kwang Gi Kim. (2019). Deep learning Deep Learning. *Healthcare Informatics Reserach*, *29*(7553), 1–73.

Nathaniel, A. S., & Butar Butar, S. (2019). Determinan Efisiensi Investasi Perusahaan Publik di Indonesia. *Jurnal Akuntansi Bisnis*, *17*(2), 192. https://doi.org/10.24167/jab.v17i2.2341

Rustam, Z., Vibranti, D. F., & Widya, D. (2018). *Predicting the direction of Indonesian stock price movement using support vector machines and fuzzy Kernel C-Means*. Author(s). https://doi.org/10.1063/1.5064205

Zhang, G., Xu, L., & Xue, Y. (2017). Model and forecast stock market behavior integrating investor sentiment analysis and transaction data. *Cluster Computing*, *20*(1), 789–803. https://doi.org/10.1007/s10586-017-0803-x