BAREKENG: Journal of Mathematics and Its Applications
December 2026 Volume 20 Issue 4 Page 2885-2900
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss4pp2885-2900

TEXT ANALYTICS AND LASSO REGRESSION FOR STOCK PRICE MOVEMENTS

Muhammad Fikri[1*], **Shantika Martha**[2], **Evy Sulitianingsih**[3], **Wirdha Andani**[4]

[1,2,3,4] *Statistics Department, Faculty of Mathematics and Natural Sciences, Universitas Tanjungpura*
Jln. Prof. Dr. Hadari Nawawi, Pontianak, 78124, Indonesia

*Corresponding author's e-mail: * **m.fikri@fmipa.untan.ac.id***

Article Info

Article History:

Received: 12th September 2025
Revised: 30th April 2026
Accepted: 25th May 2026
Available online: 24th August 2026

Keywords:

Event study;
IDX Kompas 100;
LASSO regression;
Stock price movement;
Text analytics;
XGBoost.

ABSTRACT

This study investigates the impact of news events on stock price movements in the IDX Kompas 100 index (JKKM100) by combining machine learning-based text analytics with event study analysis using LASSO regression. News data from January 1, 2019 to July 22, 2024 were collected via web scraping and used as training data for text classification. The trained model was then applied to classify news events in the period from January 1, 2024 to April 30, 2025, which were subsequently used in the event study analysis. Stock price data for the same period (January 1, 2024 to April 30, 2025) were collected to ensure consistency between predictor and response variables. Due to class imbalance, the synthetic minority over-sampling technique (SMOTE) was applied. Several machine learning algorithms were evaluated, and XGBoost achieved the highest accuracy of 72.22%, improving to 79% after hyperparameter tuning. Using weighted abnormal returns as predictors and stock closing prices as response variables, the LASSO regression results show that 13 out of 180 news events significantly influenced stock price movements. The model explains 48.17% of the variance with an RMSE of approximately 5% of the average stock price. Industry-related news contributed the most (43.20%), followed by PESTEL (3.97%) and Investment (1.00%). This study demonstrates that integrating text analytics with LASSO-based event study provides an effective framework for analyzing the impact of news on stock price movements.

How to cite this article:

M. Fikri, S. Martha, E. Sulitianingsih and W. Andani., "TEXT ANALYTICS AND LASSO REGRESSION FOR STOCK PRICE MOVEMENTS", *BAREKENG: J. Math. & App.*, vol. 20, no. 4, pp. 2885-2900, Dec, 2026.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

The development of information technology has changed the way investors and capital market players make decisions. One of the important short-term factors that influence stock prices is news, both from domestic and foreign sources, in addition to considering the macroeconomic and financial conditions of companies in the capital market in a country [1]. Retail investors are still limited in knowledge and skills, so they are vulnerable to uncertainty due to information asymmetry and are easily influenced by market emotions.

News is an important factor because it can directly influence stock prices in the short term (when the news is released) through the dissemination of information in the media. Negative sentiment news related to the global economy, international politics, geopolitical conflicts, to technological developments can influence risk perceptions and investor expectations towards the capital market, such as corporate scandals, often encouraging selling, while the herd effect amplifies price changes [2]. With the advancement of *Natural Language Processing* (NLP), text-based news analysis (*text analytics*) with the help of Machine learning algorithms can be used to group news into certain categories, such as Business, entertainment, politics, technology, Sports, and so on to be used as predictors (variable X) very quickly and easily [3].

The stock index selection system takes into account the Pareto principle and stock performance (fundamentals and performance). The IDX Kompas 100 Index (JKKM100) is selected, consisting of 100 of the most liquid stocks, with good fundamentals and performance on the Indonesia Stock Exchange (IDX). The capitalization value and transaction volume are estimated to represent 70-80% of the total market capitalization value of all stocks listed on the IDX [4]. Considering that stock price movements in this index are not only influenced by the company's fundamental factors, but also by market events/events (event studies) described in the form of news. Changes in stock prices can be assessed using the return value of the index price or the stock from the previous period [5].

The categorization of events in this news can be analyzed using news abstracts with a text analytics approach. The diversity of news that occurs causes information that should be important to be less noticed, so the process of selecting news variables for news that has a large contribution to the stock index is important to do. One relevant statistical method for analyzing the influence of many predictor variables on stock prices is LASSO regression**.** (*Least Absolute Shrinkage and Selection Operator*) [6]. This method is effective for performing feature selection on data with a large number of predictors, such as the results of feature extraction from text data.

Several previous studies have explored the relationship between news and stock price movements using different approaches. [1] analyzed more than 21 million news articles and found that news plays a significant role in explaining stock return jumps. Similarly, [2] highlighted that investor sentiment driven by news significantly affects market behavior. In the context of machine learning, [12] applied deep learning models (Bi-LSTM) to capture the impact of news on stock trends, demonstrating strong predictive capability. Furthermore, [6] emphasized that LASSO-based methods are particularly suitable for high-dimensional text data due to their ability to perform simultaneous variable selection and regularization.

Despite these advancements, most existing studies either focus on sentiment analysis using deep learning models or apply traditional event study methods separately. There is still limited research that integrates machine learning-based text categorization with event study frameworks using regularized regression techniques, particularly in emerging markets such as Indonesia.

Therefore, this study contributes to the literature by combining machine learning-based text analytics with a multivariate event study approach using LASSO regression. This integration enables not only the classification of news into meaningful categories but also the identification of the most influential events affecting stock price movements. Specifically, this study aims to analyze the influence of news events and news categories on stock price fluctuations in the IDX Kompas 100 Index (JKKM100) by considering short-term observation windows (several days before and after the news occurrence) using LASSO regression.

2. RESEARCH METHODS

Based on the objectives of the research, the following are the steps taken in analyzing stock movements based on events in the form of news.

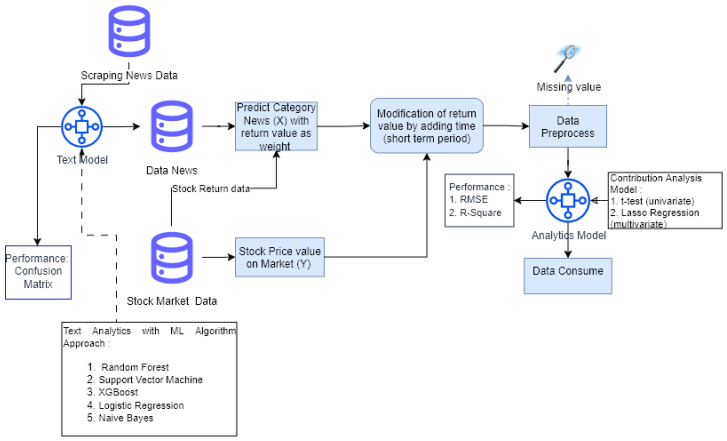


Figure 1 . Workflow Of Events (News) in Analyzing Stock Movements

2.1 Data Collection

According to the workflow in Fig. 1, there are two data collection methods, namely, first, scraping on news.google.com to obtain news data that occurs. Second, collect data through investing.com specifically for the IDX Kompas 100 (JKKM100) index shares to obtain the closing stock prices that occur each day.

According to the workflow in Figure 1, there are two data collection methods. First, news data are collected through scraping from news.google.com, which includes both national and international news sources. Second, stock price data are obtained from investing.com, specifically for the IDX Kompas 100 (JKKM100) index, to retrieve daily closing prices.

In this study, the origin of the news (national or international) is not explicitly differentiated in the analysis. All news articles are treated as a unified dataset, considering that both types of information can simultaneously influence investment sentiment and stock price movements.

2.2 Text Model

In this study, news is categorized into three main groups: Investment, Industry, and PESTEL. The PESTEL category refers to external macro-environmental factors, including Political, Economic, Social, Technological, Environmental, and Legal aspects that may influence market conditions and stock price movements.

2.2.1 Determining news categories from training data

In the context of building a machine learning model, training data is needed, which contains information about all previous events. The training data required for this study consisted of news stories, and the news categories were determined, as illustrated in Table 1.

Table 1. News Samples for Training Data

Date	Topic	Abstract	Category
2024-07-18 08:36:31	Rinse and Repeat: T-Mobile Tops 3rd Party Network Report Yet Again	Key Highlights: * T-Mobile leads in 45 states and the District of Columbia as the fastest network. * T-Mobile covers more than 330 million people with its 5G network. * T-Mobile's Ultra Capacity 5G covers over 300 million people nationwide. Original Press Release: BELLEVUE, Washington, July 17 -- T-Mobile US, Inc.	Investment
⋮	⋮	⋮	⋮
2023-10-06	To boost Indonesia's digital ambitions, NetApp	JAKARTA (IndoTelko) - Global data-centric and cloud-based software	Industry

Date	Topic	Abstract	Category
	partners with Informatics Integration Partners.	company NetApp announced that PT Mitra Integrasi Informatika (MII), a subsidiary of PT Metrodata Electronics, Tbk., one of the largest IT companies in Indonesia, has achieved Prestige status in the new NetApp Partner Sphere program.	

Table 1 shows that the training data for the category was obtained by adjusting the news abstract context, followed by validation by experts in the relevant field. The news items used in this training data were in two languages: English and Indonesian, with the consideration that global news is predominantly in English, while national news tends to be in Indonesian. This was done to reduce vocabulary selection errors during the translation and preprocessing processes in creating the text model.

2.2.2 Vectorization

Text Vectorization refers to the process of converting text (words, sentences, documents) into numeric representations (numbers) so that they can be understood and processed by Machine Learning algorithms. Vectorization is carried out because data in the form of text cannot be read by computers, so it is necessary to carry out a process of converting it into numeric. In the research process, the TF-IDF (Term Frequency - Inverse Document Frequency) vectorization method was chosen with the consideration of reducing the influence of common words (stopwords), so that the focus or keyword of a sentence can be detected properly [7]. The following is the formula for using TF-IDF which is a combination of TF and IDF;

$$TF(t,d) = \frac{f_{t,d}}{Sum\ of\ word\ in\ Document\ d} \tag{1}$$

with, $f_{t,d}$ is the number of occurrences of words t in the document d, then continued with the IDF (Inverse Document Frequency) calculation to reduce the influence of common words (stopwords) such as,

$$IDF(t) = log\frac{N}{1+df(t)} \tag{2}$$

where, N is the total number of documents, and $df(t)$ is the number of documents containing the word t. The intuition of the IDF equation illustrates that the more a word appears in a document, the less informative and lower its weight, and vice versa. These equations are then combined according to the following equation:

$$TFIDF(t,d) = TF(t,d) \times IDF(t) \tag{3}$$

From the results of this equation, a numerical weighting of each sentence is obtained, ready for further processing using a machine learning-based approach. After the vectorization process, preprocessing is performed to divide the data for modeling.

2.2.3 Data Preprocessing

The data preprocessing performed in this study involved dividing the data into two parts: 25% test data and 75% training data. This division was carried out to ensure the model was stable and there were no indications of overfitting. After the data division, the study also performed a data balancing process for categorical variables. The data balancing applied was the SMOTE method, considering that one category was very small compared to the other, and was supplemented with the interpolation concept to create synthetic data according to the following formula:

$$x_{new} = x_i + \lambda.(x_{zi} - x_i) \tag{4}$$

with, x_i is a minority data point followed x_{zi} as the nearest neighbor of x_i (from the minority class). The weight value λ is important in determining synthetic data because it must adjust to the pattern generated from the minority class with values λ ranging from $[0,1]$ [9].

2.2.4 Machine Learning Modeling

After data preprocessing, a modeling process is performed to predict the values of previously validated categories. Several machine learning algorithms will be used in the modeling process to find the best model for predicting categories. The following algorithms are used for text modeling:

SVC (Support Vector Classification)

Linear SVC is an implementation of Support Vector Machine for linear cases. The SVC model is very suitable for high-dimensional data cases [3]. In the study, the optimum hyperplane value was determined to separate the categories of each news item [8] and maximize the margin (distance to the nearest point/support vector). The stages carried out in the following algorithm are as follows:

1. Start
2. Input Data = take the vectorization result features (TF-IDF)
3. Define hyperplane
4. Maximize margin
5. Decision Boundary with decision function

$$f(x) = W^T x + b, \hat{y} = sign\big(f(x)\big) \tag{5}$$

with Optimization (regularized hinge loss),

$$\min_{w,b} \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{n} L_i \tag{6}$$

with the Loss function used,

$$L_i = max(0, 1 - y_i(W.x_i + b)^2 \tag{7}$$

6. Finished

Logistic Regression

Logistic regression is a machine learning algorithm that applies the sigmoid function (probability concept) in making predictions [8]. The following is the algorithm in logistic regression.

1. Start
2. Input Data = take the vectorization result features (TF-IDF)
3. Calculate linear combinations using

$$z = w^T x + b \tag{8}$$

4. Then change it to probability form using the sigmoid function.

$$p = \frac{1}{1 + e^{-z}} \tag{9}$$

5. Decision rule, if $p > 0{,}5$ then class 1, if $p \leq 0.5$then class 0 (case 2 categories, if more adaptable)
6. Training or weight optimization process w, busing Maximum Likelihood Estimation or gradient descent
7. Finished

Multinomial Naive Bayes

The Naïve Bayes method is a classification method based on Bayes' Theorem, which is usually called posterior probability.

$$P(y|x) = \frac{P(x|y).P(y)}{P(x)} \text{or} P(y|x) \propto P(y)\prod_i P(x_i|y) \tag{10}$$

Thus, the application of the Naïve Bayes method in this study is

$$P(y) = \frac{sum\ of\ document\ in\ class\ y}{total\ of\ document} \tag{11}$$

which is called the prior probability. Followed by conditional probability;

$$P(x_i|y) = \frac{sum\ of\ word\ i\ in\ class\ y + \alpha}{total\ of\ word\ in\ class\ y + \alpha V} \tag{12}$$

with the value αbeing the Laplace smoothing parameter [3]. The following are the stages in applying the Naïve Bayes method;

1. Start
2. Calculate the Prior Probability of each class using Eq. (9)
3. Calculate the conditional probability using Eq. (10)
4. Naïve assumption (independence)
5. Calculate the posterior probability using Eq. (8)
6. Predict the class with the highest probability
7. Finished

Random Forest Classifier

It is a type of machine learning modeling that uses ensemble methods in making decisions [9]. The following are the stages in building a model using the Random Forest method;

1. Start
2. Perform bootstrap sampling (bagging) process
3. Select variables randomly
4. Training multiple decision trees
5. Voting uses the majority concept with a formula;

$$\hat{y} = mode\{h_1(x), h_2(x), \dots, h_N(x)\} \tag{13}$$

where $h_N(x)$ is the prediction of the tree n, and $N$is the total number of trees.

6. Final prediction
7. Finished

Extreme Gradient Boosting (XGBoost) Classifer

It is a machine learning modeling method that uses the boosting-based ensemble concept. The stages of the XGBoost algorithm are as follows:

1. Start
2. Simple model initialization

   Start with an initial prediction $\hat{y}_i^{(0)}$(e.g., mean label for regression or log odds for binary classification).

   Initial loss function:

$$L^{(0)} = \sum_i l(y_i, \hat{y}_i^{0}) \tag{14}$$

3. Boosting Iteration

   For $t = 1,2, \dots, T$ (number of trees)

   i. Calculate the error of the previous model according to the loss function:

$$g_i = \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}, h_i = \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{\partial (\hat{y}_i^{(t-1)})^2} \tag{15}$$

   ii. Build a new tree to predict the error.

   iii. Update model with combination of old tree + new tree.

4. Prediction update

   Add new tree predictions to the model

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i) \tag{16}$$

   with η =learning rate, control how much each new tree contributes, and $f_t(x_i)$is the output of the tree $t$for the sample.i

5. Optimization using objective function (loss+regularization)
6. The final model is the result of adding up all the weak trees built, with the formula:

$$\hat{y} = \sum_{t=1}^{T} \eta f_t(x) \tag{17}$$

7. Finished

2.3 Stock Return

Stock returns in the study focused on the weighted value of the news variable (independent) that occurred during the return in the same period. It is known that returns are returns or the value of profits/losses received by investors due to increases/decreases in the shares invested. Based on [10], returns are influenced by the dividend yield of the shares, but in practice some shares rarely provide dividends to their investors [11]. Considering the focus of the study on the stock index, the dividend yield is assumed to be 0, so the formula for return (R_t) is [12], [13]:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}} \tag{18}$$

where, $P_t$is the current stock price, P_{t-1}as the stock price in the previous period. The return weighting calculation scheme for each news item that serves as the independent variable can be seen below:

Starting with a binary event matrix

$$X \in \{0,1\}^{T\times n} \tag{19}$$

Rows = time steps ($i = 1, \dots, T$)
Columns = variables ($j = 1, \dots, n$)
If $X_{i,j} = 1$, that means an event occurs for variable j at time i.

We want to build an expanded version $X_{expand}$where each event "spreads out" into a block of ones covering a time window of size $2b + 1$(that is, from$i - b$ to $i + b$). With $b = 3$, each event produces exactly 7 consecutive ones.

Define a matrix

$$K \in \{0,1\}^{T\times T} \tag{20}$$

with entries

$$K = \begin{cases} 1 \; if \; |p - i| \le b \\ 0 \quad otherwise \end{cases} \tag{21}$$

Each row p of K has ones in the band of indices,$[p - b, \dots, p + b]$
This makes K a banded Toeplitz matrix.

When you multiply KX, you are effectively taking each event in X and "spreading" it into its neighborhood.

$$X_{expan} = \boldsymbol{KX} \tag{22}$$

After obtaining X_{expan}, divide (or scale) by the return weights per row$r \in \mathbb{R}^T$

Formally:

$$X_{return} = diag(r)X_{expan} \tag{23}$$

or element-wise:

$$X_{return\,i,j} = r_i.(X_{expan})_{i,j.} \tag{24}$$

Here,

$X_{return\,i,j}$ is the weighted expansion matrix,
r_i is the return at time i

2.4 Event Study

According to [5], an event study is a study that examines market reactions to a publicly announced event. The most common event is news that is distributed daily. Investors need accurate and relevant information as a basis for analysis when making decisions in the capital market. These announcements serve as signals that assist investors in the decision-making process. Therefore, event studies can be used to assess the information content of these announcements. To determine whether market reactions are influenced by events, an observational window is necessary. A window here refers to the observation period, which can be

short-term or long-term. The most frequently used window is the period 3 days before the event and 3 days after, or written as [-3,3], as in Fig. 2 [10].

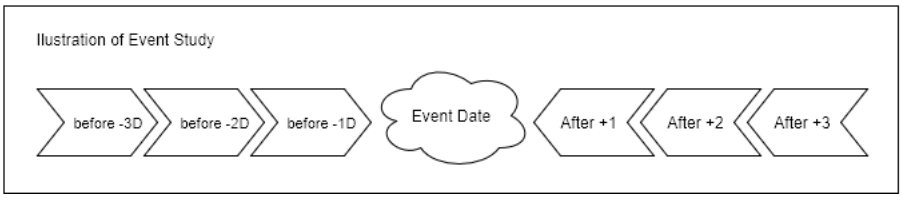


Figure 2 . Workflow of Events (News) In Analyzing Stock Movements

In terms of calculation, the weights used in Fig. 2 are the Abnormal return values. Abnormal return is the difference between the actual profit rate (R_t) and the expected profit rate $E[R_t][1]$. If the abnormal return is positive, it means the profit obtained is greater than what investors expected. Conversely, negative abnormal returns indicate that actual profits are lower than investors' expectations. Abnormal returns can be calculated using three approaches: the mean adjusted model, the market adjusted model, and the market model [5].

2.5 LASSO Regression for News Data

LASSO (Least Absolute Shrinkage Selection Operator) regression is a regression model that aims to handle cases where Pthere are many observed variables with a $n$limited sample ($P > n$). The working principle of the model is by selecting variables and reducing overfitting. The form of the LASSO regression model is as follows [6], [14], [15].,

For continuous responses $y$and features $\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{p}}$, where LASSO minimizes

$$\min_{\beta_0,\beta} \frac{1}{2n} \sum_{t=1}^{n} (y_t - \beta_0 - \boldsymbol{x}_t^T \beta)^2 + \lambda \|\beta\|_1 \tag{25}$$

The penalty L_1pushes many coefficients to exactly zero (sparsity). This process of converting coefficients to exactly zero is called the selection process. When related to the news context, the following equation can be seen:

$$\min_{\beta_0,\beta} \frac{1}{2n} \sum_{t=1}^{n} (StockPrice_t - \beta_0 - \sum_{j=1}^{p} \beta_j . News_{t,j})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{26}$$

Under the condition,

If $\beta_j = 0$, then the news does $j$not significantly affect the stock price.
If $\beta_j \neq 0$, then the news is $j$significant, the size of the influence can be seen from the valueβ_j

2.6 Model Performance

The performance model in this study refers to the goodness of the model used in forming the text model and LASSO regression. The Text Model, because it uses a Machine Learning model, uses a Confusion Matrix to see its performance [9], [16]. On the other hand, for the LASSO regression model, the R-Square approach is used to see the contribution of each variable based on the value β_j, and RMSE (Root Mean Square Error) to see the error from the difference in stock prices between the model results and the actual ones [17], [18]. The following formula is used in the Confusion Matrix

2.6.1 Confusion Matrix for Machine Learning Models

A confusion matrix is a table used to evaluate the performance of a classification model. This table compares the model's predictions with the actual labels, as shown in Table 2.

General structure:

Table 2. Confusion Matrix

Actual/Predict	Positive Predict	Negative Predict
Positive Actual	True Positive (TP)	False Negative (FN)
Negative Actual	False Positive (FP)	True Negative (TN)

Information:

True Positive (TP) : The model correctly predicts the positive class.
True Negative (TN) : The model correctly predicts the negative class.
False Positive (FP) : The model incorrectly predicts a positive when it is negative (*Type I Error*).
False Negative (FN) : The model incorrectly predicts negative when it is positive (*Type II Error*).

From the Confusion Matrix, we can calculate evaluation metrics such as:

Accuracy = (TP + TN) / Total data
Precision = TP / (TP + FP)
Recall (Sensitivity) = TP / (TP + FN)
F1 - Score = 2 * (Precision * Recall) / (Precision + Recall)

2.6.2 R-Square and RMSE for LASSO Regression Model

The measure of the goodness of a model used is using R-Square and RMSE, this takes into account the focus of the research is on the contribution of an event and the closeness of the predicted value of the stock price to the actual value.

1. R-Square (Coefficient of Determination)
   R-Square is used to measure how well the LASSO model is able to explain the variation of the actual data [17]. The equation used is as follows;
   $$R^2 = 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2} \tag{27}$$
   with,
   $R^2 = 1$, meaning the model prediction is perfect
   $R^2 = 1$, the model is not at all better than average
2. RMSE (Root Mean Squared Error)
   RMSE is used to measure how far the average model prediction is from the actual value in the same units as the data [17]. The equation used is as follows;
   $$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \tag{28}$$
   with the explanation that the smaller the RMSE, the better the model's accuracy. The RMSE value depends on the variable 's value scale Y, in this case the stock price of the IDX Kompas 100 Index (JKKM100).

3. RESULTS AND DISCUSSION

3.1 Data Collection Results

Based on the research conducted, data collection is divided into two main components: news data and stock price data. The news data used for training the text classification model were collected from January 1, 2019 to July 22, 2024. For the event study analysis, a separate dataset was used, consisting of classified news events and corresponding stock price data from January 1, 2024 to April 30, 2025 The stock price data for the IDX Kompas 100 (JKKM100) index were obtained on a daily basis, resulting in 302 observations. The closing price was used to calculate stock returns, which serve as weights for the news events occurring on the same day.

This separation ensures that the model training phase and the event analysis phase remain independent, thereby avoiding data leakage and maintaining consistency in the analysis. Table 3 presents the summary of datasets used in this study.

Table 3. Summary of Data Collection

Data Type	Purpose	Periode	Total Data
News Data (Training)	Text classification model	Jan 1, 2019 – Jul 22, 2024	10826 news
News Data (Event)	Event study analysis	Jan 1, 2024 – Apr 30, 2025	180 news
Stock Price Data	Response variable (closing)	Jan 1, 2024 – Apr 30, 2025	302 observations

The closing price was used to calculate stock returns, which serve as weights for the news events occurring on the same day. This separation ensures that the model training phase and the event analysis phase remain independent, thereby avoiding data leakage and maintaining consistency in the analysis.

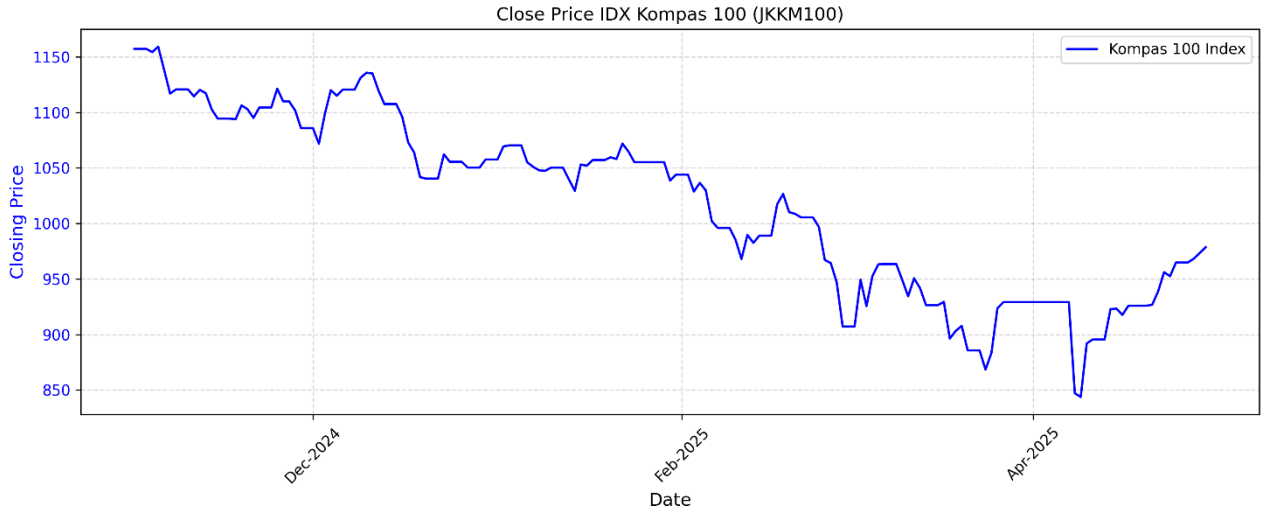


Figure 3. Close price IDX Kompas 100 (JKKM100)
(Application Source: Python)

The price of the IDX Kompas 100 (JKKM100) stock index has experienced a downward trend in the past year. This downward trend further confirms the need for further analysis of the impact of the JKKM100 index price decline. This study focuses on the impact of events occurring over the past year. The following are descriptive statistics from data collected on Investing.com and scraped from Google.News. It can be seen that the average closing price for the IDX Kompas 100 Index (JKKM100) is valued at Rp1102.53 (the dominant data size is below the Q2 value of the total available data). The highest value is Rp1211.27, while the Q3 value is only Rp1169.35, indicating that the highest value only occurs in a certain range (inconsistent), this indicates that stock price fluctuations occur over a short period of time. Related to the collection of event data through the scraping process on google.news, 10826 news data were obtained with abstracts or summaries as training data. To build a text-based news model, the following categorization was carried out;

Table 4. Result Category Training Data

News Category	Actual Count	Count with balancing
Investment	5535	5535
Industry	4182	5535
PESTEL	978	5535

3.2 Text Model Result

After handling the data imbalance process, modeling was carried out with an observation period for training data from January 1, 2019 – July 22, 2024. Data collection was carried out by scraping news.google.com with the help of python 3 software using the feedparser package. To carry out this news text analysis process, several stages were carried out including preprocessing text data, vectorization, and then modeling using Machine Learning algorithms to categorize news.

In Table 4, it can be seen that the categories for the training data have unbalanced data, where the Investment category has 5535 news and PESTEL 978 news. Considering the difference in the number of categories that is more than 50% of the most categories, it is necessary to standardize the data (adjustment), namely by carrying out the synthetic minority over-sampling (SMOTE) technique process used for over-sampling. The results of the over-sampling process obtained a number of majority categories, namely 5535 as in Table 4. After preprocessing the data, the results of modeling several machine learning algorithms were

obtained, as seen in Table 5, the best model in categorizing news is the XGBoost model with an accuracy of 72.22%.

Table 5. Descriptive Data Close Price IDX Kompas 100 (JKKM100)

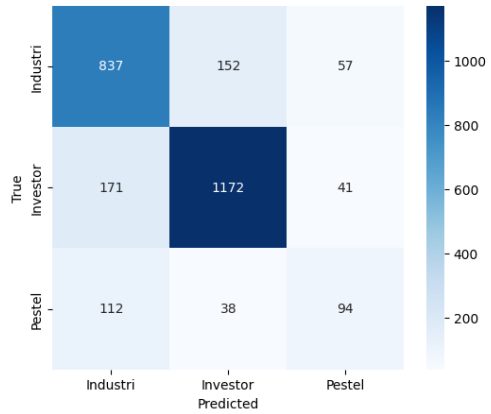
Model	Accuracy (%)
Linear SVC	71.23
Logistic Regression	71.81
Multinomial Naive Bayes	68.08
Random Forest	36.09
SVC	71.82
XGBoost	72.22

Then, from the best model results in Table 5, a parameter tuning process was carried out using RandomizedSearchCV. After the tuning process, the best parameter results were obtained according to Table 6 with an accuracy level of 79%, an increase of 7% from the model before parameter tuning.

Table 6. Optimal Tuning Hyperparameters of the XGBoost Model

Parameter	Score
tfidf__min_df	1
tfidf__ngram_range	(1, 1)
tfidf__sublinear_tf	True
xgb__colsample_bytree	np.float64(0.8270204442119109)
xgb__learning_rate	np.float64(0.1284644554526709
xgb__max_depth	9
xgb__n_estimators	188
xgb__subsample	np.float64(0.7596527212266415)

After obtaining the best model parameters, Fig. 4 shows a complete explanation of the performance of the best model in categorizing news using the Confusion Matrix. The model based on Fig. 4 (a) shows that the frequency of True predictions is more dominant except for the PESTEL category. After looking at Fig. 4 (b), the true prediction for the PESTEL category only ranges from 0.43 or around 43%. Judging from Fig. 4 (a), the model for PESTEL predictions tends to be industry-oriented. In substance, PESTEL provides a picture of external conditions that shape opportunities and threats for an industry. Therefore, when making predictions, the model reads it as news related to the industry even though it is actually related to PESTEL [19]. However, overal the model achieved an accuracy of 79%, which can be considered satisfactory for text classification tasks involving imbalanced and high-dimensional data. This result indicates that the model is able to generalize well in categorizing news into the defined classes.



(a)

	precision	recall	f1-score	support
Industri	0.75	0.80	0.77	1046
Investor	0.86	0.85	0.85	1384
Pestel	0.49	0.39	0.43	244
accuracy			0.79	2674
macro avg	0.70	0.68	0.69	2674
weighted avg	0.78	0.79	0.78	2674

(b)

Figure 4. Best Model Machine Learning Performance Result
(a) Confusion Matrix (b) Classification Report
(Application Source: Python)

Based on the text model results obtained, the next step is to apply the model to make predictions based on the latest news data that will be observed for stock movement analysis. In this case, the predicted news data covers the period from January 1, 2024, to April 30, 2025, which will be detailed in the next subsection.

3.3 LASSO Regression Results

The steps taken to see the influence of events on stock price movements are:

3.3.1 Determining Predictor and Response Variables

In this study, the predictor variables used are events that occurred during the period of January 1, 2024 – April 30, 2025, then carried out by weighting abnormal returns according to equation (24). The response variable used is the closing price of the IDX Kompas 100 (JKKM100) index shares that occurred from the period of January 1, 2024 – April 30, 2025. The following is a sample data frame formed for the LASSO regression modeling process.

	date	21 Proyek Infrastruktur Jadi Korban Efisiensi Anggaran, Ini Daftarnya - CNBC Indonesia	27 tahun Reformasi, Indonesia alami erosi kebebasan politik dan hak-hak sosial - Amnesty Indonesia	42+ Daftar Saham Teknologi di Indonesia yang Cocok Untuk Jangka Panjang - InvestasiKu	60 Tahun Mencatat Indonesia (1): Gejolak Politik dan Ekonomi - Kompas.id	AHY Tawarkan 18 Proyek Infrastruktur ke Investor, Ada 3 Tol Ini - CNBC Indonesia	ASEAN: Anak muda Indonesia paling pesimistis terhadap kondisi politik dan ekonomi Anak muda Indonesia paling pesimistis terhadap kondisi politik dan ekonomi di antara enam negara ASEAN - BBC	Ahli: Pembelajaran AI berpotensi dorong kemajuan teknologi Indonesia - ANTARA News Sulteng	Aksi Indonesia Gelap, Upaya Edukasi Kritis terhadap Kondisi Sosial, Politik, dan Demokrasi - NU Online	Alasan Infrastruktur Jadi Tantangan Indonesia dalam Menuju Negara Maju - kumparan.com - Kumparan	Anggaran PU Dipotong Rp 81,38 T, Begini Nasib Proyek Infrastruktur RI - CNBC Indonesia	Artikel Kanwil DJKN Jakarta - DJKN	Asosiasi Logistik Yakin Danantara Bisa Dorong Kemajuan Infrastruktur - MetroTVNews.com	Aturan Presidential Threshold Pasca Putusan MK - Hukumonline	BEI Awasi Pergerakan Saham Informasi Teknologi Indonesia (JATI), Ini Penyebabnya - kontan.co.id	
0	2024-11-01	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1	2024-11-02	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2	2024-11-03	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
3	2024-11-04	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
4	2024-11-05	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
5	2024-11-06	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
6	2024-11-07	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
7	2024-11-08	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
8	2024-11-09	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
9	2024-11-10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
10	2024-11-11	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
11	2024-11-12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
12	2024-11-13	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
13	2024-11-14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
14	2024-11-15	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Figure 5. Sample of Data Frame for LASSO Regression Model

It is known that the number of news events used in the event study period is 180. The analysis period spans from January 1, 2024 to April 30, 2025 (approximately 16 months), during which stock price data are observed on a daily basis. To construct the regression model, a matrix representation is formed where each column represents a news event, and each row represents a time point (day). Therefore, the resulting matrix dimension is 180 × T, where T denotes the total number of observation days in the study period. Each news event is transformed into a time-distributed variable using a fixed event window, allowing its effect to be observed across multiple days. This structure enables the model to capture the temporal impact of news on stock price movements.

3.3.2 LASSO Regression Modeling and Performance

Based on the data analysis conducted, the following news variables were produced (in this case described using news categories) that significantly influenced the share price of the IDX Kompas 100 index (JKKM100) as shown in Fig. 5.

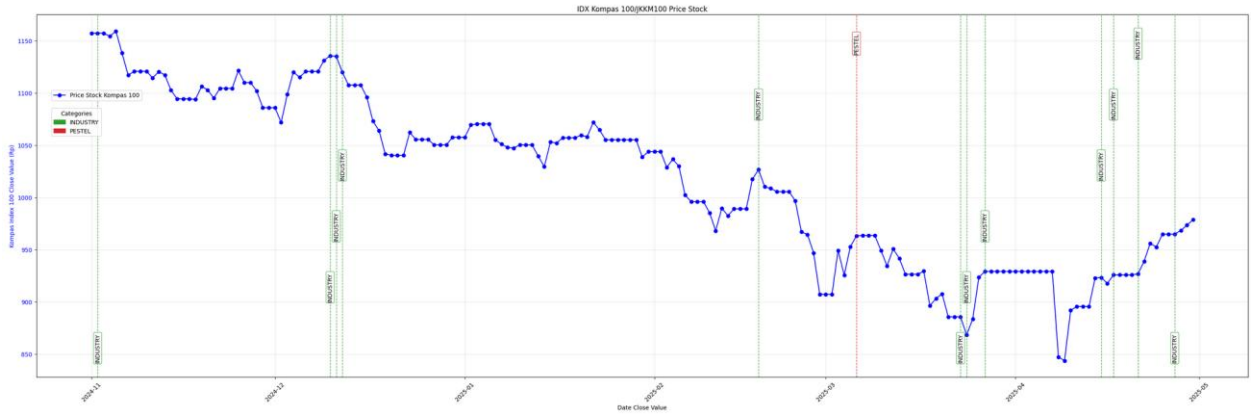


Figure 5. News Category Significant
(Application Source: Python)

It was found that after conducting the variable selection process (news events) from the original 180 news, only 13 news significantly influenced the changes in the stock price of the IDX Kompas 100 index (JKKM100) with the dominance of significant news in the Industry and PESTEL categories. Then, looking at the performance according to Table 7, the R-square value of 48.17% indicates that the model is able to explain a substantial portion of the variation in stock price movements. Although the value is below 50%, this result is considered acceptable given the complexity of stock price dynamics, which are influenced by multiple factors beyond news events, such as macroeconomic conditions, company fundamentals, and global market sentiment. In this study, the model focuses specifically on the impact of news events, which represent only a subset of the overall determinants of stock prices. Therefore, the obtained R-square value reflects the partial explanatory power of news-based variables. Furthermore, previous studies have reported that models incorporating external factors such as macroeconomic variables typically achieve R-square values in the range of 40%–50% [4], [20], [21]. This indicates that the performance of the proposed model is consistent with existing empirical findings.

Table 7. Performance Of LASSO On Model Regression

Performance	Score
R-Square	48.17%
RMSE	Rp. 56.44

To see the contribution of the news event category, according to Eq. (26), the contribution results for each category are obtained in Table 8. The highest results based on news categories are influenced by news related to industrial activities, namely 43.20 %.

Table 8. Contribution Category News to Stock Price

News Category	Contribution
Industry	43.20%
PESTEL	3.97%
Investment	1.00%

The contribution values presented in Table 8 do not sum to 100%, as they represent the partial contribution of each news category to the explained variance of the model (R-square = 48.17%). The remaining unexplained variance is influenced by other factors not included in the model, such as macroeconomic conditions, company fundamentals, and market sentiment. It is known that the shares in the IDX Kompas 100 (JKKM100) index are very liquid, therefore a company's policies and updates to an industrial sector are very relevant to the movement of shares in the related company [4].

4. CONCLUSION

This study demonstrates that the integration of machine learning-based text analytics and LASSO regression within an event study framework can effectively identify the impact of news events on stock price movements in the IDX Kompas 100 (JKKM100) index. The results show that a subset of news events

significantly influences stock price fluctuations, with Industry-related news contributing the most, followed by PESTEL and Investment categories. The model achieved an accuracy of 79% for text classification and an R-square value of 48.17% for regression analysis, indicating that news events provide meaningful, albeit partial, explanatory power for stock price movements.

However, this study has several limitations. First, the model only considers news-based variables and does not incorporate other important factors such as macroeconomic indicators, company fundamentals, or global financial conditions, which may also significantly affect stock prices. Second, the categorization of news into predefined classes may oversimplify the complexity of real-world information. Third, the analysis relies on historical data within a limited observation period, which may not fully capture long-term market dynamics.

Future research can extend this study by integrating additional variables such as macroeconomic data, financial indicators, and sentiment scores to improve the explanatory power of the model. Furthermore, advanced deep learning approaches, such as transformer-based models, may be explored to enhance the accuracy of text classification. Expanding the study to different stock indices or international markets could also provide broader insights into the relationship between news events and stock price movements.

Author Contributions

Muhammad Fikri: Conceptualization, Methodology, Formal Analysis, Writing – Original Draft. Shantika Martha: Data Curation, Software, Visualization, Writing – Review and Editing. Evy Sulistianingsih: Investigation, Resources, Validation. Wirdha Andani: Supervision, Project administration, Funding acquisition, Writing – Review and Editing. All authors reviewed the findings and participated in preparing the final manuscript.

Funding Statement

This research was funded by the Ministry of Higher Education, Science and Technology (KemenDiktiSaintek), Republic of Indonesia, through the 2025 State Budget (DIPA) allocation to faculty of Mathematics and Natural Science, University of Tanjungpura Pontianak, under Grant/Contract No. 2272/UN22.8/PT.01.05/2025.

Acknowledgment

The authors would like to express their sincere appreciation to the Faculty of Mathematics and Natural Sciences for providing research funding. The authors also extend their gratitude to the Statistics Study Program, University of Tanjungpura Pontianak, for facilitating the provision of software used in plagiarism checking.

Declarations

The authors declare no competing interest.

Declaration of Generative AI and AI-assisted technologies

Generative AI tools (e.g., ChatGPT) were used solely for language refinement (grammar, spelling, and clarity). The scientific content, analysis, interpretation, and conclusions were developed entirely by the authors. The authors reviewed and approved all final text.

REFERENCES

[1] Y. Jeon, T. H. McCurdy, and X. Zhao, "NEWS AS SOURCES OF JUMPS IN STOCK RETURNS: EVIDENCE FROM 21 MILLION NEWS ARTICLES FOR 9000 COMPANIES," *J. financ. econ.*, vol. 145, no. 2, pp. 1–17, 2022, doi: https://doi.org/10.1016/j.jfineco.2021.08.002.

[2] R. Basilone, "THE IMPACT OF NEWS ON STOCK MARKET INVESTORS," *SSRN Electron. J.*, 2021, doi: https://doi.org/10.2139/ssrn.3908662.

[3] A. Hussain, G. Ali, F. Akhtar, Z. H. Khand, and A. Ali, "DESIGN AND ANALYSIS OF NEWS CATEGORY PREDICTOR," *Eng. Technol. Appl. Sci. Res.*, vol. 10, no. 5, pp. 6380–6385, 2020, doi: https://doi.org/10.48084/etasr.3825.

[4] G. Andani and M. L. A. Kurniawan, "ANALISIS VARIABEL MAKROEKONOMI TERHADAP INDEKS SAHAM KOMPAS 100: PENDEKATAN VECM," *J. Adv. Accounting, Econ. Manag.*, vol. 1, no. 3, pp. 1–17, 2024, doi: https://doi.org/10.47134/aaem.v1i3.216.

[5] S. K. Hayumurti and K. Khomsiyah, "EVENT STUDY: CAPITAL MARKET REACTION BEFORE AND AFTER THE ANNOUNCEMENT OF PRESIDENTIAL AND VICE PRESIDENTIAL CANDIDATES 2024," *Owner*, vol. 9, no. 2, pp.

1295–1303, 2025, doi: https://doi.org/10.33395/owner.v9i2.2675.
[6] M. Freo and A. Luati, *LASSO-BASED VARIABLE SELECTION METHODS IN TEXT REGRESSION: THE CASE OF SHORT TEXTS*, vol. 108, no. 1. Springer Berlin Heidelberg, 2024.doi: https://doi.org/10.1007/s10182-023-00472-0
[7] Y. Ren, F. Liao, and Y. Gong, "IMPACT OF NEWS ON THE TREND OF STOCK PRICE CHANGE: AN ANALYSIS BASED ON THE DEEP BIDIRECTIONA LSTM MODEL," *Procedia Comput. Sci.*, vol. 174, pp. 128–140, 2020, doi: https://doi.org/10.1016/j.procs.2020.06.068.
[8] M. George and R. Murugesan, "IMPROVING SENTIMENT ANALYSIS OF FINANCIAL NEWS HEADLINES USING HYBRID WORD2VEC-TFIDF FEATURE EXTRACTION TECHNIQUE," *Procedia Comput. Sci.*, vol. 244, pp. 1–8, 2024, doi: https://doi.org/10.1016/j.procs.2024.10.172.
[9] V. Rupapara *et al*., "IMPACT OF SMOTE ON IMBALANCED TEXT FEATURES FOR TOXIC COMMENTS CLASSIFICATION USING RVVC MODEL," *IEEE Access*, vol. 9, pp. 78621–78634, 2021, doi: https://doi.org/10.1109/ACCESS.2021.3083638.
[10] D. Zaria, E. Sulistianingsih, S. Martha, and W. Andani, "PERFORMANCE EVALUATION OF THE INDF.JK STOCK PRICE MOVEMENT PREDICTION MODEL USING RANDOM FOREST METHOD WITH GRID SEARCH CROSS VALIDATION OPTIMIZATION," *Barekeng*, vol. 19, no. 3, pp. 2155–2168, 2025, doi: https://doi.org/10.30598/barekengvol19iss3pp2155-2168.
[11] V. Dogra, A. Singh, S. Verma, A. Alharbi, and W. Alosaimi, "EVENT STUDY: ADVANCED MACHINE LEARNING AND STATISTICAL TECHNIQUE FOR ANALYZING SUSTAINABILITY IN BANKING STOCKS," *Mathematics*, vol. 9, no. 24, pp. 1–18, 2021, doi: https://doi.org/10.3390/math9243319
[12] E. Nurhayati, A. Hamzah, and H. Nugraha, "STOCK RETURN DETERMINANTS IN INDONESIA," *Indones. Account. J.*, vol. 3, no. 1, p. 45, 2021, doi: https://doi.org/10.32400/iaj.32196.
[13] Heny Sidanti and Annisa Istikhomah, "THE EFFECT OF STOCK PRICE, SHARE RETURN, SHARE TRADING VOLUME, AND RETURN VARIANT ON BID-ASK SPREAD ON TEXTILE AND GARMENT COMPANIES LISTED ON THE INDONESIA STOCK EXCHANGE, 2019-2020," *Int. J. Sci. Technol. Manag.*, vol. 2, no. 4, pp. 1357–1366, 2021, doi: https://doi.org/10.46729/ijstm.v2i4.269.
[14] M. Yunus, A. Saefudin, and A. M. Soleh, "CHARACTERISTICS OF GROUP LASSO IN HANDLING HIGH CORRELATED DATA," *Appl. Math. Sci.*, vol. 11, no. June, pp. 953–961, 2017, doi: https://doi.org/10.12988/ams.2017.7276.
[15] Y. Huo, L. Xin, C. Kang, M. Wang, Q. Ma, and B. Yu, "SGL-SVM: A NOVEL METHOD FOR TUMOR CLASSIFICATION VIA SUPPORT VECTOR MACHINE WITH SPARSE GROUP LASSO," *J. Theor. Biol.*, vol. 486, 2020, doi: https://doi.org/10.1016/j.jtbi.2019.110098.
[16] S. Sathyanarayanan, "CONFUSION MATRIX-BASED PERFORMANCE EVALUATION METRICS," *African J. Biomed. Res.*, no. December, pp. 4023–4031, 2024, doi: https://doi.org/10.53555/AJBR.v27i4S.4345.
[17] M. Shanmugavalli and K. M. J. Ignatia, "COMPARATIVE STUDY AMONG MAPE, RMSE AND R SQUARE OVER THE TREATMENT TECHNIQUES UNDERGONE FOR PCOS INFLUENCED WOMEN M.," *Recent Patents on Engineering*, vol. 19. pp. 182–189, 2025, doi: https://doi.org/10.1002/1099-0992(200009/10)30:5<613::AID-EJSP11>3.3.CO;2-J.
[18] N. Safaeian Hamzehkolaei and M. Alizamir, "PERFORMANCE EVALUATION OF MACHINE LEARNING ALGORITHMS FOR SEISMIC RETROFIT COST ESTIMATION USING STRUCTURAL PARAMETERS," *J. Soft Comput. Civ. Eng.*, vol. 5, no. 3, pp. 32–57, 2021, doi: https://doi.org/10.22115/SCCE.2021.284630.1312.
[19] C. Cristina, C. Cindy, T. Wang, D. Ng, and J. Andelson, "ANALYZING EXTERNAL ENVIRONMENT MC DONALD'S," *J. Ilmu Manaj. Profitab.*, vol. 8, no. 1, pp. 93–101, 2024, doi: https://doi.org/10.26618/profitability.v8i1.14031.
[20] W. H. Lucky and T. Yuniati, "PENGARUH VARIABEL MAKRO EKONOMI TERHADAP HARGA SAHAM PADA SEKTOR PERBANKAN," *J. Ilmu dan Ris. Manaj.*, vol. 7, 2018, [Online]. Available: www.wikipedia.org.
[21] W. Puspita, "PENGARUH KINERJA KEUANGAN TERHADAP HARGA SAHAM: STUDI KASUS PADA PERUSAHAAN YANG TERDAFTAR DI BEI," *Ilmu dan Ris. Akunt.*, vol. 6, no. 5, pp. 1941–1957, 2017.