

INCORPORATING STOCK PRICES AND SOCIAL MEDIA SENTIMENT FOR STOCK MARKET PREDICTION: A CASE OF INDONESIAN BANKING COMPANY

Dhenda Rizky Pradiptyo¹, Irfanda Husni Sahid², Indra Budi³, Aris Budi Santoso⁴, Prabu Kresna Putra⁵

^{1,2,3,4,5}Faculty of Computer Science, Universitas Indonesia

email: dhenda.rizky@ui.ac.id¹, irfanda.husni@ui.ac.id², indra@cs.ui.ac.id³, aris.budi@ui.ac.id⁴, prab003@brin.go.id⁵

Abstract

Forecasting the stock market is one of the most popular topics to be discussed in many fields. Many studies, especially in information technology have been conducted machine learning algorithms to achieve a more accurate prediction of the stock market. This research aims to find the effectiveness in predicting stock market performance by utilizing social media sentiment in combination with historical data. In addition, this research uses a machine learning algorithm to train a model to predict the stock price of each bank and training the model on a dataset that included the historical stock prices of the bank, as well as the sentiment scores of the social media posts about the bank and evaluate the performance of the model by comparing the predicted stock prices to the actual stock prices. The research shows that the R2 and RMSE score model that has been built with its historical data has slightly better performance than the model that has been built with the combination of historical data and social media sentiment. The finding indicates that the research method is closely correlated and affected to the performance of the stock market prediction.

Keywords : Sentiment Analysis, Stock Market, Forecasting, Prediction, Machine Learning.

Received: 20-01-2024 | Revised: 12-03-2024 | Accepted: 14-03-2024

DOI: <https://doi.org/10.23887/janapati.v13i1.74486>

INTRODUCTION

In the early stock market, the stock prices are shown on a big board and need to be collected by telephone from different places in a country [1]. But with the help of advanced technology and the internet, all the information about stock prices can be gained through the web in real-time and processed directly. It makes it easier for the investor to predict stock prices faster and more accurately.

As the technology and internet grow rapidly, the development of social media is also growing massively. Nowadays, social media impacts the whole world by reducing barriers to communication. All social media users can interact and share information across the nation through the social media platform. This platform offers a huge potential customer for a wide range of enterprises. More significantly, this potential has expanded exponentially over time and is predicted to do so in the future. This impacted the change of the company's business strategy in the last ten years [2].

The use of machine learning helps investors to predict stock market prices rather

than the traditional approach. With large amounts of data, using the computer would process the data faster and more precisely. The most common machine learning methods are NN and SVM. Although technical data is important for stock market prediction, it needs a more advanced strategy to outperform the market. In the last decade, the use of Twitter data as additional information can increase the accuracy of DJIA prediction by up to 87.6%.[3].

Any comments on social media can be influenced to a company's stock price especially those engaged in the financial industry such as banks. Many investors recently tend to invest in the banking sector, based on the top 50 biggest market cap data from Bursa Efek Indonesia [4] shows that there are 3 out of the top 5 lists are running in the banking industry, such as BBKA (11.30%), BBRI (7.29%), and BMRI (4.71%). Negative news can significantly affect to bank stock returns, specifically the news about corporate governance [5].

In recent times, much research has been done forecasting the stock price using Twitter data, and it has become popular because

the data on Twitter is already available to the public and thus automatically and quickly influences stock prices [6]. Moreover, Twitter's user is varied from regular people to celebrities, represent people of all age groups, have 500 million tweets a day, and it has 50 million downloads, and are used by many web browsers. That is why Twitter has become a fantastic platform for expressing and forming opinions [7].

Since social media is full of raw and unprocessed data, it has made it possible to process and turn that data into meaningful information that may help most business organizations [8]. Through sentiment analysis, the raw data could be extracted into some informational data. Sentiment analysis is a finding process of the mood of the sentence which is used to determine how a person feels about a particular subject or topic. It has a wide range of applications and can be used to enhance business quality and strategy, predict political elections, raise awareness of data security, understand how people feel about a particular sport, and better locate and respond to disasters. This demonstrates how sentiment analysis is extremely important for understanding people's perceptions and for aiding in decision-making [9].

This paper aims to find the relationship between the stock prices and social media sentiment in predicting stock prices in a case of Indonesian banking company.

SOCIAL MEDIA SENTIMENT ANALYSIS

Nowadays, social media gains attention more than traditional media. This brought social media to the arena of public opinion. Hence social media is one of the sources to get investors' behavior data [10]. One of the main techniques used to extract the information of investors' behavior is by using sentiment analysis to determine the polarity of public opinion. A positive polarity means that the majority of opinion present in public supports the corresponding topics, and a negative opinion is vice versa [12].

The main technique that is used to determine the relationship of stock valuation is linear regression [11][12][5][13]. This technique is used to determine whether social media affects stock valuation by calculating the regression coefficient along with its p-value. Every paper used sentiment analysis from social media to determine the polarity of the data (positive or negative) [5][19][14].

Aside from techniques, the platform is also an important factor. According to [10], most of the papers focused on social media with a

proportion of 58% to investigate investor behavior. This is due to the ease of availability of the data. Across all social media presented in [10], Twitter ranks at number 2 just below the Google search volume index. This means that recent studies also focused on Twitter to determine investor behavior.

TECHNICAL ANALYSIS

There are many research works on stock market prediction on different technical indicators such as open price [15], closed price [15], and technical market index [16]. This approach used the price and volume of historical prices and market indices to estimate future prices or trends of the market [15]. These indicators could explain the characteristics of past and current stocks. However, historical data only could not explain behavioral factors [17]. A research conducted by Sharma [10] utilizing Random Forest combined with LSBoost outperforms Support Vector Regression in predicting CNX Nifty, and S&P Bombay Stock Exchange. Another paper [18] by Billah et.al conducted a comparison of predicting closing stock price with artificial neural networks (ANN) and Adaptive Network-Based Fuzzy Inference System (ANFIS). The result shows that ANN performed better than ANFIS with 53% less error. The state-of-the-art approach for predicting price movement based on its historical prices is the LSTM model, Fischer and Krauss [19] deploy LSTM models to predict out-of-sample the price movement of the S&P 500. There are some comparisons of methods that show LSTM's ability to perform better than other methods

PREDICTION WITH SOCIAL MEDIA SENTIMENT

Combining historical prices, indexes, and sentiment analysis provided by social media is improving the model's performance [20][1][17] and [16]. Ren [21], conducted research to predict the stock market based on social sentiment using Support Vector Machine (SVM). Other studies by Li [1] predicts Hong Kong Stock Market Data using Long Short-Term Memory (LSTM), Support Vector Machine (SVM), and Multiple Kernel Learning (MKL). Li [1] incorporates historical index (MACD, MA, RSI, etc.) and sentiment analysis in the study. The result shows that LSTM outperforms other baseline models in both accuracy and F1 Score. Moreover, LSTM which uses both sentiment analysis and historical indexes outperforms other models that only use either of them. Chen et.al. [16] develop an RNN-Boost Method that incorporates sentiment analysis from Sina Weibo and historical indexes to predict The Shanghai-

Shenzhen 300 Stock Index (HS300), the result shows that RNN Boost performed better than Artificial Neural Networks that use the same features. Chen [16] also compares the RNN-Boost method with other methods that only use historical prices namely Support Vector regression and linear regression. A study from [20] proposed a coupled matrix and tensor factorization method to integrate events, sentiments and quantitative features. This research shows that incorporating social media could improve the performance of stock market predictions.

METHOD

Data Collection

The data of this research is collected and divided into price data, news data, and sentiment data dictionary. The price data is acquired from Yahoo Finance, while the news data is sourced from Twitter. To collect data from Yahoo Finance and Twitter, this research conducts Python programming by using Python library the *bsi_sentiment*, while the sentiment data dictionary is obtained through Google search engine.

The last 1 year of the stock price from its historical data is set as the training data in this research. Additionally, the news data related to this research, collected from Twitter, is also set to encompass the last 1 year of tweets. The data dictionary covers all the words in Bahasa. The keywords that were used to acquire the data were "saham bca", "saham mandiri", and "saham bri". The number of tweets that has been collected were 3,134 tweets. Training data consisted of data ranging from November 29, 2021, to August 31, 2022, and testing data consisted of data ranging from September 1, 2022, to November 29, 2022.

Technical Feature Engineering

Technical features are obtained and calculated based on historic dataset of price market. Denote that t is a symbol for the given trading day, and $t-1$ is the previous trading day with 1-day lag. The technical features are listed in Table 1.

Sentiment Feature Engineering

Technical features are obtained and calculated based on historic dataset of price market. Denote that t is a symbol for the given trading day, and $t-1$ is the previous trading day with 1-day lag. The technical features are listed in Table 1. The tweet data were scraped from Twitter by using *bsi_sentiment* library in Python which developed by a student association called BSI Bocconi. The library itself performs scrapping the tweet by providing the user to query tweet, dates, language, location, radius, and maximum tweet.

The labelling process of each tweet's polarity were using lexicon-based VADER. The data is translated to English by using Google API, and then got tokenized. For each news N , it should be converted into word vector $W = [w_i, w_j, \dots, w_n]$. Then each word from the word vector gets checked to find a phrase using word combination.

In case a phrase still be found, then the corresponding tokenized word is removed and replaced with a phrase word. The word vector also got normalized to match similar words and get stemmed to get the root form of the word. Then stop words are excluded. The polarity score for each tweet was then calculated. Finally, the daily average polarity and tweet were calculated by aggregating each tweet by its date. The illustration for tweet polarity and sentiment feature can be seen in Table 2 and 3.

Table 1. Technical Feature

Feature	Notation	Source
Today's Opening Price	O_t	Direct Source
Today's Closing Price	C_t	Direct Source
Today's Highest Price	H_t	Direct Source
Today's Lowest Price	L_t	Direct Source
Today's Volume	V_t	Direct Source
Price Difference from Yesterday	$C_t - C_{t-1}$	Calculated
Price Limit from Yesterday	$(C_t - C_{t-1}) / C_{t-1}$	Calculated
Volume Difference from Yesterday	$V_t - V_{t-1}$	Calculated

Volume Limit from Yesterday	$(V_t - V_{t-1}) / V_{t-1}$	Calculated
Amplitude from yesterday	$(H_t - L_{t-1}) / C_{t-1}$	Calculated
Difference from Yesterday	$(C_t - O_t) / C_{t-1}$	Calculated

Table 2. Tweet Polarity

Tweet ID for The Same Day	Polarity
T1	-1
T2	0.5
Average Polarity	-0.25

Prediction

After technical and sentiment features have been constructed, both features are concatenated by the same date. Since each feature has a different scale, value normalization using Minmax value normalization is applied to this dataset. The dataset then has technical features, and sentiment features. In

order to build a prediction model, the dataset needs to have a label column. In this paper, the dataset is labeled with a closing price for the next trading day. This can be achieved by shifting the stock prices in a row to the previous row. Labelling using the closing price for the next day is done because this model works by predicting the price for the next day based on today's data and historical data. The sample of data labelling process can be seen in Table 5 or the initial forms of the tables and Table 6 for the sample result of labelling process. Hence for features with time t , a closing price $t+1$ is concatenated. The dataset's illustration can be viewed in Table 4, and it is utilized as input for the mode

Table 3. Sentiment Feature

Feature	Notation	Source
Number of Positive Tweets	P_t	Direct Source
Number of Negative Tweets	N_t	Direct Source
Average Polarity	PLR_t	Calculated

Table 4. Dataset Feature

Date	Technical Feature	Sentiment Feature	Label
T	O_t to $(C_t - O_t) / C_{t-1}$	P_t to PLR_t	C_{t+1}
T-1	O_{t-1} to $(C_{t-1} - O_{t-2}) / C_{t-2}$	P_{t-1} to PLR_{t-1}	C_t
T-2	O_{t-2} to $(C_{t-2} - O_{t-2}) / C_{t-2}$	P_{t-2} to PLR_{t-2}	C_{t-1}

Table 5. Sample Initial Forms

Date	Closing Price
22 November 2022	4590.0
23 November 2022	4670.0
24 November 2022	4720.0

Table 6. Sample Result of Labelling Process

Date	Features	Closing Price	Label
22 November 2022	Calculated	4590.0	4670.0

23 November 2022	Calculated	4670.0	4720.0
24 November 2022	Calculated	4720.0	4750.0

This paper uses the LSTM model to compare the effect of social media sentiment analysis as a feature. LSTM is a specialized type of Recurrent Neural Network (RNN) designed to capture long-range dependencies in sequential data. It utilizes memory cells and a gating mechanism consisting of input, output, and forget gates to regulate information flow. In terms of regression or forecasting, LSTM can capture temporal dependencies in time-series data. This capability of LSTM makes it suitable

for predicting stock prices because it can capture the important relationship between features and labels. The logic flow of LSTM can be seen in Figure 1.

Two models were created in this study. The first model is the baseline model, which is an LSTM model with only technical features, and the second model is an LSTM with both features. The illustration for the methodology of this paper can be seen in Figure 2.

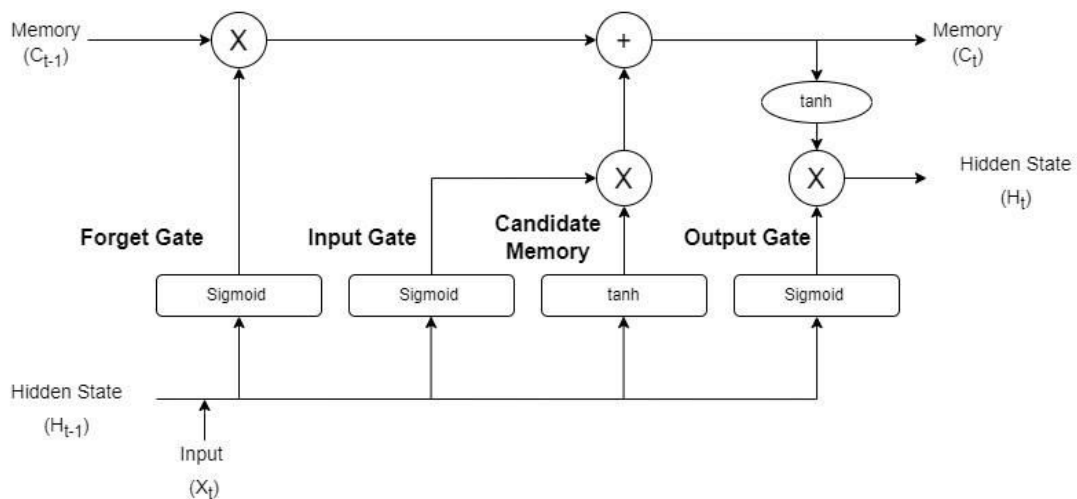


Figure 1. LSTM Logic Flow

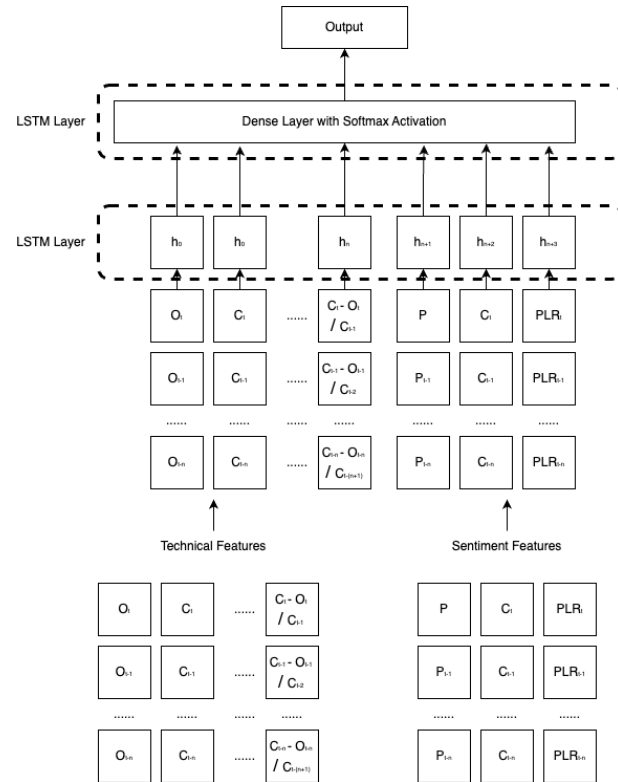


Figure 2. Research Methodology

RESULT AND DISCUSSION

The experiment has been carried out to compare between LSTM model using technical indicators and LSTM model using both technical and sentiment indicators. The price data are sourced from Yahoo Finance API with the keyword “BBCA.JK” for Bank BCA stock, “BMRI.JK” for Bank Mandiri stock, and “BBRI.JK” for Bank BRI stock. The data constructed and shaped for every stock is 246 rows and 12 columns. The tweet data were sourced by performing *bsi_sentiment* library in Python. The tweet data shape is shown in Table 7.

Table 7. Number of Tweets

Stock	Tweet Numbers
BBCA	915
BMRI	557
BBRI	1662

The price data are first converted to the technical features such as amplitude, price difference, and other features listed in Table III. Then the tweet data are also converted to the sentiment features as listed in Table 3. Based

on Li’s research [1], the model used to forecast the stock prices is the LSTM model with the hyperparameters configuration listed in Table 8. This model’s parameters are used in both scenarios.

Table 8. LSTM Parameter

Parameter	Value
Hidden Layer	1
LSTM Neuron	11
Optimizer	Adam
Maximum Epoch	100
Callback Method	True

Once the features were built and the models were trained, they were tested using the test data to evaluate their performance. The model itself is divided into 2 models. Model 1 is comparing only the technical features, while Model 2 is comparing the technical features with the sentiment features. These models are being evaluated by scoring their R2 and RMSE to see the fitness of the model that has been built. R2 score was evaluated by measuring the list of the prediction’s regression line is fit to the actual data, while RMSE was evaluated by computing the significant difference between the actual and

the prediction data. The results of R2 and RMSE metrics for the test data of the models are shown in Table 9 and Table 10.

Table 9. R2 Score Performance

Stock	R2 Model 1	R2 Model 2
BBCA	0.71	0.52
BMRI	0.14	-1.53
BBRI	0.76	0.75

Table 10. RMSE Score Performance

Stock	RMSE Model 1	RMSE Model 2
-------	--------------	--------------

BBCA	128.63	167.21
BMRI	158.97	272.80
BBRI	69.47	70.51

The result from Table 9 shows that the R2 value in Model 1 is slightly better compared to Model 2, while the RMSE value also shows that Model 1 is slightly better than Model 2. The addition of the sentiment feature played a pivotal role in this model. The plots for Model 1 and Model 2 of each stock are shown in Figure 3 and Figure 4. The plot shows the comparison of the target and the prediction price of BBCA, BMRI, and BBRI in an orderly.

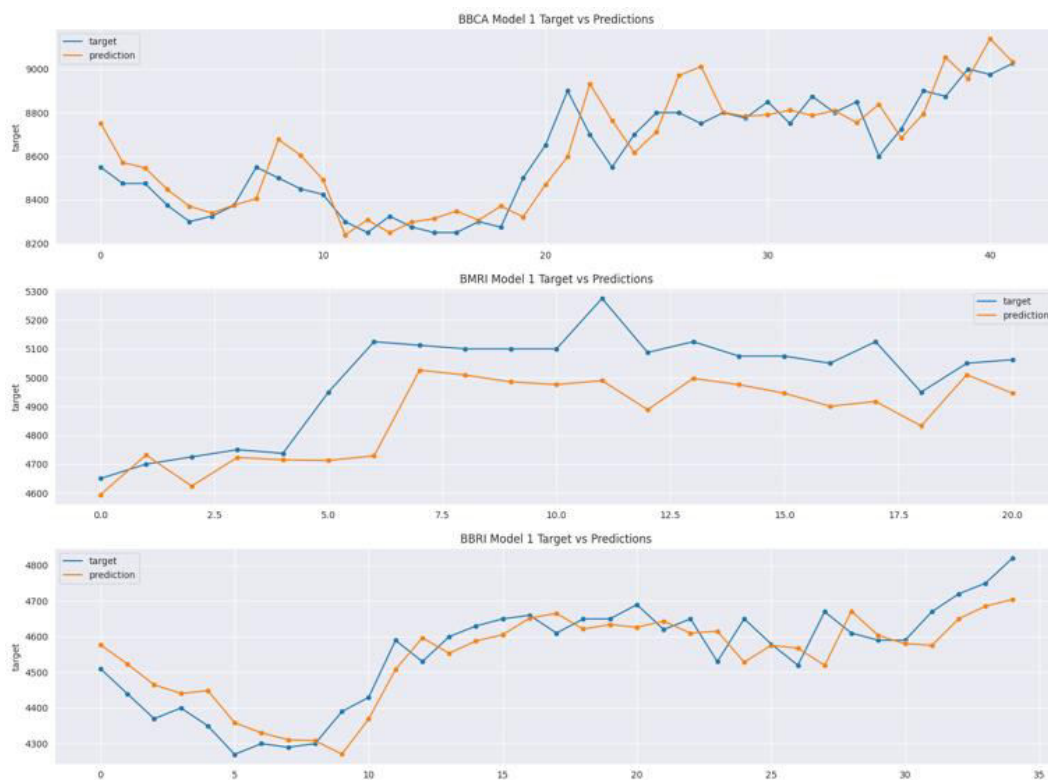


Figure 3. Target vs Prediction Model 1



Figure 4. Actual vs Prediction Model 2

Table 11. Paired t-test Performance

Stock	N	t-value	p-value	H0 Rejected (if p<0.05)	Conclusion
BBCA	41	-11.815071369392557	0.000000000000008	Yes	Different Significantly
BMRI	20	-3.8288054971038203	0.001049540888755	Yes	Different Significantly
BBRI	34	-2.3817019724982873	0.022973019458009	Yes	Different Significantly

From this table, the t-test performance result demonstrates a significant difference in the forecast's distribution. Across the examined stocks, the t-values of -11.8, -3.829, and -2.381 and the p-values below 0.05 concludes that the null hypothesis (H0) is rejected and affirming both models produce different mean and distribution. This means that there is strong enough evidence to conclude that the population of the model with sentiment is different from the model with only historical data.

CONCLUSION

The research compared two LSTM models' performance. The first LSTM was trained using technical features, while the other LSTM was trained using technical and sentiment features. The technical features consist of today's high price, low price, volume, open price, and close price. Some features are also derived from the previous day's stock price such as amplitude, price difference, and volume difference. This data was sourced from Yahoo Finance API. As for sentiment features, there are positive tweets, negative tweets, and daily average polarity. This data was sourced from Twitter using the bsi_sentiment library run in

Python. The stocks used in the research include BBCA, BBRI, and BMRI.

The model that has been constructed performs the value of R2 around 0.7 except for BMRI. The result of the LSTM performance shows that adding sentiment features decrease the performance and increased the RMSE metrics. However, the difference between each performance is small. The RMSE for BBCA and BBRI stocks respectively are 128.63 and 69.47 for Model 1, and 167.21 and 70.51 for Model 2. While the RMSE for Model 1 on BMRI is worse compared to Model 2 as the score shows 0.14 and -1.53. The research findings indicate that Model 1, which solely relies on technical features, outperforms Model 2, which incorporates both technical features and sentiment analysis, in terms of R2 value. This suggests that the inclusion of sentiment analysis does not substantially improve the model's predictive accuracy. These finding also indicates that the process of translating tweets in Bahasa into English made the scoring process of the tweet's polarity inconsistent.

Furthermore, following the addition of sentiment analysis to Model 2, a paired t-test was conducted to examine whether the sentiment feature significantly influences the mean. The results of the paired t-test show that the sentiment feature does not differ significantly from zero. This suggests that the sentiment analysis, as integrated into Model 2, does not have a statistically significant impact on the prediction of the target variable.

In summary, the research demonstrates that incorporating sentiment analysis into the predictive model does not lead to a notable enhancement in forecasting performance, as evidenced by the lower R2 value compared to Model 1. Additionally, the paired t-test results imply that the sentiment feature's effect on the model's predictions is not statistically significant. These findings have implications for refining predictive models in investment decision-making and call for further exploration of alternative approaches to incorporate sentiment analysis effectively.

Moreover, the future work is needed to this research area that would delve deeper into the dynamics between sentiment analysis and stock prices. While it may be evident that sentiment alone does not significantly influence stock prices, it's crucial to explore the role of sentiment as a medium for disseminating information related to corporate actions and announcements. Therefore, the focus will shift towards understanding whether it is the sentiment itself or the extent of information dissemination regarding these corporate actions

that exerts influence on share prices. In the technical area, the research could be improved by extending the length of stocks and tweet data, expanding the keyword of tweet data to gather more users. Then hyperparameter tuning for LSTM model needs further research to determine what architecture suits the stock price historical data.

REFERENCES

- [1] X. Li, P. Wu, and W. Wang, "Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong," *Inf Process Manag*, vol. 57, no. 5, Sep. 2020, doi: 10.1016/j.ipm.2020.102212.
- [2] B. Kaushik, H. Hemani, and P. V. Ilavarasan, "Social media usage vs. stock prices: An analysis of Indian firms," in *Procedia Computer Science*, Elsevier B.V., 2017, pp. 323–330. doi: 10.1016/j.procs.2017.11.376.
- [3] A. Porshnev, I. Redkin, and A. Shevchenko, "Machine learning in prediction of stock market indicators based on historical data and data from twitter sentiment analysis," in *Proceedings - IEEE 13th International Conference on Data Mining Workshops, ICDMW 2013*, IEEE Computer Society, 2013, pp. 440–444. doi: 10.1109/ICDMW.2013.111.
- [4] Bursa Efek Indonesia, "50 Biggest Market Capitalization - September 2022." Accessed: Oct. 23, 2022. [Online]. Available: <https://www.idx.co.id/data-pasar/laporan-statistik/digital-statistic-beta/biggest-market-cap?q=eyJ5ZWYljoimjAyMiIsIm1vbnRoiOiJ0b250aGx5In0=>
- [5] F. Carlini, D. Cucinelli, D. Previtali, and M. G. Soana, "Don't talk too bad! stock market reactions to bank corporate governance news," *J Bank Financ*, vol. 121, Dec. 2020, doi: 10.1016/j.jbankfin.2020.105962.
- [6] D. Shah, H. Isah, and F. Zulkernine, "Predicting the Effects of News Sentiments on the Stock Market," in *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, Institute of Electrical and Electronics Engineers Inc., Jan. 2019, pp. 4705–4708. doi: 10.1109/BigData.2018.8621884.
- [7] Jain College of Engineering, Institute of Electrical and Electronics Engineers. Bangalore Section., and Institute of

- Electrical and Electronics Engineers, *2020 International Conference for Emerging Technology (INCET): Belgaum, India. Jun 5-7, 2020.*
- [8] T. Mankar, T. Hotchandani, M. Madhwani, A. Chidrawar, and L. C. S \$ # Be, "Stock Market Prediction based on Social Sentiments using Machine Learning; Stock Market Prediction based on Social Sentiments using Machine Learning," 2018.
- [9] Z. Drus and H. Khalid, "Sentiment analysis in social media and its application: Systematic literature review," in *Procedia Computer Science*, Elsevier B.V., 2019, pp. 707–714. doi: 10.1016/j.procs.2019.11.174.
- [10] Siddhant College of Engineering and Institute of Electrical and Electronics Engineers, *2017 2nd International Conference for Convergence in Technology (I2CT)*.
- [11] P. Eachempati, P. R. Srivastava, A. Kumar, J. Muñoz de Prat, and D. Delen, "Can customer sentiment impact firm value? An integrated text mining approach," *Technol Forecast Soc Change*, vol. 174, Jan. 2022, doi: 10.1016/j.techfore.2021.121265.
- [12] C. Wu, X. Xiong, Y. Gao, and J. Zhang, "Does social media distort price discovery? Evidence from rumor clarifications," *Res Int Bus Finance*, vol. 62, Dec. 2022, doi: 10.1016/j.ribaf.2022.101749.
- [13] S. H. Jung and Y. J. Jeong, "Examining stock markets and societal mood using Internet memes," *J Behav Exp Finance*, vol. 32, Dec. 2021, doi: 10.1016/j.jbef.2021.100575.
- [14] E. Teti, M. Dallochio, and A. Aniasi, "The relationship between twitter and stock prices. Evidence from the US technology industry," *Technol Forecast Soc Change*, vol. 149, Dec. 2019, doi: 10.1016/j.techfore.2019.119747.
- [15] P. Y. Hao, C. F. Kung, C. Y. Chang, and J. B. Ou, "Predicting stock price trends based on financial news articles and using a novel twin support vector machine with fuzzy hyperplane," *Appl Soft Comput*, vol. 98, Jan. 2021, doi: 10.1016/j.asoc.2020.106806.
- [16] W. Chen, C. K. Yeo, C. T. Lau, and B. S. Lee, "Leveraging social media news to predict stock index movement using RNN-boost," *Data Knowl Eng*, vol. 118, pp. 14–24, Nov. 2018, doi: 10.1016/j.datak.2018.08.003.
- [17] K. Liu, J. Zhou, and D. Dong, "Improving stock price prediction using the long short-term memory model combined with online social networks," *J Behav Exp Finance*, vol. 30, Jun. 2021, doi: 10.1016/j.jbef.2021.100507.
- [18] M. Billah, S. Waheed, and A. Hanifa, "Stock market prediction using an improved training algorithm of neural network," in *ICECTE 2016 - 2nd International Conference on Electrical, Computer and Telecommunication Engineering*, Institute of Electrical and Electronics Engineers Inc., Mar. 2017. doi: 10.1109/ICECTE.2016.7879611.
- [19] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," *Eur J Oper Res*, vol. 270, no. 2, pp. 654–669, Oct. 2018, doi: 10.1016/j.ejor.2017.11.054.
- [20] X. Zhang, Y. Zhang, S. Wang, Y. Yao, B. Fang, and P. S. Yu, "Improving stock market prediction via heterogeneous information fusion," *Knowl Based Syst*, vol. 143, pp. 236–247, Mar. 2018, doi: 10.1016/j.knosys.2017.12.025.
- [21] R. Ren, D. D. Wu, and D. D. Wu, "Forecasting stock market movement direction using sentiment analysis and support vector machine," *IEEE Syst J*, vol. 13, no. 1, pp. 760–770, Mar. 2019, doi: 10.1109/JSYST.2018.2794462.