



# Analyze News Effect on Trend Stock Price in Indonesia Based on Bidirectional-Long Short Term Memory

Muhammad Azriel Satriaman, Imelda Atastina\*

Informatics, Informatics, Telkom University, Bandung, Indonesia

Email: <sup>1</sup>muhazriel@students.telkomuniversity.ac.id, <sup>2\*</sup>imeldaatastina@telkomuniversity.ac.id

Correspondence Author Email: imeldaatastina@telkomuniversity.ac.id

Submitted: 13/06/2023; Accepted: 27/06/2023; Published: 29/06/2023

**Abstract**—News platforms such as BBC, UN News, and CNN are news sites that are very global, both globally and nationally. With this site, someone can find information in other countries or their country. The news contained on the BBC website can be analyzed using sentiment analysis. Sentiment analysis is carried out to see whether the news tends to be positive, negative, or neutral so that researchers or institutions can find out how the response of the news is to other sectors such as stocks in Indonesia. With the IDX website as a list of company shares in Indonesia, sentiment analysis can be carried out on news on the BBC website that can affect the rise or fall of stock prices in Indonesia using a combination of Word2Vec and the Bidirectional- Long Short Term Memory (BiLSTM) method. The BiLSTM method is an algorithm that has a function to process text data to predict the value of stock price trends by utilizing Word2Vec for word embedding of news. In this study, the dataset used is international news on the BBC website and historical stock prices of several companies on the IDX website. This study utilizes both methods to be able to predict stock price trends. By using 15.674 data, this study shows that the BiLSTM method has an average accuracy rate of 80.03%.

**Keywords:** Sentiment Analysis; BBC; BiLSTM; Word2Vec; IDX

## 1. INTRODUCTION

IHSG is an index that measures the performance of all listed stocks on the Main Board and Development Board of the Indonesian Stock Exchange (IDX). IHSG is also known as the Indonesia Composite Index (ICI) or IDX Composite. In general, each stock has different movements in one day. There are rising, falling, and stagnant. If these stocks are combined, the average movement of the shares is reflected in the IHSG. When the IHSG rises, it can conclude that most of the shares listed on the IDX also experience an increase.

The movement of stock prices are affected by several factors, which is news. News has direct impact on short-term stock price movements according to [1]. Details of news dissemination through the media can play an important role in influencing stock prices, which causes emotional investors to be easily influence by news [2]. For example, a news piece that reveals a corporate scandal can cause negative attitudes among investors, leading to a consequent drop in the stock price because it is conceivable for investors to sell shares. In addition, investors tend to take the same investment actions as others when facing new events based on the theory of the herd effect, which makes stock price changes more trackable [3]. News on the stock market is updated daily and the type of news investors are expose to varies. Investors are susceptible to these market emotions when making investment decisions. Therefore, how to analyze the effect of news on stock prices has become a research topic. Several significant factors are considere in this study: the value of sentiment analysis and bias value [4], [5].

The beginning of this study considered the news about stocks as an object of study as important information [6]. Research has shown that the news has an impact on stock prices. In detail, through comparing the different effects of different media on stock prices. In [7] found that news on the internet media will play a more significant role in the investor decision-making process. Moreover, [8] shows that when Wall Street pessimism rises, the overall market appreciation will fall the next day. Therefore, in this paper, we further analyze the extent to the influence of positive or negative news on stock price trends in Indonesia over a certain period.

Sentiment analysis is a process of analyzing and classifying a text with data obtained from various data sources such as the Internet and social media platforms [9]. Several studies on sentiment analysis and prediction of stock prices have been carried out by researchers using various methods, feature extraction, and different deep learning approaches. For example, research conducted by Jiawei and Murata analysed the sentiment of financial news and verified that market sentiment is a very important factor in stock trends forecasting [10].

Yahya Suryana and Tjong Wan Sen also conducted research on gold price movement by comparing 3 methods namely Naive Bayes, Support Vector Machine, and K-Nearest Neighbor. By the three methods, it was found that KNN had the best accuracy, namely 61.9%, then SVM with an accuracy of 57.5%, and finally, Naive Bayes had the lowest accuracy, namely 55.5% [11].

Li et al. used a deep neural network to predict stock prices using the Differential Privacy-inspired Long Short Term Memory (DP-LSTM) technique. The research combined dan integrated different news sources using a differential privacy mechanism and showed DP-LSTM predicted stock prices accurately [12].

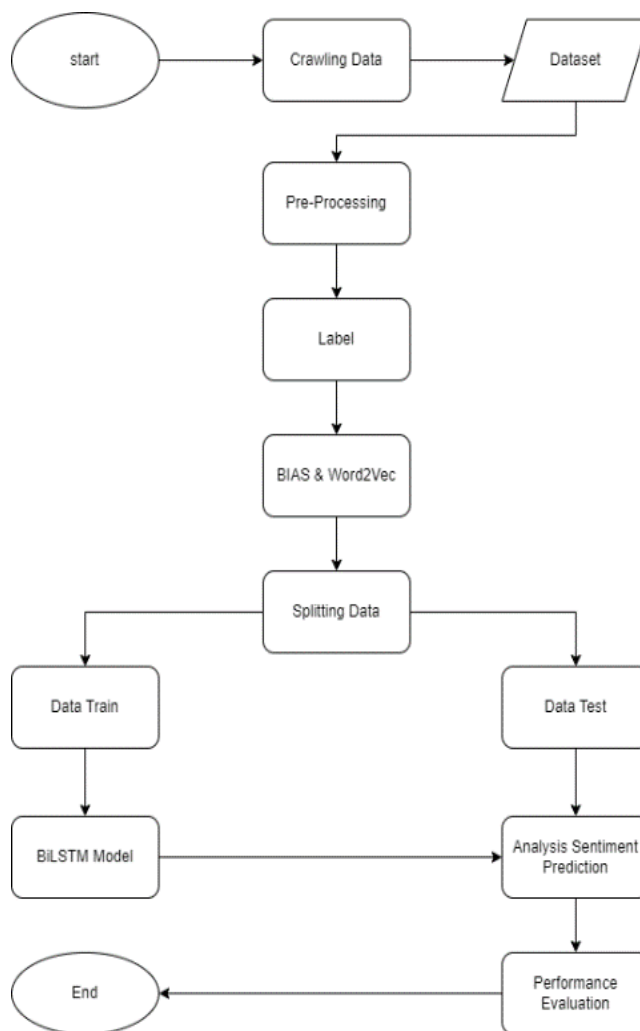
Based on the several studies that have been described, this research will apply the Bidirectional-Long Short-Term Memory method for the classification of news from BBC website and also the BIAS value from IHSG website

which will be used to predict stock price trends in several companies in Indonesia. This study will use word2vec for word embedding of news. Unlike other research, this research will focus on predicting the trend price change, not the price movement. From the results of the sentiment analysis, predictions will be made on stock price trends in several companies within a certain period.

## 2. RESEARCH METHODOLOGY

### 2.1 Architecture System

This stage is the design of the system that will be carried out to analyze sentiment on the news and BIAS using word2vec with the BiLSTM method. First, we gather information from the BBC and IHSG websites. Then, we organize and categorize the information. Next, we analyze the news articles using a special technique called word2vec, and BIAS in the stock data. After that, we divide the information into two groups: one for training and one for testing. The training group goes through a special process called BiLSTM. Finally, we make predictions about the BIAS. We also evaluate how well our process worked. The design of the system built is shown in Figure 1 below.



**Figure 1. Flowchart of the Architecture build system**

### 2.2 Data Crawling

Crawling is a data collection process to obtain information from a page on a BBC and IHSG website and then the information obtained will be stored offline or locally [13]. The crawling carried out on the BBC will use BeautifulSoup4, so that data will be obtained in the form of global English language news contained on the BBC website. The following Table 1 is the news contained in this study.

**Table 1. News**

News	Amount
Berita BBC	15674



### 2.3 Data Labelling

In this study, two labels will be used. The first is labeling for news contained on the BBC website with 2 labels, namely positive and negative. Then the second is BIAS, where this BIAS is the stock price when the news comes out and a few days after the news comes out. Therefore there are 3 labels, namely 0 for BIAS values less than 0, label 1 for BIAS values 0.1-0.9, and label 2 for BIAS values equal to or more than 1. Examples of labeling are shown in Table 2 and Table 3.

**Table 2.** News Labeling

News	Label
forget	Negative
bombing	Negative
cost	Negative
star	Positive
joining	Positive

**Table 3.** BIAS Labeling

News	Label
1,9607	2
0,3278	1
-0,3257	0
-0,9803	0
0,6514	1

### 2.4 Preprocessing Data

Most of the research on sentiment analysis has focused on user-generated text. Data pre-processing are steps that function to process data in the form of a collection of texts that are still biased or unstructured and need to be cleaned and normalized into data that has a good quality[14]. Pre-processing in this study will be carried out through several stages, namely data cleaning, case folding, removing punctuation, stopwords, tokenizing, and lemmatizer.

The first step, Data cleaning is a process that processes data by removing punctuation marks, URLs, or characters other than text. The second step, Case folding is a process that functions to make all letters the same, for example, lowercase letters by using lowercase [15]. The third step, removing punctuation is a process to remove punctuation marks. The fourth step, Stopwords is a processing function to delete words that do not have a function but do not have a clear meaning. The fifth step, Tokenizing is the process of separating each word in a sentence. Finally, the sixth step, Lemmatizer is a process that functions to turn words into sentences that have the same meaning. The following is an example of the pre-processing carried out which can be seen in Table 4.

**Table 4.** Pre-processing Data

Pre-processing	News	Result
Data Cleaning	<a href="https://www.bbc.co.uk/news/world-europe-60641873">https://www.bbc.co.uk/news/world-europe-60641873</a> , <a href="https://www.bbc.co.uk/news/world-europe-60641873?at_medium=RSS&amp;at_campaign=KARANGA">https://www.bbc.co.uk/news/world-europe-60641873?at_medium=RSS&amp;at_campaign=KARANGA</a> , "Jeremy Bowen was on the frontline in Irpin, as residents came under Russian fire while trying to flee."	Jeremy Bowen was on the frontline in Irpin, as residents came under Russian fire while trying to flee.
Case Folding	Jeremy Bowen was on the frontline in Irpin, as residents came under Russian fire while trying to flee.	jeremy bowen was on the frontline in irpin, as residents came under Russian fire while trying to flee.
Remove Punctuation	jeremy bowen was on the frontline in irpin, as residents came under Russian fire while trying to flee.	jeremy bowen was on the frontline in irpin as residents came under Russian fire while trying to flee
Stopwords	jeremy bowen was on the frontline in irpin, as residents came under Russian fire while trying to flee	jeremy bowen frontline irpin resident came russian fire trying flee
Tokenizing	jeremy bowen frontline irpin resident came russian fire trying flee	['jeremy', 'bowen', 'frontline', 'irpin', 'resident', 'came', 'russian', 'fire', 'trying', 'flee']

---

Lemmatizer	[‘jeremy’, ‘bowen’, ‘frontline’, ‘irpin’, ‘resident’, ‘came’, ‘russian’, ‘fire’, ‘trying’, ‘flee’]	jeremy bowen frontline irpin resident came russian fire trying flee
------------	--	--

---

## 2.5 Word2Vec

Word2Vec is a shallow neural network model that converts word representations which are combinations of alphanumeric characters into vectors. The vector representation has a relationship property to related words through the training process. Word2Vec is used as a model for processing news data. Word2Vec is a group of related models used to generate word embedding. It is efficient and easy so it is widely used[16], [17]. There are two models in word2vec: skip-grams and continuous bag of words (CBOW). In research[18], it is said that the skip-gram model is an efficient method for researching word vectors in large amounts of unstructured text. Meanwhile, the CBOW model predicts words based on the entire context of the word.

## 2.6 BIAS

The term BIAS refers to a calculated measure of the deviation between a stock's price and its moving average over a certain period. The bias calculation helps identify the extent to which a stock's price deviates from its moving average trend. This provides insight into possible price reversals or rebounds caused by significant deviations from the moving average[19]. Bias can help assess the reliability of stock prices that remain within their normal range of fluctuations and indicate potential continuation patterns. With the following calculation formula:

$$BIAS = \frac{P - P_i}{P_i} * 100\% \tag{1}$$

P represents the closing price on the day the news occurred, and P (i) represents the average price of the stock after the day the news occurred. BIAS is here used as an evaluation index to be able to find out the performance of existing stocks because bias is considered to be used as a benchmark to be able to find out trends in changes in the stock price of a company.

## 2.7 Bidirectional Modeling

Bidirectional Long Short – Term Memory is one of the most frequently used variants of Long Short Term Memory. Forward input and backward input are 2 types of input that are included in the Bidirectional Long Short–Term Memory architecture[20]. The outputs of this architecture are usually bundled together. With this architectural layer, the model can study past and future data for each input sequence.

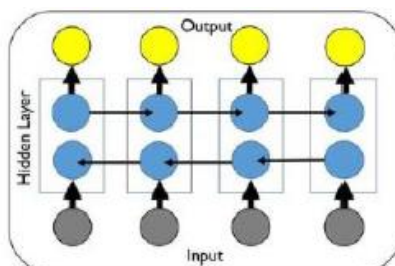


Figure 2. BiLSTM Architecture

Bidirectional LSTM makes use of prior information and subsequent information by processing data from both directions. The forward layer functions to represent previous information, and the backward layer functions to represent subsequent information.

## 2.8 Evaluation Metrics

The last stage in this research is the evaluation of the performance of the system being built. The evaluation matrix is used to assess the performance or effectiveness of a particular system. This evaluation produces a value, namely accuracy, and f1-score. F1-score is used to measure the performance of a model by having the best results from a specified method[21]. In addition to the f1-score, results in the form of precision and recall are also generated. Precision functions for the ratio of the positive prediction results of the total positive predictions and recall functions for the ratio of the predicted results of the overall positive predictions. Accuracy is the percentage of input that is successfully predicted by BiLSTM. The following is the calculation of the f1-score formula, precision, and recall.

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$



$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

### 3. RESULT AND DISCUSSION

#### 3.1 Evaluation

This study aims to apply a sentiment analysis system using the Bidirectional-Long Short Term Memory (BiLSTM) method to news labels and BIAS. There are 3 scenarios to achieve this goal, for all scenarios, 5 trials were carried out by taking the average. Scenario 1 is to determine the data splitting ratio by conducting a ratio comparison test. After that, Scenario 2 is to test using more data, namely 15,000. Scenario 3 is to determine the best timeframe to be able to predict stock price trends, namely 5 days and 20 days. Then Scenario 4 is to compare the results of all scenarios with different types of stocks where there are 3 stocks in this data, namely BRI, Unilever, and TOTO shares.

#### 3.2 Testing scenario and testing result

The first scenario was carried out to find the best baseline by comparing the splitting ratio of training data and testing data with amount of 5472 data. The splitting ratios used were 90:10, 80:20, 70:30, and 60:40. The ratio of 90:10 indicates the dataset is splitting into 90% train data and 10% test data. In this test, the most optimal results were obtained at a splitting ratio of 80:20 so that it is used for the next scenario. Here several ratio tests are carried out to get good enough accuracy so that the predictions obtained when predicting stock price trends can be maximized. And also other ratios can be seen in Table 5.

**Table 5.** Ratio splitting data

Ratio	Accuracy (%)
90:10	71,37
<b>80:20</b>	<b>74,54</b>
70:30	72,18
60:40	69,20

The second scenario is to test using more data with a total of 15 thousand. Here 5 trials were carried out with significant data additions, namely 8.000, 10.000, 12.000, 14.000, and 15.674. In this test, it is proven that with a total of 15.674, it produces the most optimal value of 78.29% and will be used for the next scenario. Shown in Table 6.

**Table 6.** Data Amount

Data Amount	Accuracy (%)
8.000	75,79
10.000	75,98
12.000	76,20
14.000	77,82
<b>15.674</b>	<b>78,29</b>

Furthermore, the third scenario is to determine the best timeframe to be able to predict stock price trends, namely 5 days and 20 days. Also in this scenario this research already use 80:20 ratio and 15.674 amount of data. In this scenario will be use 5 days period and 20 days period because based on few research predicted prices will be optimal in first week and after two week. In the experiment here you can see a scenario for predicting stock price trends is more optimal on 5 days with 7 correct results out of 10 with an average accuracy rate of 79.56% and for a comparison of the two scenarios can be seen in Tables 7 and 8.

**Table 7.** BRI

Date	5 Days	
	Predicted BIAS	Actual BIAS
01-04-2022	2	2
04-04-2022	2	2
05-04-2022	1	2
11-04-2022	0	0
18-04-2022	0	0
20-04-2022	0	0
09-05-2022	1	2



11-05-2022	2	2
20-05-2022	0	1
06-06-2022	0	0

**Table 8. BRI**

Date	20 Days	
	Predicted BIAS	Actual BIAS
01-04-2022	1	0
04-04-2022	0	0
05-04-2022	1	2
11-04-2022	1	1
18-04-2022	1	2
20-04-2022	1	1
09-05-2022	1	2
11-05-2022	0	1
20-05-2022	0	1
06-06-2022	0	0

The final scenario is a comparison of stock price trends with 3 companies namely BRI, Unilever, and TOTO.. In the last scenario, a splitting data ratio of 80:20 is used with a total of 15,674 data and a period of 5 days. From this, it can be seen that the trend of stock prices with a period of 5 days gives more optimal results even though it has compared several existing stocks. And it also has a quite different level of accuracy, namely for BRI with an average accuracy rate of 79.56%, for TOTO with an average accuracy of 78.74%, and for Unilever with an average accuracy of 80.03%. In this study, different stocks were also tested for 20 days, which can be seen in Table 9 and Table 10.

**Table 9. Comparison BIAS 3 Stock 5 Days**

Date	5 Days					
	BRI		TOTO		UNVR	
	Predicted BIAS	Actual BIAS	Predicted BIAS	Actual BIAS	Predicted BIAS	Actual BIAS
01-04-2022	2	2	2	2	0	1
04-04-2022	2	2	1	1	0	0
05-04-2022	1	2	1	0	1	2
11-04-2022	0	0	1	1	1	1
18-04-2022	0	0	1	1	1	1
20-04-2022	0	0	1	2	0	0
09-05-2022	1	2	0	0	1	0
11-05-2022	2	2	1	1	2	1
20-05-2022	0	1	1	0	2	2
06-06-2022	0	0	0	0	0	0

**Table 10. Comparison BIAS 3 Stock 20 Days**

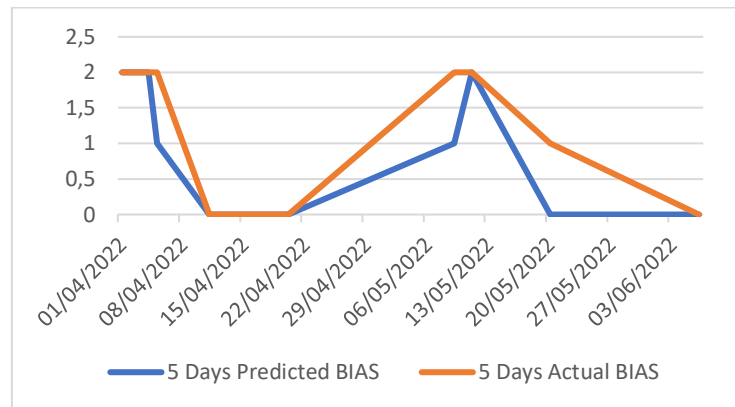
Date	20 Days					
	BRI		TOTO		UNVR	
	Predicted BIAS	Actual BIAS	Predicted BIAS	Actual BIAS	Predicted BIAS	Actual BIAS
01-04-2022	1	0	1	2	0	1
04-04-2022	0	0	1	1	0	0
05-04-2022	1	2	1	0	1	1
11-04-2022	1	1	2	2	1	0
18-04-2022	1	2	1	2	1	0
20-04-2022	1	1	1	1	2	2
09-05-2022	1	2	0	0	1	1
11-05-2022	0	1	2	1	2	1
20-05-2022	0	1	2	1	2	2
06-06-2022	0	0	0	2	0	0

In this study, testing was carried out using 4 scenarios. By determining the best ratio and also increasing the amount of data to have more certain results. This study also uses two time periods, namely 5 days and 20 days. Finally, this research compares the stock price trends of 3 companies, namely BRI, TOTO, and Unilever.

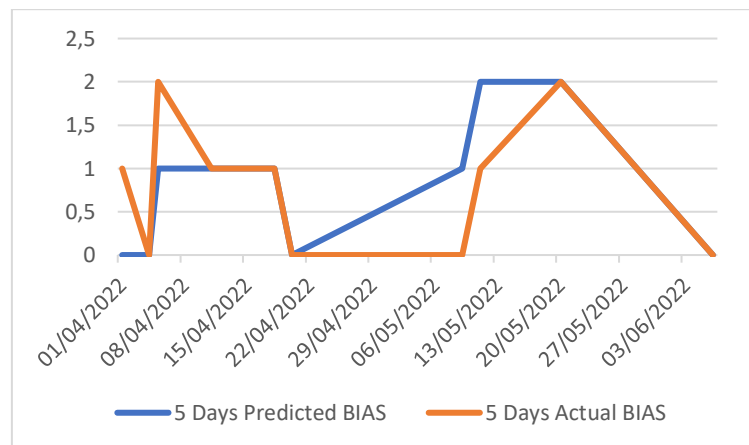
Based on the four scenarios that have been tested, the best performance results are using a ratio of 80:20 with an accuracy of **74.54%**. Followed by adding the amount of data up to 15 thousand with an accuracy of **78.29%**. These



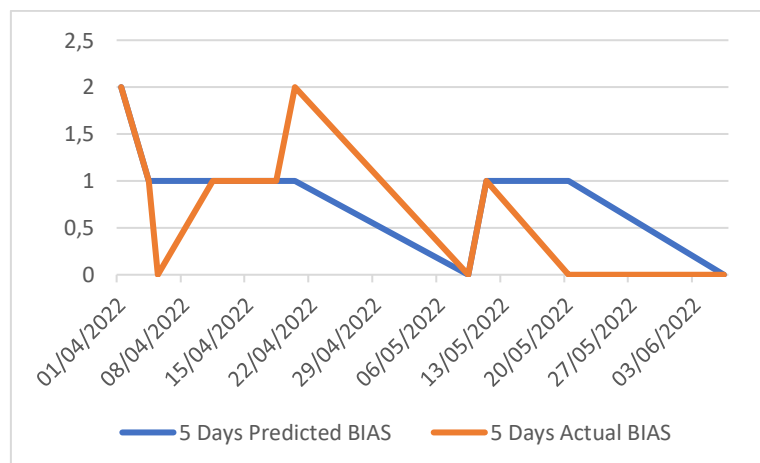
results prove that combining these 2 scenarios produces the best performance. And the last scenario is using 5 days with a greater degree of truth to the original BIAS. And it also has a quite different level of accuracy, namely for BRI with an average accuracy rate of **79.56%**, for TOTO with an average accuracy of **78.74%**, and for Unilever with an average accuracy of **80.03%**.



**Figure 3. BRI Prediction**



**Figure 4. Unilever Prediction**



**Figure 5. TOTO Prediction**

**3.3 Discussion**

Based on a model that has been made, shown splitting data ratio of 80:20 is given the highest accuracy among the others. Compared to the 90:10 splitting data ratio has slightly less accuracy because of too many unbalance labels than 80:20. This model also shown with 15.674 amount of data with an 80:20 ratio will be given the optimum accuracy. And the 5-day time period gives better results because the stock price when the news comes out and the next five days are not much different than a 20-day time period. Few stocks maybe show a slightly same pattern because there is an effect on the stock within the news.

## 4. CONCLUSION

The study uses sentiment analysis to analyze the impact of news on the BBC website and the impact of BIAS on IHSG to predict stock price trends. Experimental results show that the combined use of the BiLSTM technique and word2vec as word embeddings in messages can predict stock price trends. There are some highlights from this experiment that show that the data ratio split of 80:20 is better with 15,000 data records and 5 days time period. The final results of this study using the BiLSTM method showed an accuracy of 79.56% for BRI, 80.03% for Unilever, and 78.74% for TOTO. Based on the results obtained, by combining Word2Vec and BiLSTM, it is possible to develop a system that predicts stock price trends and performs fairly accurate sentiment analysis. For further investigation, it is recommended to conduct research using deep learning optimization techniques for better performance results.

## REFERENCES

- [1] G. Serafeim and A. Yoon, "Stock price reactions to ESG news: the role of ESG ratings and disagreement," *Review of Accounting Studies*, 2022, doi: 10.1007/s11142-022-09675-3.
- [2] A. Lazzini, S. Lazzini, F. Balluchi, and M. Mazza, "Emotions, moods and hyperreality: social media and the stock market during the first phase of COVID-19 pandemic," *Accounting, Auditing and Accountability Journal*, vol. 35, no. 1, 2022, doi: 10.1108/AAAJ-08-2020-4786.
- [3] W. Long, L. Song, and Y. Tian, "A new graphic kernel method of stock price trend prediction based on financial news semantic and structural similarity," *Expert Syst Appl*, vol. 118, 2019, doi: 10.1016/j.eswa.2018.10.008.
- [4] N. Seong and K. Nam, "Predicting stock movements based on financial news with segmentation," *Expert Syst Appl*, vol. 164, 2021, doi: 10.1016/j.eswa.2020.113988.
- [5] Y. Ren, F. Liao, and Y. Gong, "Impact of News on the Trend of Stock Price Change: an Analysis based on the Deep Bidirectional LSTM Model," *Procedia Comput Sci*, vol. 174, pp. 128–140, Jun. 2020, doi: 10.1016/j.procs.2020.06.068.
- [6] A. Sao, A. Kumar, and S. Singh, "Role of Media in Covering Stock Market and its Impact on Investors Behavioral Finance," *ANVESHAK-International Journal of Management*, vol. 10, no. 2, 2021, doi: 10.15410/aijm/2021/v10i2/166227.
- [7] B. Shantha Gowri and V. S. Ram, "Influence of news on rational decision making by financial market investors," *Investment Management and Financial Innovations*, vol. 16, no. 3, 2019. doi: 10.21511/imfi.16(3).2019.14.
- [8] J. D. Aromi, "Linking words in economic discourse: Implications for macroeconomic forecasts," *Int J Forecast*, vol. 36, no. 4, 2020, doi: 10.1016/j.ijforecast.2019.12.001.
- [9] F. F. Rachman and S. Pramana, "Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter," *Health Information Management Journal*, vol. 8, no. 2, 2020.
- [10] X. Jiawei and T. Murata, "Stock market trend prediction with sentiment analysis based on LSTM neural network," in *Lecture Notes in Engineering and Computer Science*, 2019.
- [11] Y. Suryana and T. W. Sen, "The Prediction of Gold Price Movement by Comparing Naive Bayes, Support Vector Machine, and K-NN," *JISA(Jurnal Informatika dan Sains)*, vol. 4, no. 2, 2021, doi: 10.31326/jisa.v4i2.922.
- [12] X. Li, Y. Li, H. Yang, and X.-Y. Yang Liuqing and Liu, "DP-LSTM: Differential Privacy-inspired LSTM for Stock Prediction Using Financial News," *arXiv [q-fin.ST]*. 2019.
- [13] A. Afandika, *Analisis Sentimen Teks Bahasa Indonesia Pada Media Sosial Menggunakan Algoritma Convolutional Neural Network (Studi Kasus: Operator Telekomunikasi)*, vol. 2, no. 2, 2018.
- [14] H. T. Duong and T. A. Nguyen-Thi, "A review: preprocessing techniques and data augmentation for sentiment analysis," *Comput Soc Netw*, vol. 8, no. 1, 2021, doi: 10.1186/s40649-020-00080-x.
- [15] D. T. Hermanto, A. Setyanto, and E. T. Luthfi, "Algoritma LSTM-CNN untuk Binary Klasifikasi dengan Word2vec pada Media Online," *Creative Information Technology Journal*, vol. 8, no. 1, 2021, doi: 10.24076/citec.2021v8i1.264.
- [16] Quoc Le, Tomas Mikolov, "Distributed Representations of Sentences and Documents Quoc," *Google Inc*, 1600 Amphitheatre Parkway, Mountain View, CA 94043 QVL@GOOGLE.COM, vol. 32, 2021.
- [17] J. Bhatta, D. Shrestha, S. Nepal, S. Pandey, and S. Koirala, "Efficient Estimation of Nepali Word Representations in Vector Space," *Journal of Innovations in Engineering Education*, vol. 3, no. 1, 2020, doi: 10.3126/jjee.v3i1.34327.
- [18] B. Jang, I. Kim, and J. W. Kim, "Word2vec convolutional neural networks for classification of news articles and tweets," *PLoS One*, vol. 14, no. 8, 2019, doi: 10.1371/journal.pone.0220976.
- [19] S. Huang, Y. Huang, and T. C. Lin, "Attention allocation and return co-movement: Evidence from repeated natural experiments," *J financ econ*, vol. 132, no. 2, 2019, doi: 10.1016/j.jfineco.2018.10.006.
- [20] Z. Ferdoush, B. N. Mahmud, A. Chakrabarty, and J. Uddin, "A short-term hybrid forecasting model for time series electrical-load data using random forest and bidirectional long short-term memory," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 1, 2021, doi: 10.11591/ijece.v11i1.pp763-771.
- [21] K. L. Kohsasih, M. Dipo, A. Rizky, T. Fahriyani, V. Wijaya, and R. Rosnelly, "Analisis Perbandingan Algoritma Convolutional Neural Network Dan Algoritma Multi-Layer Perceptron Neural Dalam Klasifikasi Citra Sampah," *Jurnal Technology Informatics dan Computer System*, vol. 10, no. 2, 2021.