JOIN (Jurnal Online Informatika)
p-ISSN: 2528-1682, e-ISSN: 2527-9165
Volume 10 Number 1 | June 2025: 165-176
DOI: 10.15575/join.v10i1.1422

Forecasting Shallot Prices in Indonesia Using News-Based Sentiment Indicators

Atikah Salsabila[1], Rani Nooraeni[1]
[1] Department of Statistical Computing, Politeknik Statistika STIS, Jakarta, Indonesia

Article Info

Article history:

Received November 11, 2023
Revised September 21, 2024
Accepted November 15, 2024
Published May 10, 2025

Keywords:

Sentiment analysis
Text mining
Price Forecasting
Online news
Time Series

ABSTRACT

The volatile price changes of shallots are a challenge in controlling their prices. The fluctuation in the price of shallots is always reported in the media because it affects people's lives. The news is released online via the internet and has beneficial information so it can be utilized. This study aims to provide a comparative analysis of forecasting models for shallot prices in Indonesia, evaluating the impact of using the most effective sentiment indicators derived from four lexicon-based methods. Data were collected by scraping method on three news portals and one food price information source website during the period from 2020 to 2023. The correlation and causality analysis was conducted to determine the relationship between food prices and sentiment indicators that was obtained using four sentiment analysis methods. The selected sentiment indicators for each day were used as an additional variable in forecasting using ARIMA, SARIMA, and BSTS models. The results showed that the use of news sentiment could reduce RMSE, MAPE, and MAE in forecasting shallot food prices.

Corresponding Author:

Rani Nooraeni,
Department of Statistical Computing, Politeknik Statistika STIS,
Jl. Otto Iskandardinata No.64C, Jakarta 13330
Email: raninoor@stis.ac.id

1. INTRODUCTION

The price changes of agricultural commodities have a significant impact on inflation in Indonesia [1]. Predicting the price of agricultural commodities accurately is very important for accomplishing government macroeconomic regulation effectively [2]. Besides, forecasting of the prices is needed to act as an early warning to people working in this industry, such as farmers, agricultural managers, farmworkers and so on, to forecast the prices in that state in the future to maximize their income and profit as well as avoid misinformation and further losses [3]. Shallot is one of the commodities include in volatile foods commodities which are commodities whose price movements are dominantly affected by shocks in food stuffs [4]. According to Statistics of Horticulture 2022 and 2021 released by BPS RI, Shallot has great contribution to horticulture production and inflation level which the total consumption reached 831,14 tons, higher than garlic, chillies, potato, tomato and carrot [5]. Moreover, research by Helbawanti et al [1] showed price of shallot has an impact on inflation and deflation in Indonesia since 2017 until 2021. Likewise, Strategic Plans (RENSTRA) Ministry of Agriculture for 2020-2024 mentioned that shallots are one of necessary commodities that the price stabilization is needed to maintain food security. In order to minimize losses due to price volatility, ensuring shallot price stability is important by strengthening the transparency and credibility of information provided in domestic markets [6].

Many researches studies on forecasting agricultural prices have already been conducted both regionally [7]and nationally [8], [9]. However, accuracy is still an issue due to the price fluctuations. The factors affecting the prices of shallots are very complex, and it is rather difficult to provide a comprehensive list of factors based solely upon previous research or researchers' opinions. Especially, some factors such as the weather and policy changes are difficult to quantify [10], [11]Therefore, an effective way to quantify influential factors reasonably is needed for predicting the price changes[11] Besides, the rise and fall of prices that have a direct impact on the Indonesian people encourages the media to always report the latest conditions every significant change.

Nowadays, most forms of media have an online presence and produce huge volumes of data. This data contains information in the form of opinions and sentiment about financial markets and the economy, which may not yet be reflected in macroeconomic variables [12]. Current developments in natural language processing (NLP) help in quantifying such information, which have potential to be used in forecasting models to make less mistakes predictions about the variables of interest [13]. All news articles about events that have a significant impact on the price of shallots in the future such as extreme weather [3], [11] crop damage[13], [14] pest or disease attacks[11] disrupted distribution channels[13], changes in commodity prices related to changes in food prices and etc., can provide very useful additional information[11], [15]. Relevant and valuable information from a number of online news articles as an unstructured data source can be used to improve shallot price predictions.

The news articles about changes in shallot prices, followed by expert opinions and other relevant facts, provide beneficial additional information [16]. The sentiment in each article and the frequency of those words appearing are valuable for understanding the shocks from shallot price changes. Forecasting is generally a model that combines "hard" and "soft" information. "Hard" information is an objective variable that can be measured directly in numbers, while "soft" information is a more subjective variable in text form [17], [18]. Aprigliano et al. [19] and Barbaglia et al. [20] show that news-based text data, especially in the form of sentiment indicators, can improve macroeconomic forecasts over and above hard economic indicators. Sentiment analysis of online news related to shallots will be very useful for extracting information about price changes, either positive, negative, or neutral sentiment. The indicators from the sentiment analysis help address the challenges of varying seasonal patterns annually in forecasting shallot prices. By utilizing time series and historical data, short-term predictions are made to account for frequent disruptions in agricultural commodities that often occur in a short period of time [21]. It is proven by Eugster (2024) [22] that, the use of sentiment in this news can predict the price index more accurately than the statistics traditional model.

Research by Pratap et al. [13] proposed news-based sentiment indicators using Loughran & McDonald dictionary for food inflation forecasting in India and found that they improved the ARIMA and SARIMA model performance. Besides, J, Li et al. [18], stated sentiment indicator built using Henry dictionary effectively captures the related factors and helped all models conducted for oil price trend prediction have better performance. While in Indonesia, A study conducted by Nagib and Husodo in 2022 [23] explored the relationship between news sentiment, news intensity, and price movement of LQ45 index in Indonesia using Loughran&McDonald dictionary and TextBlob Library. The findings revealed that the news sentiment has a relatively strong correlation with the index movement, and the effect of sentiment on index returns is more significant when news intensity increases. Thus, the sentiment of online news for additional information has promising potential to be used in forecasting shallot prices in Indonesia for better accuracy in the future.

This study aims to provide a comparative analysis of forecasting models for shallot prices in Indonesia, evaluating the impact of incorporating versus excluding the most effective sentiment indicators derived from four lexicon-based methods. The contribution of this research is to verify the predictive power of sentiment indicators for the change of shallot prices in order to get a more accurate prediction of the future price for being utilized in maintaining agricultural inflation in Indonesia.

2. METHOD

This research was conducted through several stages, which are presented in Figure 1. After data is collected and the preparation is conducted, the sentiment indicators are constructed using four lexicon-based sentiment analysis from the relevant information of online news releases. Among those four kinds, the one that had the best relationship with the prices of shallots, is chosen to be an additional variable in the forecasting model. The relationships are analysed by the correlation, the graphs and the causality test. The Granger causality analysis is conducted to ascertain whether the sentiment of online news articles are helpful for capturing the change in prices. Popular time-series forecasting models are employed to see how the sentiment indicators impact the performance of models such as Autoregressive Integrated Moving Average (ARIMA), seasonal ARIMA [13], and Bayesian Structural Time Series (BSTS) [24].

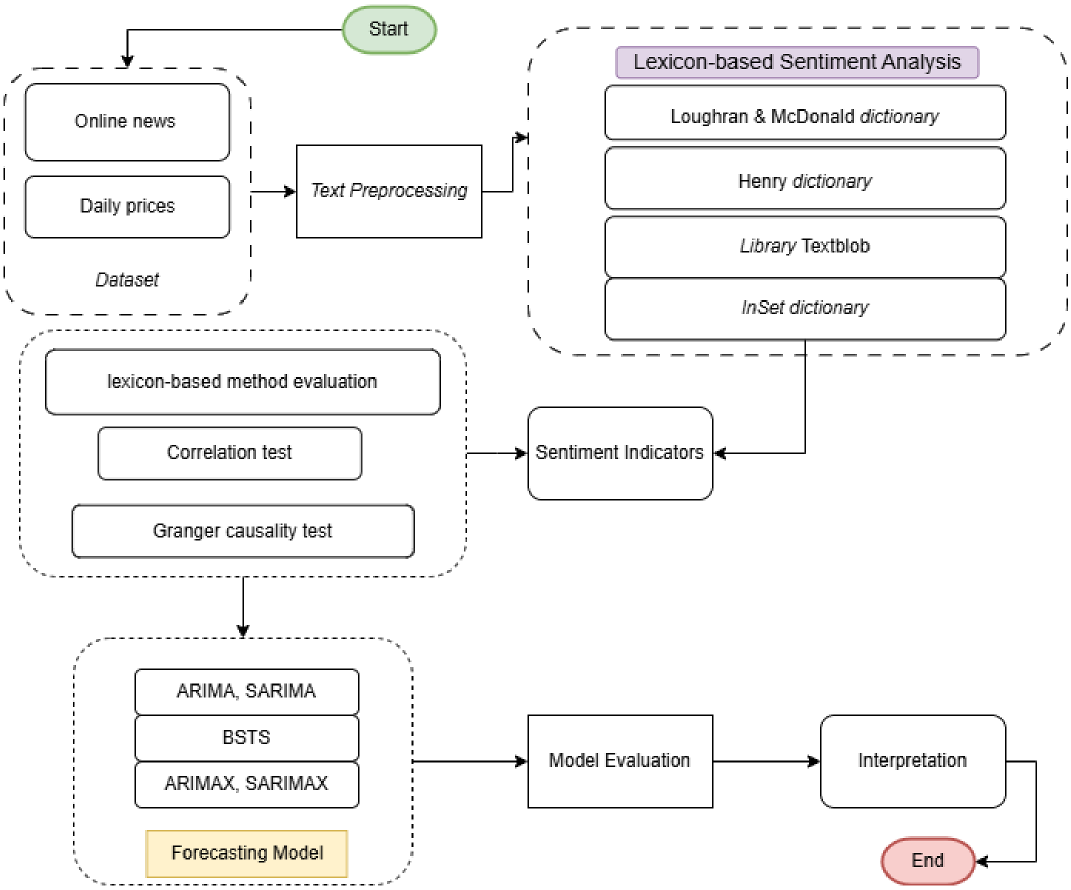


Figure 1. Research flowchart

2.1. Data Collection

The data used is obtained using the scraping method on two different data sources during the period after the existence of COVID-19 from 2020 to 2023[23]. The first dataset is the national daily price data of shallot commodities on the "Sistem Pemantauan Pasar Kebutuhan Pokok" (SP2KP) website. The data contained the date and the retail selling price of shallots per kilogram in the national average released by the Ministry of Trade. The data collected during the period from January 1, 2020, to December 31, 2023. The second dataset contains a list of national online news obtained from scraping three different sources using the keyword "shallot price". The three selected sources consist of two general news portals from the study [16], which have a search feature that is equipped with special filters for the economic/financial category, namely, the DetikFinance category from the Detik portal and okeFinance from okezone, and the third source is a special economic news portal, cnbcindonesia. All portals are publicly accessible websites. It is stated on the copyright and terms of service page of the website, respectively, that the platforms permit the use of publicly available content for academic research purposes under the condition that the content is not republished or used for commercial purposes [25], [26], [27].The news data obtained was 1412 national news articles from January 1, 2020, to December 31, 2023. Data were collected over a short period in order to avoid the shift trends due to the COVID-19 pandemic that may influence forecasting models. This range of time leads to a lack of data availability compared to previous related studies and limits the amount of news obtained. This research avoids using more news portals as sources to reduce redundancy in repeated content. The size of the dataset is acknowledged as a limitation.

2.2 Online News Preparation

The news articles' data is not yet structured. So, the next stage is text preprocessing. The steps in this stage include case folding, cleaning, filtering, tokenization, stemming and translation. A filtering process needed to eliminate news that does not have the word "shallot" in the content of the news. Also, it eliminates news that is indicated to be not national news by removing the news located not in Jakarta by the location stated in the content. The stage of translation used the Deep translator to translate the words of each news that has been tokenized into English.

2.3 Sentiment Indicators Measurement

The sentiment indicators are constructed using four lexicon-based sentiment analysis from the relevant information of online news releases. Among those four kinds, the one that had the best relationship with the prices of shallots is chosen to be an additional variable in the forecasting model. Lexicon-based approaches operate on the assumption that the semantic orientation of a text is directly linked to the polarity of the words and phrases it contains [28]. Sentiment indicators can be obtained through sentiment analysis using the lexicon method with a dictionary-based approach. This approach utilizes pre-constructed dictionaries specific to a particular language and domain of interest and is based on formalisms and rules. While it may struggle to interpret sentences with highly complex structures, it is particularly effective in scenarios with limited data. This is especially relevant for complex but low-resource languages, including the Indonesian language [28]. For our context and purposes, two dictionaries that were designed specifically for analysing economic and financial texts were used. Additionally, for comparison, the general sentiment analysis method was used and based on the language of the raw data, which is the Indonesian language, this study added one local dictionary. Thus, sentiment scores obtained using four lexicon-based approach methods: the Loughran & McDonald lexicon dictionary [13], [29], Henry's lexicon dictionary[18], [30], the lexicon-rule library Textblob [23], [31], and the Indonesian InSet dictionary [32]. In the lexicon method, the sentiment score of each $n_i$news on the t ($S_{n,t}$) is calculated from the difference between positive and negative sentences, then divided by the sum of all the words in the news, which can be seen in equation (1). The scores consider negative words, such as "*cannot*", "*never*", "*without*", etc., by checking each positive word whether the previous one or two words contain a negative word or not.

$$S_{n,t} = \frac{jumlah\ kata\ positif - jumlah\ kata\ negatif}{total\ kata\ di\ berita} \quad (1)$$

For InSet dictionary which already weights each word, the score calculation follows the weight in the dictionary, by calculating the difference between the total weight of positive words and the total weight of negative words and then dividing by the total number of words in the news.

Then, for each method, the sentiment score on news released on the same day is aggregated into a new indicator variable called Net Sentiment Score (NSS) which is the total news on the day N_t and date t, based on equation (2). The results of the NSS are obtained on every date of news release.

$$NSS_t = \sum_{n=1}^{N_t} S_{n,t} \quad (2)$$

Then, the results of sentiment analysis are evaluated by comparing the category results of the sentiment score per method with correlation test, Granger-causality test, and visualization.

2.4 AutoRegressive Integrated Moving Average (ARIMA) Model

ARIMA demonstrates high accuracy in short-term forecasting. The standard notation for ARIMA is ARIMA(p,d,q), where the parameters are replaced with integer values to specify the particular model being used. The parameter p represents the number of lag observations included in the model, also known as the lag order. The parameter d denotes the number of times the raw observations are different, referred to as the degree of differencing. The parameter q defines the size of the moving average window, also called the order of the moving average [33].

A linear regression model is constructed with the specified number and type of terms, and the data is different to achieve stationarity, eliminating trends and seasonal structures that could negatively impact the regression model. A non-seasonal ARIMA consists of differencing, autoregressive, and moving average components. The ARIMA(p,0,q) model can be expressed as a linear equation:

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-k} \quad (3)$$

where ϕ_1, ϕ_2 are parameters for Auto Regressive and θ_q is the parameter of Moving Average. When we include either explanatory variable into the ARIMA model it is named ARIMAX model. ARIMA is suitable for short-term time series forecasting [9], [13], [18], [34]. The ARIMA model without sentiment indicators will be compared by using sentiment indicators as exogenous variables in the ARIMAX model.

2.5 Bayesian Structural Time Series (BSTS) Model

BSTS is a model with parameters that evolve over time, as it incorporates past information on variables, enabling it to achieve higher forecasting accuracy than ARIMA. [24], [33], [35].
One of the advantages of Bayesian modeling is to accounts for the uncertainty associated with parameter estimates and provides exact measures of uncertainty on the posterior distributions of these parameters, which is traditionally ignored in classical estimation models. In addition, Bayesian estimation and inference provide confidence intervals on parameters and probability values on hypotheses that are more in line with commonsense interpretations [33].
BSTS integrates Bayesian inference with structural time series by combining both approaches. A model can be constructed by incorporating a regression component into the widely used "basic structural model.". This model can be written

$$y_t = \mu_t + \tau_t \beta^{\tau} X_t + \varepsilon_t \quad (4)$$

Where y_t, μ_t, τ_t, ω_t, ξ_t and ε_t representing target time series, local linear trend component, seasonal component, regression component and observation error terms respectively.

$$\begin{aligned} \mu_t &= \mu_{t-1} + \delta_{t-1} + u_t \qquad u_t \sim N(0, \sigma_u^2) \\ \tau_t &= \Sigma_{s=1}^{S-1} \tau_{t-s} + w_t \qquad w_t \sim N(0, \sigma_\tau^2) \\ \delta_t &= \delta_{t-1} + v_t \qquad v_t \sim N(0, \sigma_v^2) \end{aligned} \quad (5)$$

S represent the number of season for y and τ_t denotes their joint contribution to the observed target time series y_t. As is common in Bayesian data analysis, forecasts from our model rely on the posterior predictive distribution. Given draws of model parameters and state from their posterior distribution, simulating from the posterior predictive distribution becomes straightforward. Let $\underline{y}$ denote the set of values to be forecast.

2.6 Seasonal ARIMA (SARIMA) Model

The ARIMA model was enhanced to the SARIMA model to better handle time series data by incorporating both seasonal and non-seasonal components for processing univariate time series data. The seasonal ARIMA *(p,d,q)*(P, D, Q)*^{S} are the non-negative integers for handling seasonality, X_t is the observed value at time *t,* and *s* is the number of periods per season. Equation (6) represents the general form of the SARIMA prediction model [36].

$$\varphi_P(G)\varphi_P(G^s)(1-G)^d(1-G^s)^D X_t = \gamma_q(G)\,\omega_Q(G^s)e_t \quad (6)$$

Where the coefficients $\varphi_P(G)$and $\gamma_q(G)$ are the orders of the non-seasonal AR and non-seasonal MA components' characteristic polynomials, and the polynomials $\varphi_P(G^s)$ and $\omega_Q(G^s)$are the seasonal autoregressive (SAR) and seasonal moving average (SMA) polynomials, respectively. The non-seasonal and seasonal time series are (1-G) and (1-G^s), respectively, which are the different components. In addition, d and D are the non-seasonal ARIMA model's ordinary differential terms and the SARIMA

p-ISSN: 2528-1682
e-ISSN: 2527-9165

model's seasonal differences terms, respectively; e_t is the prediction error; s is the duration of the seasonal pattern (e.g., s= 12 monthly series); and G is the backshift operator coefficient.

2.7 Model evaluation

This study applied a 75:25 split to the data to create separate training and testing groups. This data split was considered to strike a balance between having enough data for training while also having sufficient data to reliably test the model's performance. The data used are sentiment indicators and national price data for shallots in daily and weekly form for 4 years. The training set used data from January 2020 to December 2022, while the test set used data from January to December 2023. As a preliminary step, we used MinMax scaler to transform the data so that it ranged from 0 to 1. The performance of the models was then measured against the testing group. Root mean square error (RMSE), mean absolute percentage error (MAPE), and mean absolute error (MAE) were some of the performance measures we used to assess the models.

The RMSE serves as a measure to assess the average error's magnitude. The formula for RMSE as follows:

$$RMSE = \sqrt{\Sigma_{t=1}^{n} \left(\frac{\hat{y}_t - y_t}{n}\right)^2} \tag{7}$$

The MAE is a simple and intuitive metric that expresses the forecast error of the actual value,while the MAPE is the percentage of the actual value. The formula for MAE and MAPE follows:

$$MAE = \frac{1}{n} \Sigma_{t=1}^{n} \left| \frac{\hat{y}_t - y_t}{} \right| \tag{8}$$

$$MAPE = \Sigma_{t=1}^{n} \left| \frac{\hat{y}_t - y_t}{y_t} \right| \times 100 \tag{9}$$

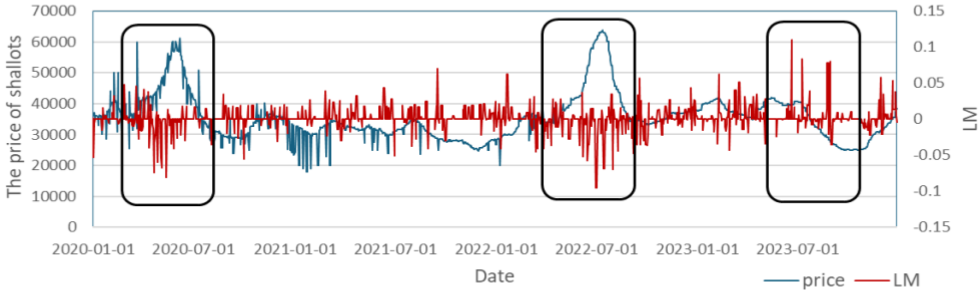
These measurements were used to evaluate the efficiency of the proposed sentiment indicator implementation technique and to make comparisons between the various models' outputs.

3. RESULT AND DISCUSSION

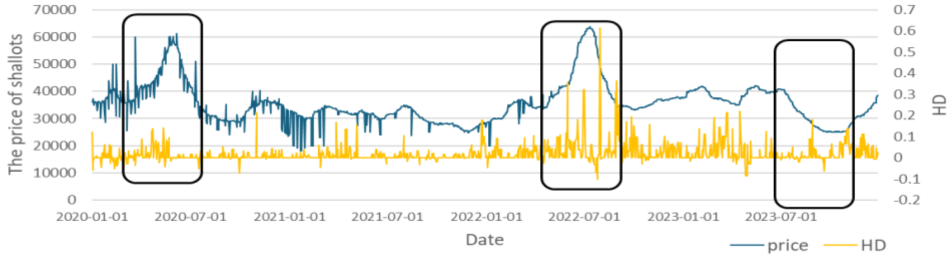
This section tries to analyze the performance of sentiment analysis techniques. Then we see an inverse relationship between news sentiment and the price of shallots. Finally, we try to see whether sentiment impacts the price and what model gives the best performance.

3.1. Performance comparison of four lexicon-based sentiment analysis for analyzing the relationship between the shallot price and sentiment score

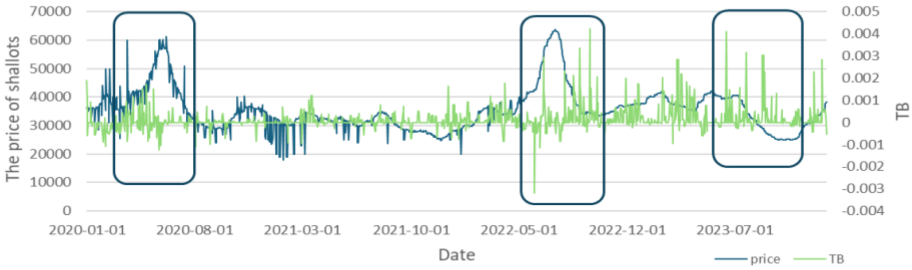
Sentiment analysis conducted for obtaining sentiment scores of each news. Then from the score of each news, the score of each day was obtained as sentiment indicators denoted by NSS (Net Sentiment Score). NSS was obtained from methods using a dictionary from Loughran & McDonald denoted by LM, dictionary from Henry denoted by HD, lexicon-rule library TextBlob denoted by TB, and an Indonesian lexicon dictionary named InSet denoted by IN. There has been no theory or previous research that examines which of the four methods is better, so this study explores the implementation of the four methods in the analysis of news sentiment related to the price of shallots. First, this study analyzes the relationship between price and the four methods by double-y axis visualization.



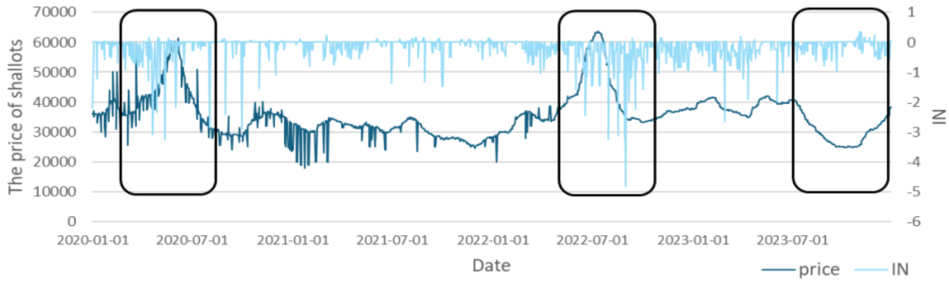
(a)



(b)



(c)



(d)

Figure 2. Relationship between major price shocks and sentiment indicators on all four lexicon approaches: (a) Loughran & McDonald's dictionary, (b) Henry's dictionary (c), TextBlob, (d) InSet dictionary

In Figure 2, it shows that each of the four methods has a different relationship to price. There were three spikes in the price of shallots during the 4 year research period, namely mid-2020, mid-2022, and late 2023. This significant change occurred due to production changes such as large harvests or crop failures affected by weather and changes in market demand, especially on some big events in Indonesia. For example, in mid-2020 and 2022, price spikes occurred due to crop failure problems due to weather and climate when demand increased because it was close to the Muslim holidays, namely Eid al-Fitr and Eid al-Adha, these weather changes also affected access to supply distribution to various places [37], [38]. In the sharp increase, the sentiment of LM, TB, and IN showed the opposite, namely a sharp decline even though with different days and ranges of numbers, while the sentiment of HD showed an increase just like the price. Likewise, when prices experienced a significant decline in mid-2023 due to the big harvest [5], [39], sentiment also showed significant changes although not as obvious as the sharp increase and the longer day gap. This is such a limitation of this research topic, it shows in 2021 where prices are quite stable and below the four-year average of the research period, price declines or stability

p-ISSN: 2528-1682
e-ISSN: 2527-9165

are less often reported than price increases. The increase in prices that coincides with the decline in sentiment in these three lexicon methods can be said to be reasonable because the high price is very likely to cause a lot of negative sentiment in related news. The negative relationship between sentiment and shallot prices is supported by the results of a correlation test between price and sentiment with four methods.

Based on Table I, the highest correlation found between LM and price as well as IN and price, both of which are negatively marked. In addition, it can also be seen that in years where prices have increased sharply such as 2020 and 2022, the correlation of sentiment with prices is higher than in 2021 and 2023 where prices tend to be low. Overall, the correlation value is negative, indicating that the decline in sentiment can be associated with the increase in onion prices.

Table 1. The result of the correlation test between price and sentiment with four methods

	LM	HD	TB	IN
Year 2020-2023	-0.174	0.090	-0.010	-0.216
Year 2020	-0.201	0.090	-0.017	-0.160
Year 2021	0.026	0.058	0.072	-0.036
Year 2022	-0.245	0.091	-0.069	-0.170
Year 2023	0.061	-0.092	0.006	-0.138

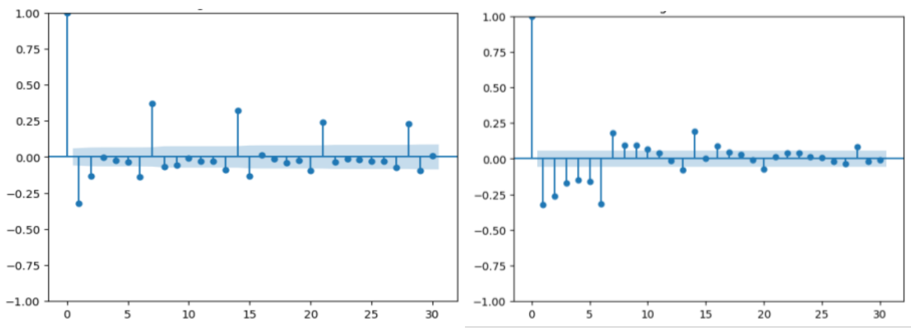
The relationship between sentiment and price was also proxied by looking at the cause-and-effect relationship between sentiment scores and daily prices. The Granger-causality test is conducted to see whether the news-based sentiment indicators help in capturing price changes. While H0 sentiment does not granger cause price and lag value of 1, the p-value of LM is 0.000, HD p-value is 0.853, 0.261 for TB, and the p-value of IN is 0.268. By using a significant level of 0.05, it is concluded that only LM does granger-cause of onion prices. Thus, the results of this test support the argument that news-based sentiment indicators help in capturing shallot price changes in the future. Therefore, sentiment with the lexicon method of the Loughran & McDonald dictionary is used for price forecasting modeling.

3.2. Performance comparison of forecasting models for shallot prices in Indonesia for analyzing the impact of using Sentiment Indicators

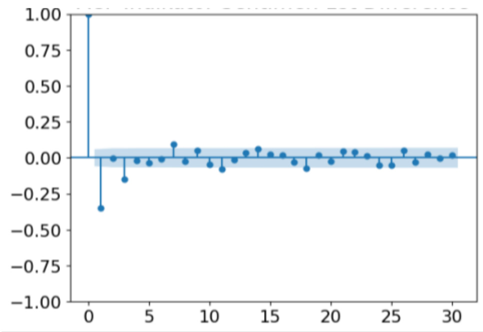
Forecasting models are built variously with and without sentiment indicators on three time series methods. The models will be compared in performance with the ARIMA model as the selected benchmark model. For ease of comparison, the evaluation results are shown in terms of performance relative to the ARIMA benchmark model. If the relative RMSE is less than 1, it indicates that the actual RMSE of a model is less than the benchmark model. The ARIMA model is determined using autoarima to obtain the best model by selecting the best model based on the smallest AIC value. The selected model is ARIMA(5,1,2) which has met the white noise property from the p-value of the Ljung-Box(LB). Then, we added selected sentiment indicator LM as exogenous variable to the same exact parameter model, The following is how ARIMAX (5,1,2) with LM can be written as an equation in mathematics:

$$Y_t' = 0.513Y_{t-1}' - 0.614Y_{t-2}' - 0.238Y_{t-3}' - 0.231Y_{t-4}' - 0.213Y_{t-5}' - 1.086\epsilon_{t-1} + 0.776\epsilon_{t-2} + 0.026LM + \epsilon_t \quad (10)$$

From the model above, it can be interpreted that the exogenous variable LM has a coefficient of 0.75 with a p-value of 0.023, indicating a statistically significant positive impact on the dependent variable at the 5% level and the beta-value which is 0.026 as coefficient of LM suggests a positive impact on the price [40]. It is known from the summary model that the estimate is reliable since the coefficient is greater than twice its standard error. The second model SARIMA is determined by comparing tentative model using range of parameter obtain from the ACF and PACF plot shown in Figure



(a)



(b)

Figure 3. (a) ACF and PACF Plot first *differencing* of the shallots price data (b) ACF plot first differencing of sentiment indicator LM.

Figure 3 shows the pattern on the plots in lag=7, 14, 21, etc. as well as the LM ACF plot. Therefore, the value of seasonal component m is set as 7. The parameters p and q range from 0 to 6, to make tentative models in order to obtain the best model by selecting the best model based on the smallest AIC value. The selected SARIMA model is SARIMA$(3,1,2)*(0, 1, 2)^{7}$. Similar to ARIMA model, selected sentiment indicator LM was added as an exogenous variable to the same exact parameter model. From the summary SARIMAX $(3,1,2)*(0, 1, 2)^{7}$ model , it can be interpreted that the exogenous variable LM has a small positive effect (coefficient = 0.019) on the dependent variable. The third model is BSTS, the selected BSTS model added LM as exogenous variable so the equation extend to:

$$Y_t = -3.599\times 10^{5}\text{LM} + \varepsilon_t \qquad \varepsilon_t \sim N(0, 0.044^2) \qquad (10)$$

The performance of the six models was being compared by the evaluation results. The results of the evaluation of the daily price forecasting model can be seen in Table 2.

Table 2.The Results of Evaluation of Daily Price Forecasting Model

Model	RMSE relative	MAPE	MAE
ARIMA	1.000	16.612%	4.898
ARIMA + LM	0.983	16.433%	4.871
SARIMA	1.057	17,468%	5.214
SARIMA + LM	1.031	17.112%	5.035
BSTS	0.998	16.599%	4.904
BSTS + LM	0.971	16.250%	4.838

Based on Table 2, it is concluded that the BSTS model with news sentiment has better performance than the benchmark model. In addition, it can also be seen that in each method, the model with the sentiment indicator LM performs better than the model without sentiment. Furthermore, forecasting modeling is carried out on weekly prices to see the influence of sentiment in price forecasting with more cumulative data. This is also based on the ACF and PACF plot mentioned above and the fact that news released in the same week tends to have similar content, so that both price and sentiment can be accumulated within a week. Similar procedure to daily price, the selected ARIMA model is ARIMA(0,0,4), the selected SARIMA model is SARIMA(*2,0,1)*(0, 0, 0*)^{52}, and the coefficient value of the LM in BSTS Model is 7.331×10^5. The seasonal component of SARIMA and SARIMAX was set as 52 because of the phenomena explained below Figure 2. The results of the evaluation of the weekly price forecasting model can be seen in Table 3.

Table 3. Results of Evaluation of Weekly Price Forecasting Model

Model	RMSE relative	MAPE	MAE
ARIMA	0.882	15.230%	4.744
ARIMA + LM	0.881	15.120%	4.729
SARIMA	0.871	14.917%	4.609
SARIMA + LM	0.870	14.879%	4.592
BSTS	0.974	16.217%	4.805
BSTS + LM	0.918	15.324%	4.597

Based on Table 3, using ARIMA(5,1,2) on daily prices as a benchmark model similar to table 2, it is concluded that the SARIMAX model has the best performance of all models conducted based on RMSE, MAPE, and MAE values. All results show that models with sentiment indicators LM on all models perform better than models without sentiment.

The predictive power of sentiment indicator proved by the comparison of accuracy evaluation of each model. This goes along with the study [13] which also found empirical evidence that sentiment indicators derived from news data enhance the accuracy of consumer price index forecasting of tomato, onion and potato commodities. Additionally. It is also similar to [13] that the analysis results show the inverse relationship between the sentiment indicator and the commodity prices. Research [11] also proposed a forecasting framework using massive online news headlines and found that variables built from news were effective in chinese soybean future price prediction. Therefore, it is suggested that this similar kind of research, using online news to capture the price change, probably applied in Asian countries especially agrarian countries with large populations with dependence on seasonal climate.

4. CONCLUSION

Research findings show that news-based sentiment indicators help in capturing shallot price changes in the future. Specifically, sentiment analysis using Loughran&McDonald's dictionary resulted in sentiment indicators that have closest relationship with the shallot prices than other methods using Henry's dictionary, TextBlob, and InSet dictionary. In addition, the analysis shows the inverse relationship between the shallot prices and sentiment indicators using Loughran&McDonald's dictionary. By using the news sentiment indicator as a variable input in a shallot price forecasting model in Indonesia, the results of the model estimation using ARIMA, BSTS, and SARIMA show that the news sentiment indicator data helps improve the performance of the price forecasting model

However, it should be noted that the study faced limitations in data availability, resulting in a smaller dataset compared to previous research. Consequently, a traditional model was chosen, as deep learning models are more suitable for larger datasets. To enhance future research, we recommend obtaining a more diverse and varied data source and employing more advanced model processing techniques for improved accuracy and performance.

REFERENCES

[1] O. Helbawanti, W. A. Saputro, and A. N. Ulfa, "Pengaruh Harga Bahan Pangan Terhadap Inflasi di Indonesia," 2021.
[2] Y. Zhang and S. Na, "A novel agricultural commodity price forecasting model based on fuzzy information granulation and MEA-SVM model," *Math Probl Eng*, vol. 2018, 2018, doi: 10.1155/2018/2540681.

[3] C. Z. Yuan and S. K. Ling, "Long Short-Term Memory Model Based Agriculture Commodity Price Prediction Application," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Aug. 2020, pp. 43–49. doi: 10.1145/3417473.3417481.
[4] Bank Indonesia, "Penjelasan Indikator, Data dan Informasi PIHPS (Pusat Informasi Harga Pangan Strategis) Nasional (Frequently Asked Questions). ." Accessed: Jun. 10, 2024. [Online]. Available: https://www.bi.go.id/hargapangan/Informasi/FAQ
[5] Badan Pusat Statistik, "Statistik Holtikultura 2022," Jakarta, 2023.
[6] dagmar Matoskova, "agricecon_age-201101-0005," *Agric. Econ. – Czech, 57, 2011 (1): 34–40*, vol. 57, pp. 34–40, 2011.
[7] J. A. Wiralodra, A. Hasyim, A. Rosyid, C. Dhanes, N. Viana, and W. A. Saputro, "Penerapan Model Box Jenkins (ARIMA) Dalam Peramalan Harga Konsumen Bawang Merah di Provinsi Jawa Tengah," 2021.
[8] M. A. Zen, S. Wahyuningsih, A. Tri, and R. Dani, "Aplikasi Pendekatan Agglomerative Hierarchical Time Series Clustering untuk Peramalan Data Harga Minyak Goreng di Indonesia (Application of Agglomerative Hierarchical Time Series Clustering Approach for Forecasting Cooking Oil Price Data in Indonesia)," *Seminar Nasional Official Statistics*, 2022.
[9] A. M. Windhy *et al.*, "Peramalan Harga Cabai Merah Indonesia: Pendekatan ARIMA Forecasting Indonesian Red Chilli Prices: The ARIMA Approach," 2021.
[10] Z. Wang, O. Kwon, and F. Liu, "Applying Keyword Analysis to Predicting Agriculture Product Price Index: The Case of the Chinese Farming Market," *Asia Pacific Journal of Business Review*, vol. 1, no. 1, pp. 1–22, Aug. 2016, doi: 10.20522/apjbr.2016.1.1.1.
[11] J. Li, G. Li, M. Liu, X. Zhu, and L. Wei, "A novel text-based framework for forecasting agricultural futures using massive online news headlines," *Int J Forecast*, vol. 38, no. 1, pp. 35–50, Jan. 2020, doi: 10.1016/j.ijforecast.2020.02.002.
[12] S. Tilly, M. Ebner, and G. Livan, "Macroeconomic forecasting through news, emotions and narrative," *Expert Syst Appl*, vol. 175, p. 114760, Aug. 2021, doi: 10.1016/j.eswa.2021.114760.
[13] B. Pratap, A. Ranjan, V. Kishore, and B. B. Bhoi, "Forecasting Food Inflation using News-based Sentiment Indicators," 2022. [Online]. Available: https://static.pib.gov.in/WriteReadData/specificdocs/documents/2021/oct/
[14] J. S. Bandara and Y. Cai, "The impact of climate change on food crop productivity, food prices and food security in South Asia," *Econ Anal Policy*, vol. 44, no. 4, pp. 451–465, Oct. 2014, doi: 10.1016/j.eap.2014.09.005.
[15] J. Ha, S. Lee, and S. Kim, "Influence Relationship between Online News Articles and the Consumer Selling Price of Agricultural Products—Focusing on Onions," *Agriculture (Switzerland)*, vol. 13, no. 9, Sep. 2023, doi: 10.3390/agriculture13091707.
[16] F. Khairani, A. Kurnia, M. N. Aidi, and S. Pramana, "Predictions of Indonesia Economic Phenomena Based on Online News Using Random Forest," *SinkrOn*, vol. 7, no. 2, pp. 532–540, Apr. 2022, doi: 10.33395/sinkron.v7i2.11401.
[17] A. H. Shapiro, M. Sudhof, and D. Wilson, "Measuring News Sentiment," *Federal Reserve Bank of San Francisco, Working Paper Series*, pp. 01–49, Mar. 2020, doi: 10.24148/wp2017-01.
[18] J. Li, Z. Xu, H. Xu, L. Tang, and L. Yu, "Forecasting Oil Price Trends with Sentiment of Online News Articles," *Asia-Pacific Journal of Operational Research*, vol. 34, no. 2, Apr. 2017, doi: 10.1142/S021759591740019X.
[19] V. Aprigliano, S. Emiliozzi, G. Guaitoli, A. Luciani, J. Marcucci, and L. Monteforte, "The power of text-based indicators in forecasting the Italian economic activity," 2021.
[20] L. Barbaglia, S. Consoli, and S. Manzan, "Forecasting with Economic News," *Journal of Business and Economic Statistics*, vol. 41, no. 3, pp. 708–719, 2023, doi: 10.1080/07350015.2022.2060988.
[21] C. Elleby, I. P. Domínguez, M. Adenauer, and G. Genovese, "Impacts of the COVID-19 Pandemic on the Global Agricultural Markets," *Environ Resour Econ (Dordr)*, vol. 76, no. 4, pp. 1067–1079, Aug. 2020, doi: 10.1007/s10640-020-00473-6.
[22] P. Eugster and M. W. Uhl, "Forecasting inflation using sentiment," *Econ Lett*, vol. 236, p. 111575, Mar. 2024, doi: 10.1016/j.econlet.2024.111575.
[23] M. Nagib and Z. A. Husodo, "News Sentiment, News Intensity, and Price Movement of Indonesia's 45 Most Liquid Stock Index," *The 5th International Conference on Business, Economics, Social Sciences, and Humanities*, 2022.
[24] N. Feroze, "Forecasting the patterns of COVID-19 and causal impacts of lockdown in top five affected countries using Bayesian Structural Time Series Models," *Chaos Solitons Fractals*, vol. 140, p. 110196, Nov. 2020, doi: 10.1016/j.chaos.2020.110196.
[25] detik.com, "copyright," https://www.detik.com/copyright. Accessed: Jul. 01, 2024. [Online]. Available: https://www.detik.com/copyright
[26] okezone, "Term of Service," https://management.okezone.com/term-of-service. Accessed: Jul. 01, 2024. [Online]. Available: https://management.okezone.com/term-of-service
[27] cnbc, "Terms," https://www.cnbc.com/terms/. Accessed: Jul. 01, 2025. [Online]. Available: https://www.cnbc.com/terms/
[28] R. Catelli, S. Pelosi, and M. Esposito, "Lexicon-Based vs. Bert-Based Sentiment Analysis: A Comparative Study in Italian," *Electronics (Basel)*, vol. 11, no. 3, p. 374, Jan. 2022, doi: 10.3390/electronics11030374.
[29] T. Loughran and B. Mcdonald, "When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks Journal of Finance, forthcoming," 2011.
[30] E. Henry, "Are Investors Influenced By How Earnings Press Releases Are Written?," *Journal of Business Communication*, vol. 45, no. 4, pp. 363–407, Oct. 2008, doi: 10.1177/0021943608319388.
[31] P. , Suanpang, P. , & Jamjuntr, P. Kaewyong, and P. Kaewyong, "Sentiment analysis with a textblob package implications for tourism," *Journal of Management Information and Decision Sciences*, vol. 24, no. S6, pp. 1–9, 2021.
[32] F. Koto and G. Y. Rahmaningtyas, "Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs," in *2017 International Conference on Asian Language Processing (IALP)*, IEEE, Dec. 2017, pp. 391–394. doi: 10.1109/IALP.2017.8300625.
[33] R. Nooraeni, N. P. Yudho, and N. S. Purba, "Using Google Trend Data as an Initial Signal Indonesia Unemployment Rate," *Conference: 62nd ISI World Statistic Congress* , vol. 3, 2019, [Online]. Available: https://www.researchgate.net/publication/335320380
[34] Á. D. Hartvig, Á. Pap, and P. Pálos, "EU Climate Change News Index: Forecasting EU ETS prices with online news," *Financ Res Lett*, vol. 54, Jun. 2023, doi: 10.1016/j.frl.2023.103720.

[35] S. Scott and H. Varian, "Bayesian Variable Selection for Nowcasting Economic Time Series," Cambridge, MA, Oct. 2013. doi: 10.3386/w19567.
[36] F. R. Alharbi and D. Csala, "A Seasonal Autoregressive Integrated Moving Average with Exogenous Factors (SARIMAX) Forecasting Model-Based Time Series Approach," *Inventions*, vol. 7, no. 4, p. 94, Oct. 2022, doi: 10.3390/inventions7040094.
[37] S. R. D. Setiawan, "Harga Bawang Merah Mahal, Ini Penyebabnya.Kompas. ." Accessed: Jun. 10, 2024. [Online]. Available: https://money.kompas.com/read/2020/06/23/154419726/harga-bawang-merah-mahal-ini-penyebabnya?page=all
[38] K. D. Utami, "Penurunan Produksi Picu Gejolak Harga Bawang Merah." Accessed: Jun. 10, 2024. [Online]. Available: https://www.kompas.id/baca/nusantara/2022/07/03/penurunan-produksi-picu-gejolak-harga-bawang-merah
[39] I. Silfia, "Bawang Merah Turun Harga di Seluruh Wilayah Pada Agustus." Accessed: Jun. 10, 2024. [Online]. Available: https://megapolitan.antaranews.com/berita/257532/bawang-merah-turun-harga-di-seluruh-wilayah-pada-agustus
[40] H. Wang, R. Yao, L. Hou, J. Zhao, X. Zhao, and S. HANA Core, "The 34th Canadian Conference on Artificial Intelligence A Methodology for Calculating the Contribution of Exogenous Variables to ARIMAX Predictions," 2021.