Online-ISSN 3025-3125

Long Short Term Memory Method and Social Media Sentiment Analysis for Stock Price Prediction

Nurkholis Amanullah [a,1], Agung Mustika Rizki [a,2], I Gede Susrama Mas Diyasa [b,3,*]

[a] Department of Informatics, Faculty of Computer Science, UPN "Veteran" Jawa Timur, Surabaya, Indonesia
[b] Master of Information System, Faculty of Computer Science, UPN "Veteran" Jawa Timur, Surabaya, Indonesia

[1] 20081010046@student.upnjatim.ac.id; [2] agung.mustika.if@upnjatim.ac.id; [3] igsusrama.if@upnjatim.ac.id*
* Corresponding author

ARTICLE INFO

Keywords
Long Short Term Memory
Sentiment Analysis
Stock Price Prediction
Social Media

ABSTRACT

The stock market is a complex arena of interest yet uncertainty. Trading stocks, binaries, gold, and bitcoin is growing in popularity, but is prone to price fluctuations influenced by economic and political factors. Social media, particularly Twitter, is where views on companies are shared. Social media sentiment analysis can provide additional insights to evaluate potential future stock price movements, preventing unwanted speculation. The purpose of this research is to develop a Tesla stock price prediction model by integrating the Long Short-Term Memory (LSTM) method and social media sentiment analysis from Twitter to improve prediction accuracy. Stock price data is obtained from Kaggle and Twitter sentiment data is processed through pre-processing. Evaluation values such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are lower in the model with sentiment indicating the ability of the model to more accurately model the dynamics of stock price movements. Lower MSE and RMSE indicate that the model's predictions are closer to the true values, and therefore, the model can be considered more reliable in projecting future stock price changes. These results provide support for the use of Twitter sentiment analysis as a useful source of additional information in improving the prediction accuracy of LSTM regression models in the context of stock market analysis.

1. Introduction

The stock market, as a realm full of complexity and uncertainty, stands out as one of the leading fields in finance. Shares, in the form of Limited Liability Companies (PT) or issuers, have gained high popularity as securities in the financial market [1][2]. Stock investment has become a common practice, with the selection of stock instruments requiring careful consideration of various aspects [3][4]. Investors can gain profits within a short period of time after investing their capital.

Trading, which includes buying and selling activities in financial markets, especially stocks, has become popular as it is considered profitable and attractive to the general public [5][6]. Nonetheless, trading activities also carry great risks due to price fluctuations influenced by economic factors, politics, and global news. Various types of trading, including stocks, binary, gold, and bitcoin, are increasingly favored by financial market enthusiasts.

.

The increasing public interest in capital market investments, especially stocks, has an important role in long-term economic growth [7-9]. Although stocks are an alternative source of financing, significant price fluctuations are the basis of research to prevent speculation in stock transactions [10].

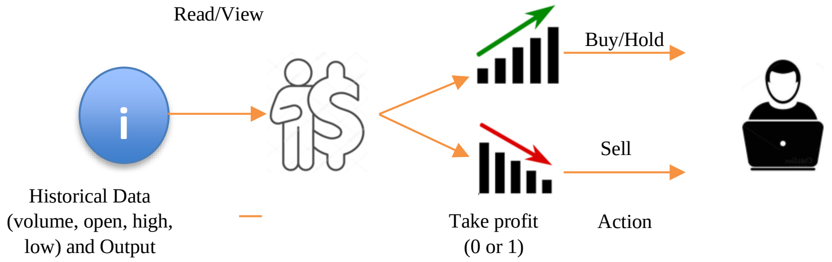


Figure 1. Research Target [11]

Stock price prediction is a focus of research with a significant impact on the global economy and individual decisions. Deep learning methods [12], such as Long Short-Term Memory (LSTM), offer promising results in predicting stock prices. LSTMs, as Artificial Neural Networks (ANNs), are designed to overcome long-term dependencies in data sequences such as stock prices [13-15]. RNNs, including LSTMs, are suitable for modeling time series data and are used in various applications such as stock market prediction, language translation, and signal processing [16][17].

Social media sentiment analysis is becoming an important focus for predicting stock price movements, especially on platforms such as Twitter [18]. Most previous studies tend to focus on using LSTM or social media sentiment analysis separately, leaving gaps related to data noise and bias in sentiment on social media [19]. This research integrates the LSTM method with social media sentiment analysis, specifically on Tesla (TSLA) stock, with more sophisticated pre-processing methods such as tokenization [20]. The goal of this research is to produce a stock price prediction model that combines the power of LSTM and social media sentiment analysis to help investors make better investment decisions [21].

2. Method

In this research, two algorithms namely Long Short-Term Memory and Sentiment Analysis are used for this research, both algorithms are used to predict stock prices with the LSTM function to predict then Sentiment Analysis to integrate the prediction results with social media sentiment, as shown in Figure 2.

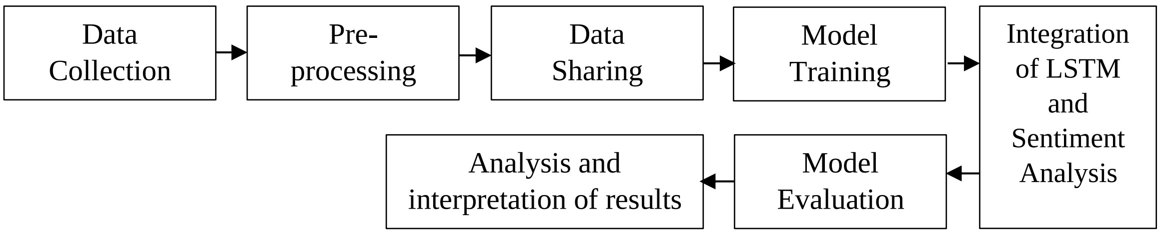


Figure 2. Block Diagram of Stock Analysis with LSTM and Social Media Sentiment

3.1. Acquisition Data

The dataset used in this study is the Tesla stock price dataset from October 4, 2018 to October 3, 2023 obtained by the author from Kaggle. This data includes important information such as daily closing prices, trading volume, and changes in stock prices during the period. In addition, social media sentiment data from the Twitter platform was also acquired for sentiment analysis related to Tesla stock during the period September 30, 2021 to September 30, 2022. This data acquisition process includes collecting, filtering, and simplifying the data required for further analysis, as shown in Figure 3 (Tesla Stock Acquisition Data) and Figure 4 (Twitter Sentiment Acquisition Data).

Date	Close/Last	Volume	Open	High	Low
10/03/2023	$246.53	101985300	$248.61	$250.02	$244.45
10/02/2023	$251.60	123810400	$244.81	$254.2799	$242.62
09/29/2023	$250.22	128522700	$250.00	$247.55	$246.35
09/28/2023	$246.38	117058900	$240.02	$245.33	$238.65
09/27/2023	$240.50	136597200	$244.262	$249.55	$241.6601
..	..	..	..	..	..

Figure 3. Tesla Stock Acquisition Data

Date	Tweet	Stock Name	Company
2022-09-29 23:41:16+00:00	Mainstream media has done an amazing job at brainwashing people. Today at work, we were asked what c...	TSLA	Tesla, Inc.
2022-09-29 23:24:43+00:00	Tesla delivery estimates are at around 364k from the analysts. $tsla	TSLA	Tesla, Inc.
2022-09-29 23:18:08+00:00	3/ Even if I include 63.0M unvested RSUs as of 6/30, additional equity needed for the RSUs is 63.0M ...	TSLA	Tesla, Inc.
2022-09-29 22:40:07+00:00	@RealDanODowd @WholeMarsBlog @Tesla Hahaha why are you still trying to stop Tesla FSD bro! Get your ...	TSLA	Tesla, Inc.
2022-09-29 22:27:05+00:00	@RealDanODowd @Tesla Stop trying to kill kids, you sad deranged old man	TSLA	Tesla, Inc.
..	..	..	..

Figure 4. Twitter Sentiment Acquisition Data

2.1. Pre-Processing Data

Data preprocessing involves cleaning and transforming Tesla stock data and Twitter sentiment data. This process includes removal of irrelevant data, handling of missing data, normalization of stock data, and text cleaning and tokenization of Twitter data. After the data was processed, both datasets were divided into training and testing sets. By performing efficient preprocessing, in this study to ensure good data quality for training and testing stock price prediction models with the LSTM method and social media sentiment analysis, the flow of data pre-processing is as shown in Figure 5.

2.2. Recurrent Neural Network

Long Short-Term Memory (LSTM) is a type of architecture in Recurrent Neural Network (RNN) developed by Hochreiter and Schmidhuber in 1997. LSTM is specifically designed to address the problem of long- and short-term dependencies in time-sequence data, a common challenge in stock price prediction. In this context, LSTM has the ability to recognize complex patterns in time sequence data and retain crucial information over a longer period of time [6].

Repair Data → Scaling Data → Create Data Train → Reshape Data

Figure 5. Data Pre-Processing Flow

The importance of using the LSTM method is seen in many studies, however, research conducted by Bhandari et al. (2022) [9] pointed out the shortcomings of using LSTM in stock price prediction, namely imperfections in the preprocessing method. One of the weaknesses identified was the non-use of data tokenization methods. Therefore, this study evaluates and improves the preprocessing process by applying the tokenization method. The tokenization method is used to improve the quality of data that will be used in LSTM modeling and social media sentiment analysis.

After training both models (LSTM and sentiment analysis), integration is performed by combining stock price features and social media sentiment features. Model performance evaluation is performed using evaluation metrics, where we choose Mean Squared Error and Root Mean Square Error algorithms as performance indicators. The Mean Squared Error algorithm measures the average of the squares of the difference between the predicted and actual values, while the Root Mean Square Error gives the average value of the square root of the difference. The selection of these evaluation metrics helps assess the extent to which the LSTM model and sentiment analysis can predict stock prices accurately and effectively, as in equation (1)

$$MSE = \frac{1}{m \times n} \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} [f(i,j) - g(i,j)]^2$$

$$RMSE = \sqrt{\frac{1}{m \times n} \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} [f(i,j) - g(i,j)]^2} \tag{1}$$

3. Results and Discussion

The research results are divided into four sections, namely: Data Pre-processing Results, Data Training Results, Data Evaluation Results, and Prediction Accuracy Results.

3.1. LSTM Pre-Processing Results

At this stage, data preprocessing is carried out on the stock price dataset that has been obtained on Kaggle, it is necessary to do processing and filtering so that it is easy to do computing or modeling, in Pre-Processing the following steps are carried out:

1. Create Dataset. This dataset refers to the type of dataset used, namely the tesla stock dataset for the LSTM algorithm and also the twitter sentiment dataset or X, this dataset is then put into two folders which will be processed as a dataset that is ready to be filtered and then used for research, as shown in Figure 6.

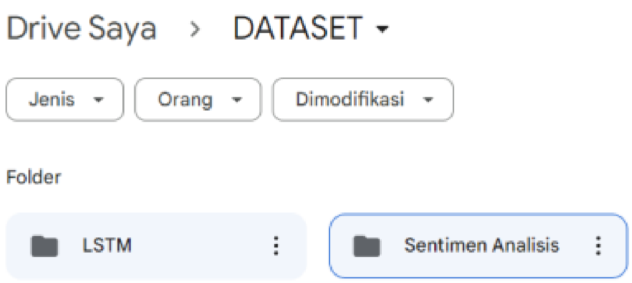


Figure 6. Dataset Folder Creation

2. Dataset Repair. After the dataset is divided, the process is then added to the close stock data, because in this study the data used is close data for research needs to predict stock prices, not only that at this stage there is a dataset repair, namely sorting date data, the following is the appearance of the dataset after repair and also sorting, as in Figure 7.

dates	date	close
10/03/2023	09/27/2023	$240.50
10/02/2023	09/28/2023	$246.38
09/29/2023	09/29/2023	$250.22
09/28/2023	10/02/2023	$251.60
09/27/2023	10/03/2023	$246.53

Figure 7. Dataset after repair and sorting

3. Data Filtering. Data filtering will produce a dataset that is ready to be scaled, this data can be said to be clean data because the data has been filtered or filtered, the following data results are before filtering and those that have been filtered, as shown in Figure 8 (a and b)

Date	Open	High	Low	Close	Adj Close	Volume
2017-01-03	28.950001	29.082500	28.690001	29.037500	27.297693	115127600
2017-01-04	28.962500	29.127501	28.937500	29.004999	27.267138	84472400
2017-01-05	28.980000	29.215000	28.952499	29.152500	27.405800	88774400
2017-01-06	29.195000	29.540001	29.117500	29.477501	27.711332	127007600
2017-01-09	29.487499	29.857500	29.485001	29.747499	27.965151	134247600
...	...	...	...	...	...	...
2021-12-27	177.089996	180.419998	177.070007	180.330002	180.100540	74919600
2021-12-28	180.160004	181.330002	178.529999	179.289993	179.061859	79144300
2021-12-29	179.330002	180.630005	178.139999	179.380005	179.151749	62348900
2021-12-30	179.470001	180.570007	178.089996	178.199997	177.973251	59773000
2021-12-31	178.089996	179.229996	177.259995	177.570007	177.344055	64062300

1259 rows × 6 columns

(a)

Date	Close
2017-01-03	29.037500
2017-01-04	29.004999
2017-01-05	29.152500
2017-01-06	29.477501
2017-01-09	29.747499
...	...
2021-12-27	180.330002
2021-12-28	179.289993
2021-12-29	179.380005
2021-12-30	178.199997
2021-12-31	177.570007

1259 rows × 1 columns

(b)

Figure 8. (a) Dataset before filtering. (b) Dataset after filtering

4. Data Scaling. Normalizing the actual data into values with a range of intervals [0,1], this process is used to make it easier to do data training, the results are as shown in Figure 9 (a)

Date	Close
2017-01-03	29.037500
2017-01-04	29.004999
2017-01-05	29.152500
2017-01-06	29.477501
2017-01-09	29.747499
...	...
2021-12-27	180.330002
2021-12-28	179.289993
2021-12-29	179.380005
2021-12-30	178.199997
2021-12-31	177.570007

1259 rows × 1 columns

```
[[0.16751901]
 [0.15289266]
 [0.16751901]
 [0.14701679]
 [0.14995473]
 [0.14995473]
 [0.15178588]
 [0.14894862]
 [0.15860752]
 [0.17156657]
 [0.17297513]
 [0.1513087 ]
 [0.1513087 ]
 [0.18631652]
 [0.17770399]
 [0.17500754]
 [0.1828554 ]
 [0.18987828]
 [0.15813127]
 [0.15813127]
 [0.15813127]
 [0.18553176]
 [0.18780564]
 [0.19088442]
 [0.1890331 ]
```

(a)

```
Data Train : 90%
[*********************100%***********************]  1 of 1 completed
Jumlah data train : 657
Jumlah data test : 73
Data Train : 80%
[*********************100%***********************]  1 of 1 completed
Jumlah data train : 584
Jumlah data test : 146
Data Train : 70%
[*********************100%***********************]  1 of 1 completed
Jumlah data train : 511
Jumlah data test : 219
```

(b)

Figure 9. (a) Scalling Data 0-1 (b) Splitting Data

5. Splitting Data. Divide the data into 2 parts, namely training data (Train Data) and test data (Test Data). This division is done with a composition of 90% train data and 10% test data, 80% train data and 20% test data, 70% train data and 30% test data, as shown in Figure 9 (b).

3.2. LSTM Data Training Results

After preprocessing the data, the next process is data training. Before conducting LSTM data training and Sentiment Analysis, it is necessary to build the architecture first to determine the neurons and layers used. the data training process consists of several stages, namely: model training, model evaluation, integration with sentiment, and saving the trained model, as show in Figure 10

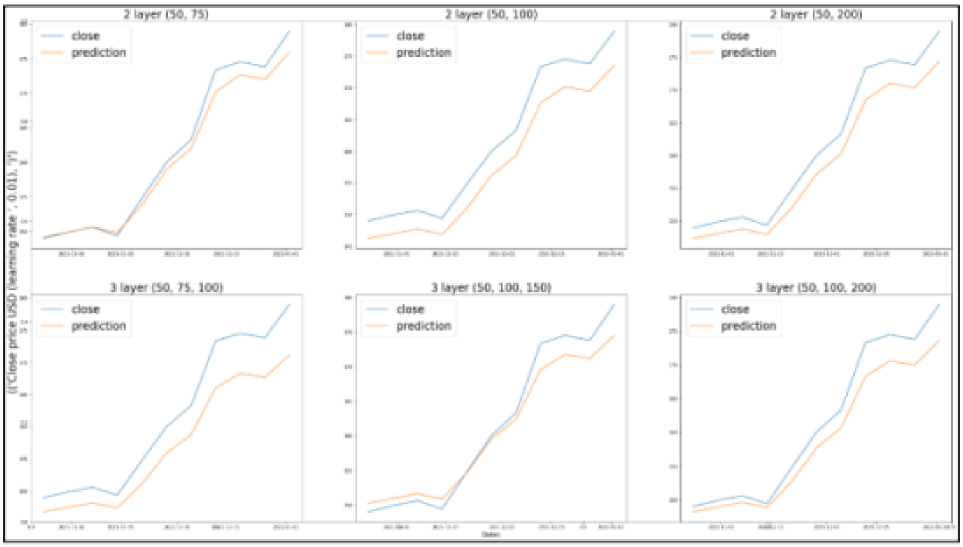
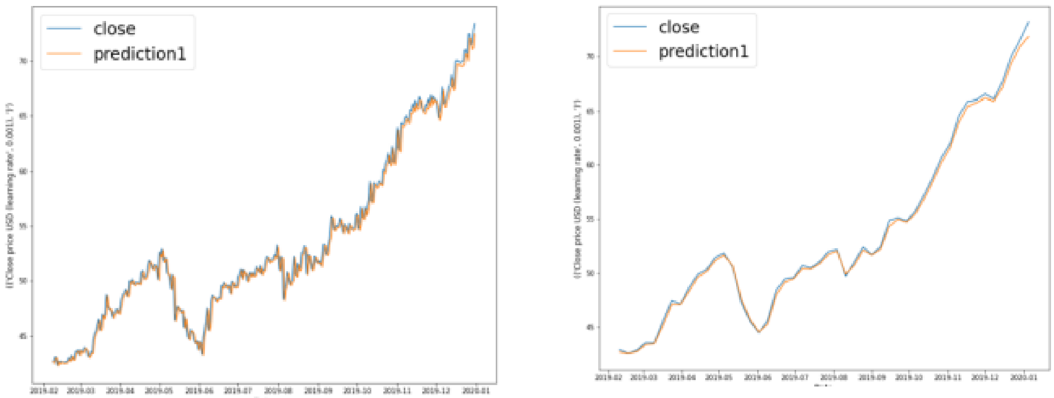


Figure 10. LSTM Training Schema

It can be seen in the figure above that LSTM training and sentiment analysis go hand in hand, the yellow line is the LSTM model training, the blue line is the sentiment analysis model training. The graph above proves that LSTM modeling has a positive value (the graph rises upwards) and can be integrated well with sentiment analysis as show in Figure11 (a). After training the model with the LSTM algorithm, sentiment analysis will be performed to obtain appropriate and accurate accuracy, as show in Figure 11 (b).



(a) (b)

Figure 11. (a) LSTM Training Results (b) Integration result of LSTM and Sentiment Analysis

3.3. Data Evaluation

After training both models (LSTM and sentiment analysis), integration is performed by combining stock price features and social media sentiment features. Model performance evaluation is performed using evaluation metrics, auration evaluation using Mean Squared Error and Root Mean Square Error algorithms as performance indicators. The Mean Squared Error algorithm measures the average of the squares of the difference between the predicted value and the actual value, while the Root Mean Square Error provides the average value of the square root of the difference. The selection of these evaluation metrics helps assess the extent to which the LSTM model and sentiment analysis can predict stock prices accurately and effectively, as show in Figure 12.

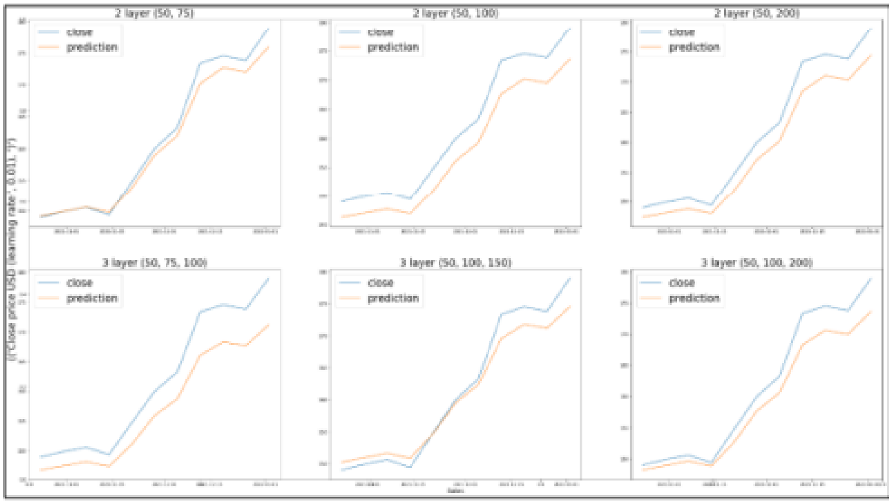


Figure 12. Training Result Graph

It can be seen in the figure 12. above that the blue line (with sentiment) as well as the yellow line (without sentiment) have different graphical increments and show accuracy, thus, the integration of Twitter sentiment into the LSTM model can improve the model's ability to understand stock market dynamics and provide more accurate predictions. While these results show progress, it is important to continue monitoring the performance of the model and consider further developments, such as parameter optimization and the use of additional features, to improve prediction performance. This evaluation provides an indication that Twitter sentiment analysis can add value in predicting stock price movements. The following are the accuracy results of LSTM calculations with sentiment and without sentiment, as show in Table 1.

Table 1. Accuracy Test Results

Model	MSE	RMSE
Sentimentless	0,000123	0,011089
With sentiment	0,000098	0,009898

This Table 3. presents a comparison of MSE and RMSE between the model without sentiment and the model with sentiment. The model with sentiment shows better performance, indicated by lower MSE and RMSE. Using evaluation metrics such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), we can see that the MSE and RMSE values of the model with sentiment are lower than those of the model without sentiment. The MSE value shows the average of the squared difference between the prediction and the true value, while the RMSE gives an idea of how large the overall error is by giving more weight to larger errors. Therefore, the lower values in the model with sentiment indicate the ability of the model to more accurately predict the closing price of the stock. This conclusion supports the view that the integration of Twitter sentiment can positively contribute to the prediction quality of LSTM regression models in the context of stock price analysis.

3.4. Prediction Accuracy Results

After compiling and comparing the actual price with the predicted price, the next step is to conclude the prediction accuracy results of the LSTM regression model evaluation for Tesla stock closing price prediction, it can be concluded that the model that utilizes Twitter sentiment as an additional feature shows an increase in prediction accuracy, as show in Table 2. Evaluation values such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are lower in the model with sentiment indicating the ability of the model to more accurately model the dynamics of stock price movements. Lower MSE and RMSE indicate that the model's predictions are closer to the true values, and therefore, the model can be considered more reliable in projecting future stock price changes. These results provide support for the use of Twitter sentiment analysis as a useful source of additional information in improving the prediction accuracy of LSTM regression models in the context of stock market analysis. Nonetheless, it is important to continuously monitor the performance of the model and perform additional validation to ensure the reliability of predictions under changing market conditions.

Table 2. Accuracy Presentation

Range	MSE	RMSE	Category
< 10%	156	170	Excellent forecasting model capability
10-20%	5	6	Good forecasting model capability
20-50%	1	3	The ability of the forecasting model is feasible
>50%	0	1	Poor forecasting model capability

3. Conclusion

This research successfully optimized the LSTM parameter configuration and integrated Twitter sentiment analysis as an additional feature in the LSTM regression model. This integration positively affects the accuracy of stock closing price prediction, by reducing the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) values. Results show that Twitter sentiment information contributes significantly to improving model performance, enabling a better understanding of the psychological aspects of the market that influence stock price movements. This research confirms the relevance of Twitter sentiment analysis in a financial context, providing a holistic view of the factors that can influence stock price movements, providing a foundation for market participants in making investment decisions.

References

[1] H. Hamilah, “The Influence of Upstream and Downstream Supply Chain Management on the Indonesia Stock Exchange,” Int. J. Sup. Chain. Mgt, vol. 9, no. 4, pp. 988, 2020.

[2] N. Choirunnisa and N. Fadilah, “Legal Protection for Investors in Crowdfunding Services Through Information Technology Offers (Equity Crowdfunding),” Syiah Kuala Law Journal, vol. 4, no. 2, Aug. 2020.

[3] S. Dewi and T. K. Pertiwi, “Analysis of Stock Investment Decisions on Investors in Surabaya,” Budapest International Research and Critics Institute (BIRCI-Journal): Humanities and Social Sciences, vol. 4, no. 2, pp. 2748-2759, 2021.

[4] S. Biswas, G. Bandyopadhyay, D. Pamucar, and N. Joshi, “A Multi-criteria-based stock selection framework in emerging market,” Operational Research in Engineering Sciences: Theory and Applications, vol. 5, no. 3, pp. 153-193, 2022.

[5] K. Vasista, “Role of a Stock Exchange in Buying and Selling Shares,” International Journal of Current Science (IJCSPUB), vol. 12, no. 1, pp. 2250-1770, 2022.

[6] R. T. Williams, An Introduction to Trading in the Financial Markets: Trading, Markets, Instruments, and Processes. Mill Valley, CA: University Science, 2011.

[7] R. Adewiyah and A. Bawono, “The Effect of Islamic Capital Market Instruments on Indonesia's Economic Growth is Moderated by The Money Supply,” ISLAMICONOMIC: Jurnal Ekonomi Islam, vol. 14, no. 1, 2023.

[8] J. Wuri, “The Role of Investment to the Indonesian Economic Growth,” Journal of Business & Finance in Emerging Markets, Sanata Dharma University, Indonesia, 2018.

[9] L. D. N. Kant, “Global Investment Policy Issues During the Covid-19 Pandemic and Policy Employment On Investment,” in Proc. International Conference Faculty of Law, vol. 1, no. 1, pp. 12-20, 2021.

[10] A. Slobodianyk and G. Abuselidze, “Influence of speculative operations on the investment capital: An empirical analysis of capital markets,” E3S Web of Conferences, vol. 234, p. 00084, 2021.

[11] Z. Syafitri, A. W. Suryani, “Stock Information on Social Media and Stock Return”, The Indonesian Journal of Accounting Research, Vol. 25 (3), pp. 383-412, 2022

.

[12] P.A. Riyantoko, I. G. Susrama, Sugiarto, Kraugusteeliana, ""F.Q.A.M" Feyn-QLattice Automation Modelling: Python Module of Machine Learning for Data Classification in Water Potability", International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), IEEE-xplrore, pp. 135-141, 2021

[13] D. A. Prawinata, A. D. Rahajoe, and I. G. S. M. Diyasa, "Analisis Sentimen Kendaraan Listrik Pada Twitter Menggunakan Metode Long Short-Term Memory," SABER: Jurnal Teknik Informatika, Sains dan Ilmu Komunikasi, vol. 2, no. 1, pp. 300-313, 2024.

[14] S. Jie and Azhar, "Stock Price Prediction Using Long Short-Term Memory," Applied Information Technology and Computer Science, vol. 4, no. 1, pp. 1311-1324, 2023.

[15] L. Lv, W. Kong, J. Qi, and J. Zhang, "An improved long short-term memory neural network for stock forecast," MATEC Web of Conferences, vol. 232, p. 01024, 2018.

[16] H. Hewamalage, C. Bergmeir, K. Bandara, "Recurrent Neural Networks for Time Series Forecasting: Current status and future directions", International Journal of Forecasting, pp. 1-40, 2020

[17] U. P. Iskandar and M. Kurihara, "Long Short-term Memory (LSTM) Networks for Forecasting Reservoir Performances in Carbon Capture, Utilisation, and Storage (CCUS) Operations," Scientific Contributions Oil and Gas, vol. 45, no. 1, pp. 35-51, 2022.

[18] A. Derakhshan and H. Beigy, "Sentiment analysis on stock social media for stock price movement prediction," Engineering Applications of Artificial Intelligence, vol. 85, pp. 569-578, 2019.

[19] R. A. Mendoza-Urdiales, J. A. Núñez-Mora, R. J. Santillán-Salgado, and H. Valencia-Herrera, "Twitter sentiment analysis and influence on stock performance using transfer entropy and EGARCH methods," Entropy, vol. 24, no. 7, p. 874, 2022.

[20] A. Bhadkamar and S. Bhattacharya, "Tesla Inc. Stock Prediction using Sentiment Analysis," Australasian Accounting, Business and Finance Journal, vol. 16, no. 5, pp. 52-66, 2022.

[21] M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos, and P. M. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms," Expert Systems with Applications, p. 119862, 2023.[1] H. Hamilah, "The Influence of Upstream and Downstream Supply Chain Management on the Indonesia Stock Exchange," Int. J. Sup. Chain. Mgt, vol. 9, no. 4, pp. 988, 2020.

.