

SKRIPSI

<JUDUL>



**UNIVERSITAS
MERCU BUANA
YOGYAKARTA**

Disusun Oleh:

Nama : <nama lengkap>

NIM : <nim>

**PROGRAM STUDI SISTEM INFORMASI
FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS MERCU BUANA YOGYAKARTA
TAHUN <tahun>**

<JUDUL>

Disusun Oleh:

Nama : <nama lengkap>

NIM : <nim>

**Skripsi ini diajukan untuk memenuhi persyaratan akademik sarjana pada
Program Studi Sistem Informasi, Fakultas Teknologi Informasi,
Universitas Mercu Buana Yogyakarta.**

Yogyakarta, <tgl-bulan-tahun>
Menyetujui Pembimbing,

<pembimbing>
NIDN

HALAMAN PENGESAHAN

<JUDUL>

Oleh:

<Nama Lengkap>

<NIM>

Telah dipertanggungjawabkan dan diterima
oleh Tim Penguji pada tanggal
<tgl-bulan-tahun>

Mengetahui
Dekan,

Dosen Pembimbing,

(<nama beserta gelar>)

(<dosen pembimbing>)

NIDN.

NIDN.

Dosen Penguji,

1. <dosen penguji 1>
NIDN.

2. <dosen penguji 2>
NIDN.

3. <dosen penguji 3>
NIDN.

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI SKRIPSI UNTUK KEPENTINGAN AKADEMIS

Sebagai mahasiswa Universitas Mercu Buana Yogyakarta, saya yang bertanda tangan di bawah ini:

Nama : <nama lengkap>
NIM : <nim>
Program Studi : Sistem Informasi
Fakultas : Teknologi Informasi
Jenis Karya : Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Mercu Buana Yogyakarta Hak Bebas Royalti Non-eksklusif (*None-exclusive Royalty Free Right*) atas karya ilmiah saya yang berjudul:

<judul>

Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Royalti Non-eksklusif ini Universitas Mercu Buana Yogyakarta berhak menyimpan, mengalih-media-kan, mengelola dalam bentuk pangkalan data (*database*), merawat dan mempublikasikan skripsi saya selama tetap mencantumkan nama saya sebagai penulis dan sebagai pemilik hak cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : YOGYAKARTA
Pada Tanggal : <tgl-bulan-tahun>

Yang menyatakan,

(<nama lengkap>)

HALAMAN PERNYATAAN ORISINALITAS KARYA

**Skripsi ini adalah hasil karya saya sendiri,
dan semua sumber baik yang dikutip maupun yang dirujuk,
telah saya nyatakan dengan benar.**

Nama Lengkap : <nama lengkap>

NIM : <NIM>

Judul Skripsi : <judul>

Program Studi : Sistem Informasi

Tanggal : <tgl-bulan-tahun>

Tanda Tangan : <ttd>

HALAMAN PERSEMBAHAN

Tuliskan beberapa kalimat puitis persembahkan karya tulis ini untuk beberapa orang terdekat dengan Anda.

Format tulisan bebas, termasuk jenis, ukuran.

HALAMAN MOTTO

Tuliskan motto Anda.

Format tulisan bebas, termasuk jenis, ukuran.

KATA PENGANTAR

Paragraf pertama berisi ungkapan rasa syukur atas terselesaikannya karya tulis ini.

Selanjutnya dapat dituliskan ucapan terima kasih kepada beberapa pihak yang terkait. Diawali dengan kalimat pembuka, kemudian tuliskan beberapa pihak yang tersebut dalam pointer angka arab (1,2,3,4,...). Contoh:

1. Pihak pertama, sebagai, atas bantuan dan kerjasamanya dalam hal.....
2. Pihak pertama, sebagai, atas bantuan dan kerjasamanya dalam hal.....
3. Pihak pertama, sebagai, atas bantuan dan kerjasamanya dalam hal.....
4. Pihak pertama, sebagai, atas bantuan dan kerjasamanya dalam hal.....
5. Pihak pertama, sebagai, atas bantuan dan kerjasamanya dalam hal.....

Paragraf terakhir adalah paragraf penutup. Beri ungkapan seperlunya.

Yogyakarta, <tgl-bulan-tahun>
Penulis,

<nama lengkap>

TEMPEL DOKUMEN ARTIKEL ILMIAH DI BAGIAN INI

Note : ketentuan penulisan, format dan layouting artikel tergantung pada template jurnal yang dituju. Silakan menyesuaikan dengan format artikel yang bersesuaian.



Analysis of Forecasting Methods on Rice Price Data at Milling Level According to Quality

Indira Dhekawanti Aulia✉, Irfan Pratama

Department of Information Systems, Faculty of Information Technology, Universitas Mercu Buana Yogyakarta, Indonesia

Article Info

Article History:

Received: 13 May 2024

Revised: 31 July 2024

Accepted: 28 August 2024

Keywords:

Forecasting, Random Forest Regression, Rice Prices, RMSE

Abstract

Rice is a primary source of carbohydrates for many Indonesians, and its prices often surge due to uncontrolled demand. Therefore, the government is crucial in monitoring rice prices to maintain stability. Information technology, particularly data mining such as forecasting, is essential for providing accurate information on future rice prices. It will assist various stakeholders in making informed pricing policy decisions. This study employs Random Forest Regression and Gradient Boosting Regressor methods to predict rice prices using a dataset that includes monthly average rice prices at milling levels, categorized by quality (Premium and Medium), spanning from January 2013 to April 2024. The dataset consists of 136 rows, each representing a unique combination of year, month, and quality, and is stored in CSV format. Methodological steps include data collection, preprocessing, modeling, and model evaluation using monthly average rice prices at milling levels based on quality, including premium and medium grades. The results from Random Forest Regression indicate Root Mean Square Error (RMSE) values of 24.90 for premium rice and 25.47 for medium rice. The study reveals that Random Forest Regression outperforms Gradient Boosting Regressor in this context. Future research should explore additional prediction methods and consider other variables influencing rice prices to enhance model accuracy.

INTRODUCTION

Rice is one of the agricultural products and a staple food for a significant portion of the Indonesian population. As Indonesia is an agricultural country, with nearly 90% of its people consuming rice as their primary carbohydrate source, rice plays a crucial role in economic and political stability. The price of rice commodities is continuously monitored and intervened by the government. It was because rice prices contribute to food security, poverty alleviation, macroeconomic stability, and the country's economic growth (Jiuhardi, 2023).

Information technology plays a pivotal role in monitoring and predicting rice prices at the milling level based on quality. Artificial intelligence in the food sector is an innovative technology that supports the management of staple foods, such as price prediction, food quality determination, and demand mapping (Putra & Sinaga, 2022). These technologies offer innovative solutions that enhance the management of staple foods like rice, ensuring that supply and demand dynamics are efficiently addressed.

Data mining is a discipline that studies methods for extracting knowledge or discovering patterns from large datasets (Sumarni & Rustam, 2020). The government also plays a crucial role in ensuring price regulation to prevent drastic fluctuations. Such fluctuations usually occur during major holidays, caused by increased goods such as food, particularly rice, leading to a rise in prices. Fluctuations are changes in certain variables that generally occur due to market mechanisms (R. Amalia et al., 2023).

Therefore, it is necessary to monitor and predict rice prices to maintain stability and prevent burdening disadvantaged community groups, which can be achieved by applying data mining techniques. In this context, data mining is used in the form of forecasting. According to (Yudianto et al., 2023), forecasting methods are divided into quantitative and qualitative categories. Qualitative methods are based on opinions and descriptive analysis, while quantitative methods rely on mathematical calculations.

Forecasting methods have been widely applied in various aspects, particularly in pricing. Research on price prediction has been conducted by several researchers, such as Putra and Sinaga (2022), Mukhlisin, Imrona, and Murdiansyah (2019), Saadah and Salsabila (2021), and Amalia et al. (2022). However, existing studies have yet to focus on simultaneously predicting rice prices based on different types, such as premium or medium.

The previous research titled 'Estimation of Premium Rice Prices in DKI Jakarta using Linear

Regression' aimed to predict the prices of premium rice in DKI Jakarta, resulting in a Mean Absolute Error (MAE) of 275.55 and a Mean Squared Error (MSE) of 103169.10. When the MSE was squarely rooted, the result was an RMSE (Root Mean Squared Error) of approximately 321.199 (Putra & Sinaga, 2022).

Previous research was also conducted by (Adjie Setyadj et al., 2023) with the titled 'Forecasting Rice Commodity Prices in East Kalimantan Using Neural Network Algorithm.' This study used daily premium rice price data obtained from the community in East Kalimantan. The study yielded a Root Mean Square Error (RMSE) value of 52.846 for premium rice.

Similar research was also conducted previously by (Mukhlisin et al., 2019) titled 'Prediction of Premium Rice Prices using the K-Nearest Neighbor Algorithm' using data from the Central Statistics Agency of Bandung and weather data from BMKG Bandung. The study resulted in an RMSE of 352.450 for non-normalized data and an RMSE of 174.38 for normalized data.

Another research study is titled 'Bitcoin Price Prediction Using Random Forest Method.' The research utilizes several attributes: low, high, and price. The Random Forest Regression method produces a MAPE value of 1.50% or achieves an accuracy of around 98% using random data. The data used in this study had high fluctuation characteristics, so the Random Forest Regression method could provide fittings that match the actual data (Saadah & Salsabila, 2021).

Similarly, research conducted by (A. Amalia et al., 2022) with the title 'Car Price Prediction using Regression Algorithm with Hyper-Parameter Tuning.' This study developed three regression models: Linear Regression, Random Forest Regression, and Gradient Boosting Regression. Hyperparameter tuning was applied to enhance the accuracy of the models. Parameters added included an intercept for Linear Regression, 'sqrt' for max features, 'gini' for criterion in Random Forest, and 'sqrt' for max features, and 'friedman_mse' for criterion in Gradient Boosting. The results of this study showed that Gradient Boosting Regression achieved the highest model accuracy, with a training accuracy of 99.58% and a testing accuracy of 96.75%.

Based on the previous explanation, the Random Forest Regression and Gradient Boosting Regression methods have shown promising results. This study aims to test these models on a rice price dataset, comparing them with other methods from previous research to evaluate their effectiveness. The results from

evaluating the best model can then be used to predict rice prices at the milling level based on quality.

RESEARCH METHODS

The research materials used in this study are monthly average rice prices at the Milling

Level according to quality sourced from the Central Bureau of Statistics of Indonesia from 2013 to 2024. This research is conducted using a quantitative method. Quantitative research is a process of discovering knowledge using numerical data as a tool to analyze information about what is sought (Wantari, 2021). The research flowchart can be seen in Figure 1.

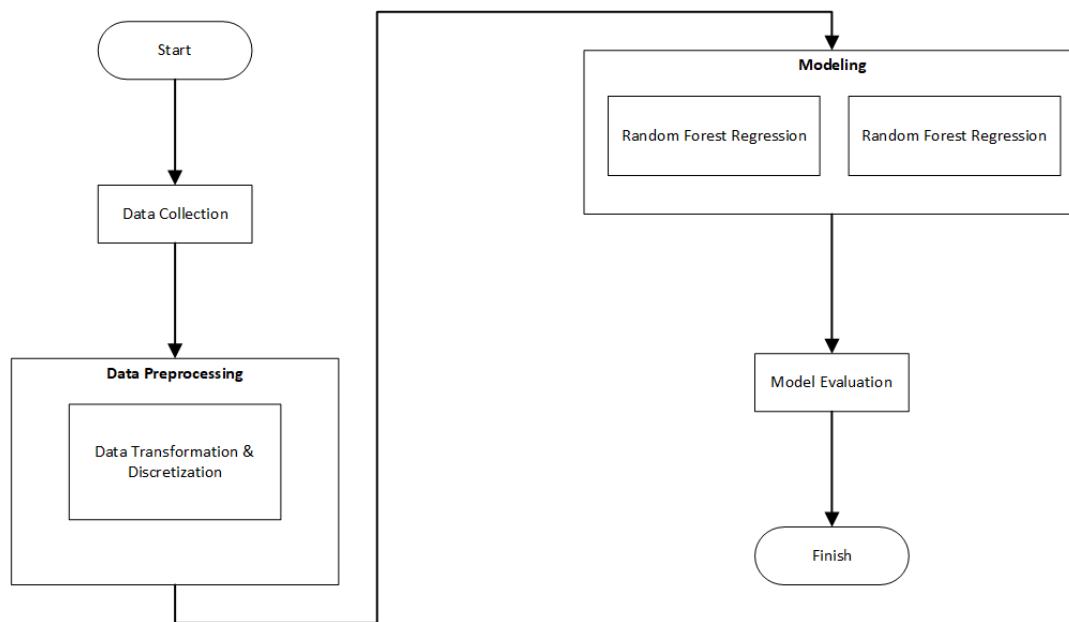


Figure 1. Research flowchart

Based on Figure 1 above, the research flow/path can be explained as follows:

A. Dataset Collection

The data collection in this study involves obtaining data from the Central Bureau of Statistics of Indonesia, which can be accessed at <https://www.bps.go.id/id/statistics-table/2/NTAwIzI=/rata-rata-harga-beras-bulanan-di-tingkat-penggilingan-menurut-kualitas.html>. The dataset includes monthly average rice prices at the Milling Level, categorized by quality (Premium and Medium), spanning from January 2013 to April 2024. It consists of 136 rows, each representing a unique combination of year, month, and quality (Premium and Medium), and is stored in CSV format. The data is utilized to analyze rice prices at the Milling Level according to quality, aiming to assist in building a rice price prediction system. The data format is in the form of a time series. The example dataset used is shown in Table 1.

Table 1. Sample Data of Rice Prices at the Milling Level According to Quality in 2013

| Bulan | Tahun | Premium | Medium |
|-------|-------|---------|---------|
| 1 | 2013 | 7797.3 | 7697.37 |
| 2 | 2013 | 7773.26 | 7645.05 |
| 3 | 2013 | 7576.27 | 7503.27 |
| 4 | 2013 | 7420.72 | 7290.96 |
| 5 | 2013 | 7545.5 | 7261.71 |
| 6 | 2013 | 7548.22 | 7419.63 |
| 7 | 2013 | 7823.68 | 7553.54 |
| 8 | 2013 | 7761.29 | 7524.03 |
| 9 | 2013 | 7746.17 | 7652.87 |
| 10 | 2013 | 7846.05 | 7702.05 |
| 11 | 2013 | 7919.98 | 7732.05 |
| 12 | 2013 | 7976.72 | 7871.21 |

Source: <https://www.bps.go.id/id/statistics-table/2/NTAwIzI=/rata-rata-harga-beras-bulanan-di-tingkat-penggilingan-menurut-kualitas.html>

B. Data Preprocessing

In the data preprocessing stage, the data obtained from the Central Bureau of Statistics of Indonesia, which consists of monthly average rice prices at the Milling Level according to quality,

undergoes several steps. The preprocessing steps applied to this research data include Data Transformation and Discretization.

Data Transformation is a preprocessing stage where data is modified or combined into a suitable data format for processing in Data Mining (Rayuwati et al., 2022). Meanwhile, Discretization is a technique in transformation aimed at converting numerical attributes into categorical attributes, thereby creating several levels or hierarchies (Alghifari & Juardi, 2021).

The data used can improve the model further through this data preprocessing stage. In this stage, what needs to be done is adjusting the delimiter in each data field, converting the month format into numbers, and combining the month and year columns into a date format to form the Year-Month-Date, which is stored in the date column. Next, the premium and medium data are divided into training and testing data using time series splitting.

C. Modeling

This stage begins by determining the prediction method suitable for predicting rice prices at the mill level according to quality. The methods or models used to evaluate this study's best models are Random Forest Regression and Gradient Boosting Regressor.

Random Forest Regression is a supervised machine learning algorithm that repeatedly builds decision trees, thus forming a forest (Saadah & Salsabila, 2021). Random Forest is a classification consisting of several decision trees, each constructed using a random vector (Mambang & Byna, 2017).

Gradient Boosting Regressor is a machine learning model that can be used for regression and classification, and it generates a predictive model consisting of an ensemble of weak prediction models on decision trees that result in shallow prediction errors when using the median as the prediction method (Riyadi et al., 2023).

The selection of these two models is because each model has its advantages. For instance, the Random Forest Regression method has several strengths, such as its ability to improve accuracy when dealing with incomplete data and its robustness against extreme data variations. Therefore, Random Forest Regression can effectively handle large datasets with complex parameters (Mardiyanti Elsa Nurul & Dewi Tresna, 2021).

Furthermore, the Gradient Boosting Regressor is a decision tree classification algorithm for addressing prediction and classification issues. This algorithm is also tree-based, which helps avoid overfitting

(Kraugusteeliana et al., 2023). Those advantages are among the reasons for selecting the Random Forest Regression and Gradient Boosting Regressor methods for modeling rice prices at the Milling Level according to quality.

D. Model Evaluation

The evaluation stage is the most crucial as it aims to assess how well the prediction model performs. The steps in evaluating the model here involve comparing the best models to determine which will be used to predict rice prices at the Milling Level according to quality.

In this evaluation, a loop is used to repeat the training and model evaluation process 30 times, which is used to calculate the average of MAPE and MAE. Thus, in this evaluation stage, Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Error (MAE) are generated.

Root Mean Square Error (RMSE) is the square root of the Mean Square Error (MSE) value obtained from the calculation of a method (Syakir et al., 2022). A high RMSE score means low forecasting accuracy. On the other hand, a low RMSE score means high forecasting accuracy (Sabar Sautomo & Hilman Ferdinandus Pardede, 2021). The formula for calculating RMSE can be seen in Equation (1).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (1)$$

Where:

\hat{y}_i = forecasted value
 y_i = observed value at i-th observation
 n = number of data points

Mean Absolute Percentage Error (MAPE) is the average percentage error value obtained from the sum of each % error value divided by the number of periods in the data (Alfarisi, 2017). The Equation of MAPE can be seen in eq (2).

$$MAPE = \frac{100\%}{n} \sum \left| \frac{y - \hat{y}_i}{y_i} \right| \quad (2)$$

Where:

\hat{y}_i = forecasted value
 y_i = observed value at i-th observation
 n = number of data points

Mean Absolute Error (MAE) is one method used to measure a model's accuracy by intuitively calculating the average error with equal weighting given to all data (Suryanto, 2019). The formula for calculating MAE is here, as seen in Equation (3).

$$MAE = \frac{1}{n} \sum |f_i - y_i| \quad (3)$$

n = number of data points

Where:

f_i = forecasted value

y_i = observed value at i-th observation

From the evaluation results, we can see which method performs better while comparing the evaluation outcomes from previous studies as a reference for selecting the best model.

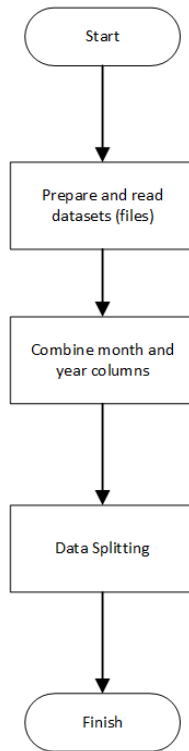


Figure 2. The flowchart of data preprocessing

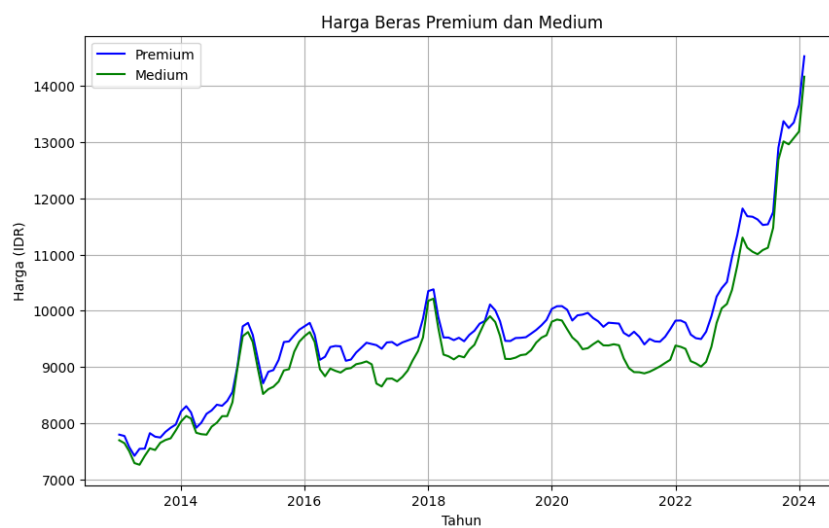


Figure 3. Visualization of actual prices of premium and medium rice

RESULT AND DISCUSSION

In this research, the accuracy results of the models or methods used will be obtained, along with comparisons with previous research that has been conducted. The findings and discussion of this study are as follows:

A. Data Preprocessing

The stages of data preprocessing in this research are crucial as they aim to prepare raw data in a suitable format for further analysis or modeling. The following is the data preprocessing workflow, as depicted in Figure 2.

In Figure 2, the first stage of this data preprocessing involves preparing and reading the dataset. The data obtained from the Central Bureau of Statistics of Indonesia, consisting of monthly average rice prices at the Milling Level according to quality, undergoes this initial phase. In this stage, delimiter=';' is also used, which serves as a separator between values in each column, ensuring that the data system can be interpreted correctly and the columns read accurately. Below is the visualization of actual premium and medium rice prices, as depicted in Figure 3.

Next, the stage involves combining the month and year columns. This combination transforms the 'Bulan' (Month) and 'Tahun' (Year) columns into a DateTime object unified in a single column named 'Tanggal' (Date). Subsequently, the 'Tanggal' column is set as the primary index of the 'data' data frame. The format of the combined 'Bulan' and 'Tahun' data can be seen in Table 2 and Table 3.

Table 2. Sample Data of Month and Year Before Combination

| Month | Year |
|-------|------|
| 1 | 2013 |
| 2 | 2013 |
| 3 | 2013 |
| 4 | 2013 |
| 5 | 2013 |
| 6 | 2013 |
| 7 | 2013 |
| 8 | 2013 |
| 9 | 2013 |
| 10 | 2013 |
| 11 | 2013 |
| 12 | 2013 |

Table 2 presents the sample data of 'Bulan' (Month) and 'Tahun' (Year) before the combination process. The data is in its original form and does not yet reflect the unified 'Tanggal'

(Date) column resulting from the combination of month and year.

Table 3. Sample Data Resulting from Month and Year Columns After Combination

| Date | Month | Year |
|------------|-------|------|
| 2013-01-01 | 1 | 2013 |
| 2013-02-01 | 2 | 2013 |
| 2013-03-01 | 3 | 2013 |
| 2013-04-01 | 4 | 2013 |
| 2013-05-01 | 5 | 2013 |
| 2013-06-01 | 6 | 2013 |
| 2013-07-01 | 7 | 2013 |
| 2013-08-01 | 8 | 2013 |
| 2013-09-01 | 9 | 2013 |
| 2013-10-01 | 10 | 2013 |
| 2013-11-01 | 11 | 2013 |
| 2013-12-01 | 12 | 2013 |

Table 3 shows the sample data after combining the 'Bulan' (Month) and 'Tahun' (Year) columns into a single 'Tanggal' (Date) column. The 'Tanggal' column is created by merging the month and year into a DateTime format. For example, January 2013 is represented as '2013-01-01' in the 'Tanggal' column. This table illustrates the resulting unified date format for the year 2013.

Next, the final stage in this data preprocessing involves data splitting. In this study, data splitting utilizes the technique of time series split, dividing the data into five parts. Each iteration provides index sets within the for loop for training and testing data based on the previously performed time series split. Below are sample training and testing data, as shown in Table 4, Table 5, Table 6, and Table 7.

Table 4. Sample Data for 'y_train_premium'

| Date | y_train_premium |
|------------|-----------------|
| 2021-12-01 | 9672.54 |
| 2022-01-01 | 9824.23 |
| 2022-02-01 | 9826.88 |
| 2022-03-01 | 9786.63 |
| 2022-04-01 | 9576.75 |

Table 5. Sample Data for 'y_test_premium'

| Date | y_test_premium |
|------------|----------------|
| 2022-05-01 | 9512.63 |
| 2022-06-01 | 9497.40 |
| 2022-07-01 | 9628.57 |
| 2022-08-01 | 9901.15 |
| 2022-09-01 | 10252.31 |

Table 4 and Table 5 represent examples of sample training and testing data from the premium dataset. In contrast, Table 6 and Table 7 illustrate

examples of some sample training and testing data from the medium dataset.

Table 6. Sample Data for 'y_train_medium'

| Date | y_train_medium |
|------------|----------------|
| 2021-12-01 | 9128.44 |
| 2022-01-01 | 9381.24 |
| 2022-02-01 | 9358.61 |
| 2022-03-01 | 9323.35 |
| 2022-04-01 | 9104.35 |

Table 7. Sample Data for 'y_test_medium'

| Date | y_test_medium |
|------------|---------------|
| 2022-05-01 | 9065.18 |
| 2022-06-01 | 9007.86 |
| 2022-07-01 | 9091.92 |
| 2022-08-01 | 9358.34 |
| 2022-09-01 | 9785.04 |

B. Modeling

In the modeling stage of this research, two methods or models were chosen to predict rice prices at the milling level according to quality: random forest regression and gradient boosting regression.

In the modeling process, Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Error (MAE) are calculated. The parameter 'n_estimators' in the chosen methods or models, Random Forest Regression and Gradient Boosting Regressor is not explicitly specified, so the default value of 100 will be used for 'n_estimators' parameter settings.

Additionally, the data is divided, where the targets used are premium and medium rice prices. The variables used to store these prediction targets are 'y_premium' and 'y_medium.' The model is trained using different

training and testing data in each iteration. In the final iteration, the training data consists of 114 records, and the testing data consists of 22.

Thus, the training data encompasses approximately 83.21% of the total data, while the testing data comprises about 13.25%. The modeling outcomes will yield RMSE, MAPE, and MAE values for each tested model.

C. Model Evaluation

In the model evaluation stage, the data is split using the time series splitting technique. The 'cross_val_score' function is used to calculate the RMSE. The 'scoring' parameter used is 'neg_root_mean_square_error', which results in negative RMSE values. The negative values are converted to positive using the function 'np.sqrt(-cross_val_score(...))' to address this. The calculation of RMSE is performed five times, corresponding to the number of splits in the time series splitting ('n_splits=5'), and the results are then averaged to obtain the RMSE value used for model evaluation.

Additionally, a loop is employed to obtain MAPE and MAE results, indicating that the model evaluation process is repeated 30 times to measure these metrics. MAPE and MAE are calculated in each iteration, stored in a list, and then averaged to obtain the final values. This iterative approach ensures a robust assessment of model performance across different data splits.

The model evaluation results, consisting of RMSE, MAPE, and MAE, were obtained using the mentioned technique. The library used for model evaluation was sklearn. The model testing results for predicting rice prices at the milling level according to quality, including premium and medium rice, can be seen in Table 8.

Table 8. Model Evaluation Results

| Testing Results | Random Forest Regression | Gradient Boosting Regressor |
|-----------------|--------------------------|-----------------------------|
| | Premium | Premium |
| RMSE | 24.90 | 25.48 |
| MAPE | 18.88% | 18.95% |
| MAE | 2390.67 | 2398.03 |
| | Medium | Medium |
| | Premium | Premium |
| RMSE | 25.47 | 26.24 |
| MAPE | 19.57% | 19.40% |
| MAE | 2392.18 | 2371.62 |

Table 9. Previous Research Results

| Testing Results | K-Nearest Neighbor (Mukhlisin et al., 2019) | | Linear Regression (Putra & Sinaga, 2022) | Neural Network (Adjie Setyadj et al., 2023) |
|-----------------|--|--------------|---|--|
| | Non-Normalization | Denormalized | | |
| | Premium | Premium | Premium | Premium |
| RMSE | 352,450 | 174,38 | 321,199 | 52,846 |
| MAPE | - | - | - | - |
| MAE | - | - | 275,55 | - |

Table 9 highlights the RMSE results from previous research studies, focusing solely on the performance of different prediction models used in those studies. The comparison is limited to RMSE values without accounting for factors such as the volume of data used, data sources, or other evaluation metrics. Concentrating on RMSE, this comparison provides a snapshot of the accuracy of various models employed in past research to predict rice prices.

From Table 8, it can be observed that the Random Forest Regression method has the

lowest RMSE compared to Gradient Boosting Regression and other techniques used in previous studies, with RMSE values of 24.90 for the Premium Rice dataset and 25.47 for the Medium Rice dataset. Therefore, in general, the Random Forest Regressor method applied without changing hyperparameter values is quite effective compared to similar methods for rice price data.

Based on the testing results in Table 8, here are visualizations of the actual prices of premium and medium rice with the tested models, as shown in Figure 4 and Figure 5.

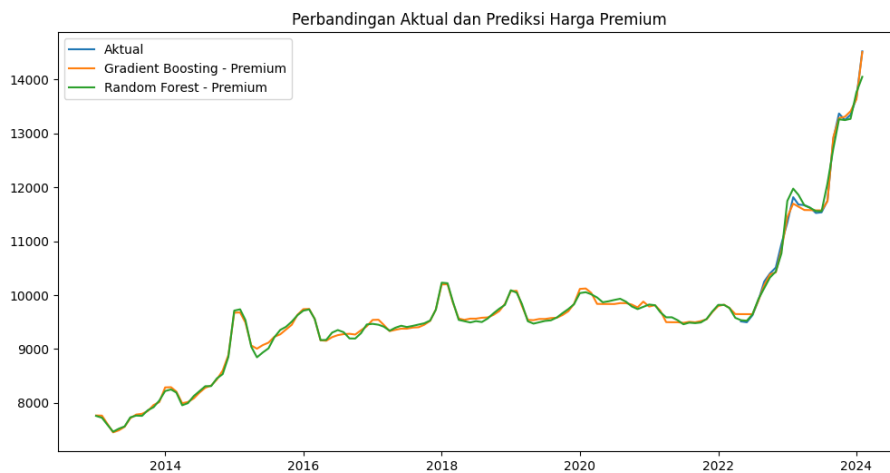


Figure 4. Visualization of Model and Actual Prices of Premium Rice

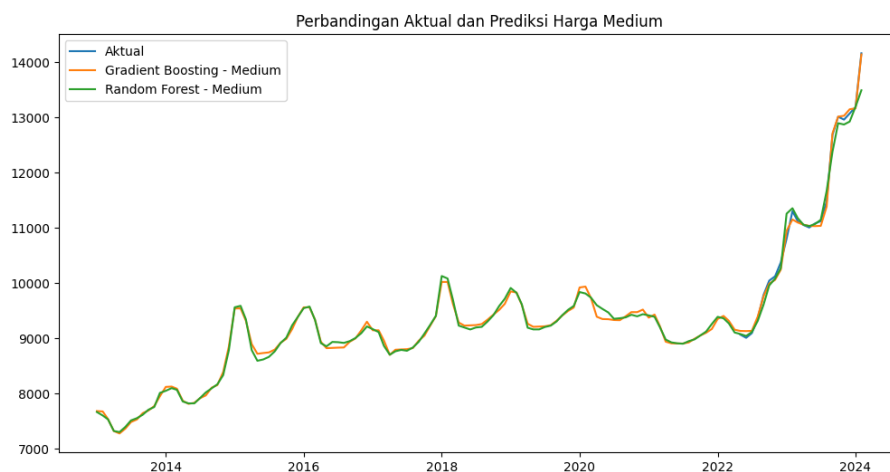


Figure 5. Visualization of Model and Actual Prices of Medium Rice

From the visualizations shown in Figure 4 and Figure 5, the results from the Random Forest Regression model align well with the actual data of both premium rice and medium rice prices. The price movement graphs in these

visualizations use data from 2013 – 2024. In 2022, there is a significant increase in both premium and medium rice prices compared to previous years, indicating a drastic upward trend.

The results of this study provide new insights by showing that the Random Forest Regression method for predicting rice prices at the milling level, according to quality, yields lower RMSE values compared to previous research.

CONCLUSION

From the results of this research, it can be concluded that the Random Forest Regression model demonstrates good performance when implemented on monthly average rice prices at the Milling Level according to quality from 2013 – 2024 data obtained from the Central Bureau of Statistics of Indonesia website.

Data splitting using Time Series Splitting significantly impacts model evaluation. The use of 'cross_val_score' in this study is employed to calculate the RMSE, which is tailored to time series data format utilizing Time Series Splitting for data splitting.

In this research, model testing involved experimenting with two models: Random Forest

Regression and Gradient Boosting Regressor. From the model testing results, the Random Forest Regression method produced the smallest Root Mean Square Error (RMSE), with values of 24.90 for premium rice and 25.47 for medium rice. The MAPE was 18.88% for premium rice and 19.57% for medium rice, while the MAE was 2397.67 for premium rice and 2392.18 for medium rice.

The results of this research are expected to be further developed to enhance the prediction of rice prices at the Milling Level according to quality, using either the same method or different methods to achieve optimal performance. Future research should consider expanding the dataset by integrating additional sources such as regional price variations, production data, and economic indicators to address the limitation of the current dataset size. Additionally, extending the period or incorporating more granular data could provide a more comprehensive view and improve the model's accuracy.

REFERENCES

- Adjie Setyadj, M., Faqih, A., & Arie Wijaya, Y. (2023). Peramalan Harga Komoditas Beras Di Kalimantan Timur Menggunakan Algoritma Neural Network. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 7(1), 320–324.
<https://doi.org/10.36040/jati.v7i1.6327>
- Alfarisi, S. (2017). Sistem Prediksi Penjualan Gamis Toko QITAZ Menggunakan Metode Single Exponential Smoothing. *JABE (Journal of Applied Business and Economic)*, 4(1), 80.
<https://doi.org/10.30998/jabe.v4i1.1908>
- Alghifari, F., & Juardi, D. (2021). Penerapan Data Mining Pada Penjualan Makanan Dan Minuman Menggunakan Metode Algoritma Naïve Bayes. *Jurnal Ilmiah Informatika*, 9(02), 75–81.
<https://doi.org/10.33884/jif.v9i02.3755>
- Amalia, A., Radhi, M., Sinurat, S. H., Sitompul, D. R. H., & Indra, E. (2022). Prediksi Harga Mobil Menggunakan Algoritma Regressi Dengan Hyper-Parameter Tuning. *Jurnal Sistem Informasi Dan Ilmu Komputer Prima (JUSIKOM PRIMA)*, 4(2), 28–32.
<https://doi.org/10.34012/jurnalsisteminf ormasidanilmukomputer.v4i2.2479>
- Amalia, R., Masdiyanti, & Nadir Nadir. (2023). Analisis Fluktuasi Dan Trend Harga Komoditas Telur Ayam Ras Di Kabupaten Bulukumba. *Jurnal Sains Agribisnis*, 3(1), 21–28.
- Hilmi, N., & Saputra, W. A. (2023). Implementasi HE, AHE, dan CLAHE Pada Metode Convolutional Neural Network untuk Identifikasi Citra X-Ray Paru-Paru Normal atau Terinfeksi Covid19. *Edu Komputika Journal*, 10(1), 1–9.
<https://doi.org/10.15294/edukomputika.v10i1.57237>
- Jiuhardi. (2023). Analisis Kebijakan Impor Beras Terhadap Peningkatan Kesejahteraan Petani di Indonesia. *INOVASI: Jurnal Ekonomi, Keuangan Dan Manajemen*, 19(1), 1–13.
- Kraugusteeliana, K., Muis, S., Nugroho, F., Karim, A., & Siagian, Y. (2023). Data Mining Klasifikasi Breast Cancer Menerapkan Algoritma Gradient Boosted Trees. *JURNAL MEDIA INFORMATIKA BUDIDARMA Volume 7, Nomor 2, April 2023, Page 881-890*, 7(April), 881–890.
<https://doi.org/10.30865/mib.v7i2.6095>
- Magnolia, C., Nurhopipah, A., & Kusuma, B. A. (2023). Penanganan Imbalanced Dataset untuk Klasifikasi Komentar Program Kampus Merdeka Pada Aplikasi Twitter. *Edu Komputika Journal*, 9(2), 105–113.
<https://doi.org/10.15294/edukomputika.v9i2.61854>
- Mambang, & Byna, A. (2017). Analisis

- Perbandingan Algoritma C.45, Random Forest Dengan Chaid Decision Tree Untuk Klasifikasi Tingkat Kecemasan Ibu Hamil. *Semnasteknomedia Online*, 5(1), 103–108. <https://ojs.amikom.ac.id/index.php/semnasteknomedia/article/view/1771>
- Mardiyanti Elsa Nurul, Dewi Tresna, O. Y. (2021). *Analisa Prediksi Tegangan Input Sensor Capacitive Soil Moisture dengan Random Forest untuk Mendukung Pertanian Pintar*. 2(1), 13–23. <http://journal.isas.or.id/index.php/JASENS>
- Mukhlisin, Imrona, M., & Murdiansyah, D. T. (2019). Prediksi Harga Beras Premium dengan Metode Algoritma K-Nearest Neighbor. *E-Proceeding of Engineering*, 7(1), 2714–2724.
- Putra, R. E., & Sinaga, A. S. (2022). Perkiraan Harga Beras Premium DKI Jakarta Menggunakan Regresi Linier. *Journal of Information Engineering and Educational Technology*, 6(2), 80–85. <https://doi.org/10.26740/jieet.v6n2.p80-85>
- Rayuwati, Husna Gemasih, & Irma Nizar. (2022). IMPLEMENTASI ALGORITMA NAIVE BAYES UNTUK MEMPREDIKSI TINGKAT PENYEBARAN COVID. *Jurnal Riset Rumpun Ilmu Teknik*, 1(1), 38–46. <https://doi.org/10.55606/jurritek.v1i1.127>
- Riyadi, A. S., Wardhani, I. P., Irfan, & Perdana, A. (2023). Aplikasi Perbandingan Prediksi Harga Bitcoin Menggunakan Deep Learning Dengan Metode Arima, Sarima, Ltsm Dan Gradient Boosting Regressor. *Seminar Nasional Teknologi Informasi Dan Komunikasi STI&K (SeNTIK)*, 7(1), 192–199.
- Saadah, S., & Salsabila, H. (2021). Prediksi Harga Bitcoin Menggunakan Metode Random Forest. *Jurnal Komputer Terapan*, 7(1), 24–32. <https://doi.org/10.35143/jkt.v7i1.4618>
- Sabar Sautomo, & Hilman Ferdinandus Pardede. (2021). Prediksi Belanja Pemerintah Indonesia Menggunakan Long Short-Term Memory (LSTM). *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 5(1), 99–106. <https://doi.org/10.29207/resti.v5i1.2815>
- Sumarni, S., & Rustam, S. (2020). Klasifikasi Topik Tugas Akhir Mahasiswa menggunakan Algoritma Particle Swarm Optimization dan K-Nearest Neighbor. *ILKOM Jurnal Ilmiah*, 12(2), 168–175. <https://doi.org/10.33096/ilkom.v12i2.604.168-175>
- Suryanto, A. A. (2019). Penerapan Metode Mean Absolute Error (Mea) Dalam Algoritma Regresi Linear Untuk Prediksi Produksi Padi. *Saintekbu*, 11(1), 78–83. <https://doi.org/10.32764/saintekbu.v11i1.298>
- Syakir, Y., Iman Hermanto, T., Ramadhan, Y. R., Studi, P., Informatika, T., Teknologi, S. T., & Purwakarta, W. (2022). Analisis Marketplace Shopee Untuk Memprediksi Penjualan dengan Algoritma Regresi Linier. *Jurnal Sains Komputer & Informatika (J-SAKTI)*, 6(2), 904–915.
- Wantari, N. K. (2021). Aplikasi Persamaan Linier Dalam Matematika Bisnis:(Model Persamaan Linier/Harga Keseimbangan Pasar/Suplus Konsumen Atau *Jurnal Dunia Ilmu*, 1(3), 1–8. <http://duniailmu.org/index.php/repo/article/view/46>
- Yudianto, F., Herlambang, T., Anshori, M. Y., Adinugroho, M., & Rulyansah, A. (2023). Sosialisasi Perhitungan Numerik Terkait Forecasting Pengunjung Hotel (Studi di Hotel Primebiz Surabaya). *Indonesia Berdaya*, 4(3), 989–996. <https://doi.org/10.47679/ib.2023511>

Perbandingan Metode TF-IDF dan Bag of Words dalam Analisis Sentimen Diet KopiAmericano di Media Sosial Twitter Menggunakan Naïve Bayes

Rahmatika Suryanti, Putri Taqwa Prasetyaningrum*

Fakultas Teknologi Informasi, Sistem Informasi, Universitas Mercu Buana Yogyakarta, Yogyakarta, Indonesia

Email: ¹201210119@student.mercubuana-yogya.ac.id, ^{2,*} putri@mercubuana-yogya.ac.id

Email Penulis Korespondensi: putri@mercubuana-yogya.ac.id

Submitted: 30/04/2025; Accepted: 31/05/2025; Published: 01/06/2025

Abstrak—Popularitas diet kopi, khususnya varianAmericano, meningkat seiring berkembangnya gaya hidup sehat di masyarakat. Fenomena ini melahirkan berbagai opini publik di media sosial yang perlu dianalisis untuk memahami persepsi konsumen. Penelitian ini bertujuan untuk membandingkan dua metode representasi fitur teks yang umum digunakan, yaitu Term Frequency-Inverse Document Frequency (TF-IDF) dan Bag of Words (BoW), dalam analisis sentimen menggunakan algoritma Naïve Bayes. Data dikumpulkan dari Twitter dengan kata kunci yang relevan dan melalui tahapan preprocessing meliputi case folding, cleansing, tokenizing, stopwords removal, dan stemming. Proses labeling dilakukan secara manual berdasarkan kata kunci sentimen, lalu data diklasifikasikan ke dalam kategori positif, negatif, dan netral. Hasil evaluasi menunjukkan bahwa model klasifikasi dengan TF-IDF memberikan akurasi sebesar 85%, lebih unggul dibandingkan dengan BoW yang memperoleh akurasi 64%. Perbedaan performa ini mengindikasikan bahwa pemilihan metode representasi fitur berperan penting dalam keberhasilan analisis sentimen. Penelitian ini diharapkan dapat menjadi acuan dalam optimalisasi teknik representasi teks untuk memahami opini publik berbasis media sosial, khususnya dalam konteks produk diet dan minuman rendah kalori..

Kata Kunci: Analisis Sentimen; Diet Kopi; Naïve Bayes; TF-IDF; Bag of Words

Abstract—The popularity of diet coffee, particularly theAmericano variant, has risen alongside the growing trend of healthy lifestyles in society. This phenomenon has led to various public opinions circulating on social media, which need to be analyzed to better understand consumer perceptions. This study compares two commonly used text feature representation methods, Term Frequency-Inverse Document Frequency (TF-IDF) and Bag of Words (BoW), in sentiment analysis using the Naïve Bayes algorithm. Using relevant keywords, data were collected from Twitter and underwent preprocessing stages including case folding, cleansing, tokenizing, stopwords removal, and stemming. Sentiment labeling was conducted manually based on keyword indicators, and the data were classified into positive, negative, and neutral categories. The evaluation results show that the TF-IDF model achieved an accuracy of 85%, outperforming BoW which obtained 64%. This performance gap indicates that the choice of feature representation method plays a crucial role in the success of sentiment classification. This research is expected to serve as a reference for optimizing text representation techniques to analyze public opinion on social media, particularly concerning diet products and low-calorie beverages.

Keywords: Sentiment Analysis; Coffee Diet; Naïve Bayes; TF-IDF; Bag of Words

1. PENDAHULUAN

Perkembangan gaya hidup sehat semakin memengaruhi pola konsumsi masyarakat, termasuk dalam memilih jenis minuman. Salah satu tren yang berkembang adalah konsumsi kopi diet, dengan jenisAmericano menjadi pilihan utama karena kandungannya yang rendah kalori serta bebas gula, susu dan krimer. Meningkatnya popularitas kopi ini terlihat dari banyaknya di sosial media, seperti Twitter, Instagram, Youtube dan TikTok, yang membahas manfaat, rasa, serta efek dari jenis kopi ini. Namun, opini publik terhadap diet kopiAmericano cukup beragam. Sebagian mendukung karena rasa yang ringan dan manfaat kesehatannya, tidak sedikit juga merasa rasanya terlalu pahit dan tidak memberikan efek signifikan terhadap program diet. Dengan banyaknya persepsi ini, dibutuhkan suatu metode untuk mengklasifikasikan opini-opini tersebut secara sistematis melalui analisis sentimen. Perkembangan teknologi digital telah memungkinkan analisis sentimen menjadi alat strategis untuk memahami opini publik, sebagaimana dijelaskan dalam penelitian analisis sentimen pada ulasan Google Review oleh Rustiawan dan Prasetyaningrum[1].

Sebagai solusi, analisis sentimen dapat digunakan untuk mengidentifikasi dan mengelompokkan opini publik secara otomatis melalui pendekatan Natural Language Processing (NLP). Dalam penelitian ini, analisis sentimen digunakan untuk mengkaji opini publik terhadap diet kopiAmericano. Langkah awal dalam analisis sentimen adalah mengubah data teks menjadi representasi numerik melalui teknik ekstraksi fitur. Dua metode yang populer digunakan adalah Term Frequency-Inverse Document Frequency (TF-IDF) dan Bag of Words (BoW). TF-IDF menilai pentingnya suatu kata dengan membandingkan frekuensinya dalam satu dokumen terhadap keseluruhan korpus, sedangkan BoW menghitung jumlah kemunculan kata dalam sebuah dokumen tanpa mempertimbangkan konteks atau urutan.

Di Indonesia, Riskesdas tahun 2018 mencatat angka obesitas sebesar 21,8%, meningkat dibandingkan tahun-tahun sebelumnya. Kondisi ini mendorong masyarakat untuk mencari solusi penurunan berat badan secara praktis dan alami, salah satunya melalui konsumsi kopi, terutama kopi hijau dan kopiAmericano, yang dikenal rendah kalori dan dianggap mampu mendukung program diet[2]. Media sosial turut memfasilitasi penyebaran tren ini, menjadikan kopi, termasuk varian kopiAmericano sebagai objek diskusi, promosi, dan penilaian publik. Fenomena diet dengan kopiAmericano pun menjadi perbincangan yang hangat, diwarnai oleh opini yang beragam, baik yang mendukung maupun

yang skeptis[3]. Berbagai penelitian menunjukkan bahwa kopi, baik yang berkafein maupun tidak, memiliki potensi sebagai agen anti-obesitas dengan mekanisme seperti penghambatan akumulasi lemak, peningkatan metabolisme, serta pengaruh terhadap mikrobiota usus [4].

Penelitian Dedy Sugiarto et al. membandingkan performa algoritma Naïve Bayes dan Logistic Regression menggunakan dua metode ekstraksi fitur teks, yaitu TF-IDF dan Bag of Words (BoW), dalam mengklasifikasikan opini publik terhadap kebijakan BLT Minyak Goreng berdasarkan data dari Twitter. Hasilnya menunjukkan bahwa model Logistic Regression dengan BoW menghasilkan akurasi terbaik sebesar 72% dan F1-score 70%[5]. Kurniawan et al. membandingkan metode TF-IDF dan Bag of Words (BoW) dengan algoritma Support Vector Machine (SVM) pada data Twitter yang membahas layanan JNE. Pada penerapan teknik TF-IDF yang dipadukan dengan metode SVM memiliki hasil yang lebih baik dengan nilai Accuracy 86%, Precision 85%, Recall 85% dan F1-Score 85% sedangkan penerapan teknik BOW yang dipadukan metode SVM hanya unggul pada nilai Recall sebesar 89% [6]. Random Forest dan k-NN dikenal sebagai algoritma klasifikasi yang efektif dalam berbagai penelitian prediksi berbasis data kesehatan seperti yang ditunjukkan oleh Supoyo dan Prasetyaningrum [7]. Decision Tree (C4.5) sebagai metode klasifikasi juga terbukti efektif dalam mengelola stok farmasi rumah sakit. Studi oleh Perkasa dan Putri menunjukkan bahwa Naïve Bayes mampu mencapai akurasi hingga 87% dalam klasifikasi sentimen, memperkuat pemilihan metode dalam penelitian ini[8]. Penerapan teknik data mining seperti FP-Growth dan clustering telah terbukti efektif dalam memahami perilaku konsumen di sektor gamifikasi perbankan[9]. Dalam penelitian ini, pendekatan berbasis Data Mining digunakan, sejalan dengan metode prediksi kelulusan mahasiswa menggunakan Multilayer Perceptron (MLP) yang diungkapkan oleh Windarti dan Prasetyaningrum[10]. Metode Support Vector Machine dengan kernel Radial Basis Function (RBF) dan Linear telah dibandingkan sebelumnya dalam analisis kepuasan pelanggan mobile banking[11]. Gamifikasi dalam mobile banking meningkatkan keterlibatan pengguna, sebagaimana diungkapkan melalui pendekatan clustering dan evaluasi classifier yang dilakukan pada pengguna di Indonesia. Dalam kasus dataset tidak seimbang, penggunaan teknik SMOTE-Tomek Links terbukti mampu meningkatkan akurasi klasifikasi hingga 98,7%[12]. Metode K-Nearest Neighbor (KNN) digunakan dalam penelitian ini untuk prediksi klasifikasi berbasis layanan, sebagaimana telah sukses diterapkan dalam prediksi produk layanan Indihome dengan akurasi hingga 99,99%[13]. Penggunaan algoritma Random Forest telah terbukti efektif dalam prediksi kinerja karyawan dan penentuan atribut terbaik dalam dataset, memperkuat penerapan metode ensemble dalam penelitian ini[14]. Dalam penelitian sejenis, analisis sentimen menggunakan SVM kernel RBF pada data mobile banking juga menghasilkan akurasi tinggi, mencapai 93%, menegaskan keunggulan teknik ini dalam klasifikasi teks[11].

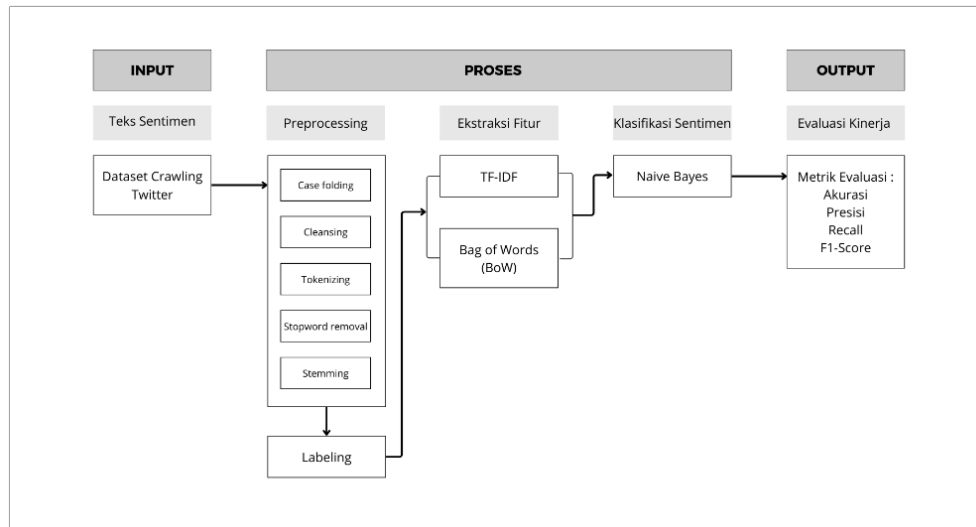
Penelitian oleh Darmawan et al. yang mengkaji opini masyarakat terhadap kebijakan kenaikan harga BBM di Twitter menggunakan algoritma SVM menemukan bahwa BoW memiliki performa terbaik secara keseluruhan berdasarkan F1-Score rata-rata, meskipun TF-IDF lebih unggul dalam precision pada beberapa label[15]. Fauzi dan Yunial melakukan analisis sentimen pada data Twitter maskapai penerbangan AS menggunakan kedua metode vektorisasi tersebut dan membandingkannya dengan berbagai algoritma klasifikasi seperti SVM, Naïve Bayes, dan ensemble voting. Hasilnya menunjukkan bahwa TF-IDF yang dikombinasikan dengan SVM menghasilkan akurasi tertinggi (95%), sementara BoW paling optimal saat digunakan dengan Naïve Bayes (akurasi 92%)[16]. Penelitian Hadi dan Utami mengevaluasi performa algoritma K-Nearest Neighbors (K-NN) menggunakan tiga metode ekstraksi fitur, yaitu Bag of Words (BoW), TF-IDF, dan N-Grams pada klasifikasi ujaran kebencian di Twitter. Hasilnya menunjukkan bahwa kombinasi TF-IDF dan K-NN dengan nilai $k=3$ memberikan performa terbaik dengan akurasi 86,88% dan F1-score 86,50%, mengungguli BoW dan N-Grams dalam semua metrik evaluasi [17]. Dade dan Hani melakukan penelitian analisis sentimen review film menggunakan Naive Bayes dan TF-IDF sebagai ekstraksi fitur memberikan hasil akurasi 86,48% [18]. Meskipun telah banyak dilakukan penelitian analisis sentimen pada berbagai topik, belum ditemukan studi yang secara eksplisit membandingkan performa TF-IDF dan BoW dalam konteks diet kopi Americano di media sosial. Oleh karena itu, penelitian ini memberikan kontribusi dalam memilih teknik representasi teks yang optimal dalam analisis sentimen terhadap produk diet, khususnya kopi Americano di media sosial.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini dilakukan dengan membandingkan dua teknik ekstraksi fitur teks, yaitu TF-IDF (Term Frequency–Inverse Document Frequency) dan Bag of Words (BoW), dalam analisis sentimen ulasan diet kopi Americano di Twitter menggunakan algoritma klasifikasi Naïve Bayes. Data dikumpulkan dari Twitter menggunakan teknik *scraping* atau API Twitter dengan kata kunci “diet kopi”, “kopi Americano”, “kopi hitam”, dan “kopi untuk diet”. Sebanyak 1200 tweet dikumpulkan, dan setelah melalui tahapan preprocessing serta labeling, 236 data digunakan sebagai dataset untuk pelatihan dan pengujian model dengan rasio 70%:30%. Kemudian dilakukan tahapan pra-pemrosesan berupa *case folding*, *cleansing*, *tokenizing*, *stopword removal*, dan *stemming*. Setelah dibersihkan, data diberi label sentimen secara manual menjadi tiga kategori: positif, negatif, dan netral. Data kemudian diubah menjadi vektor numerik menggunakan TF-IDF dan BoW sebagai representasi fitur. Dataset dibagi menjadi data latih dan data uji dengan rasio 70:30. Proses klasifikasi dilakukan menggunakan *Multinomial Naïve Bayes*, dan evaluasi model

dilakukan dengan menggunakan metrik akurasi, presisi, recall, dan F1-score untuk membandingkan performa kedua metode. Diagram alur lengkap penelitian dapat dilihat pada Gambar 1.



Gambar 1. Tahap Penelitian

Pada Gambar 1 menampilkan diagram alur proses penelitian yang menggambarkan langkah-langkah sistematis yang dilakukan oleh peneliti dalam analisis sentimen diet kopiAmericano menggunakan metode TF-IDF dan BoW. Berikut penjabaran setiap tahap yang tergambar dalam diagram tersebut.

2.2 Pengumpulan Data

Data dikumpulkan melalui *web scraping* dari platform media sosial Twitter menggunakan library Python seperti *snsrcape*. Kata kunci pencarian mencakup frasa seperti “diet kopiAmericano”, “kopi diet”, “diet kopi hitam” dalam Bahasa Indonesia. Data yang dikumpulkan difilter berdasarkan rentang waktu satu tahun terakhir agar relevan dengan tren saat ini.

2.3 Pre-processing

Preprocessing adalah tahap dalam analisis sentiment untuk membersihkan dan menyiapkan data teks mentah dari social media Twitter sebelum proses ekstraksi fitur menggunakan TF-IDF dan BoW. Proses ini untuk meningkatkan akurasi model dalam memahami sentimen dari data yang diambil dari Twitter[19]. Berikut tahapan dari preprocessing:

- Case Folding**
Pada proses *Case Folding*, semua huruf kapital dalam tweet diubah menjadi huruf kecil menggunakan fungsi *lower* dalam bahasa pemrograman Python. Proses ini diterapkan setelah pengumpulan data guna memastikan konsistensi teks, sehingga model dapat mengenali pola dan makna analisis dengan lebih baik[20].
- Cleansing**
Cleansing adalah proses yang bertujuan untuk menghilangkan elemen-elemen yang tidak relevan dalam teks, teks dibersihkan dari unsur-unsur yang tidak diperlukan seperti tanda baca, angka, link, mention, hashtag, dan emotikon. Pembersihan dilakukan menggunakan metode regex pada Python[21].
- Tokenizing**
Dalam proses tokenizing, teks dipisahkan menjadi potongan-potongan kecil, biasanya berupa kata atau simbol. Teknik ini diterapkan menggunakan fungsi *word tokenize* dari pustaka NLTK di Python, sehingga setiap elemen teks dapat dianalisis secara terpisah [22].
- Stopword removal**
Stopword Removal merupakan tahap penghapusan kata-kata umum yang dianggap tidak berkontribusi secara signifikan terhadap pemahaman makna teks. Kata-kata seperti "dan", "atau", "yang", "di" dihapus menggunakan daftar stopwords Bahasa Indonesia. Proses ini membantu meningkatkan kualitas data sebelum analisis sentimen dilakukan.[8]
- Stemming**
Stemming merupakan proses normalisasi teks dengan mengonversi kata berimbuhan ke bentuk dasarnya. Pada penelitian ini, teknik stemming dilakukan menggunakan library Sastrawi untuk Bahasa Indonesia guna meningkatkan efektivitas analisis model terhadap struktur kata [23].

2.4 Labeling

Proses pelabelan data tweet yang telah dikumpulkan dianalisis secara manual untuk menentukan apakah memiliki sentimen positif atau negatif terhadap diet kopiAmericano. Hasil labeling ini digunakan sebagai data latih dan data uji pada model klasifikasi[24].

2.5 Ekstraksi Fitur

2.5.1 Term Frequency – Inverse Document Frequency (TF-IDF)

TF-IDF (*Term Frequency-Inverse Document Frequency*) adalah metode statistik yang digunakan untuk menilai seberapa penting sebuah kata dalam suatu dokumen, dengan mempertimbangkan juga keberadaannya di seluruh koleksi dokumen. *Term Frequency (TF)* mengukur seberapa sering sebuah kata muncul dalam dokumen, sedangkan *Inverse Document Frequency (IDF)* mengukur tingkat keunikan kata dalam seluruh koleksi dokumen[25]. Rumus metode *Term Frequency-Inverse Document Frequency (TF-IDF)*:

$$TF - IDF(t, d) = \left(\frac{\text{Jumlah kemunculan kata } t \text{ dalam dokumen } d}{\text{Jumlah total kata dalam dokumen } d} \right) \times \log \left(\frac{\text{Jumlah total dokumen}}{\text{Jumlah dokumen yang mengandung kata } t} \right) \quad (1)$$

2.5.2 Bag of Words (BoW)

Ekstraksi fitur *Bag of Words (BoW)* adalah metode sederhana dalam proses mengubah teks menjadi representasi numerik berupa vektor yang merepresentasikan frekuensi kemunculan kata-kata unik dalam dokumen, tanpa memperhatikan urutan kata [26].

2.6 Klasifikasi Sentimen

Algoritma Naïve Bayes digunakan dalam penelitian ini sebagai metode klasifikasi yang menghitung probabilitas suatu teks termasuk ke dalam kelas sentimen tertentu dengan memanfaatkan frekuensi kemunculan kata-kata pada data teks. Algoritma ini dipilih karena kesederhanaannya serta kemampuan komputasi yang efisien, sehingga sangat cocok untuk analisis sentimen dalam skala besar maupun pada lingkungan dengan keterbatasan sumber daya komputasi[27]. Implementasi klasifikasi sentimen menggunakan Naïve Bayes dilakukan melalui beberapa tahapan, yaitu:

a. Pembagian dataset

Dataset dibagi menjadi dua subset, yaitu data latih sebesar 70% dan data uji sebesar 30%, untuk memastikan evaluasi kinerja model dilakukan secara adil pada data yang tidak terlihat sebelumnya.

b. Pelatihan Model

Klasifier Naïve Bayes dilatih menggunakan data latih yang telah diproses, sehingga model dapat mempelajari pola distribusi kata-kata terhadap label sentimen.

2.7 Evaluasi Model

Evaluasi model bertujuan untuk mengukur kinerja algoritma klasifikasi dalam menentukan kelas sentimen data. Dalam evaluasi ini, beberapa metrik yang digunakan meliputi akurasi, presisi, recall, dan F1-score. Empat komponen utama dalam hasil klasifikasi dijelaskan pada Tabel 1 berikut.

Tabel 1. Confusion Matric

| | Prediksi Positif | Prediksi Negatif |
|-------------------|--------------------|--------------------|
| Kenyataan Positif | True Positif (TP) | False Negatif (FN) |
| Kenyataan Negatif | False Positif (FP) | True Negatif (TN) |

Dimana *True Positive (TP)* jumlah data yang benar-benar termasuk dalam kelas positif dan diklasifikasikan dengan benar sebagai positif oleh model, *True Negative (TN)* jumlah data yang benar-benar termasuk dalam kelas negatif dan diklasifikasikan dengan benar sebagai negatif oleh model, *False Positive (FP)* jumlah data yang sebenarnya negatif tetapi diklasifikasikan secara salah oleh model sebagai positif, *False Negative (FN)* jumlah data yang sebenarnya positif tetapi diklasifikasikan secara salah oleh model sebagai negatif. Model yang telah dilatih kemudian diuji menggunakan data uji. Evaluasi kinerja dilakukan dengan menghitung beberapa metrik utama, yaitu:

a. Akurasi

Akurasi mengukur proporsi prediksi yang benar dibandingkan dengan total data yang diuji. Metrik ini memberikan gambaran umum mengenai seberapa baik model dalam mengklasifikasikan seluruh data uji. Akurasi dihitung dengan rumus. Metrik ini memberikan indikasi umum, tetapi tidak cukup untuk kasus dengan data yang tidak seimbang (misalnya, kelas yang sangat dominan), di mana metrik lain lebih berguna. Secara konseptual, akurasi dihitung dengan menjumlahkan nilai *True Positive (TP)* dan *True Negative (TN)*, kemudian dibagi dengan total data uji yang terdiri dari *TP*, *TN*, *False Positive (FP)*, dan *False Negative (FN)*. Secara matematis ditulis sebagai berikut:

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

b. Presisi

Presisi: Presisi mengukur ketepatan model dalam memprediksi kelas positif (atau negatif) dibandingkan dengan seluruh prediksi yang dibuat sebagai kelas positif. Dengan kata lain, presisi memberi tahu kita seberapa banyak dari semua prediksi positif yang benar-benar positif. Presisi dihitung sebagai rasio antara jumlah prediksi positif yang benar (*True Positive*) terhadap seluruh prediksi yang diklasifikasikan sebagai positif, yaitu jumlah dari *True Positive (TP)* dan *False Positive (FP)*. Rumus presisi secara matematis dinyatakan sebagai berikut:

$$\text{Presisi} = \frac{TP}{TP+FP} \quad (3)$$

c. Recall

Recall mengukur kemampuan model dalam menangkap semua instance kelas positif. Dengan kata lain, recall memberi tahu kita seberapa banyak dari seluruh contoh yang seharusnya diprediksi sebagai positif yang benar-benar diprediksi positif. Recall dihitung sebagai rasio antara jumlah prediksi positif yang benar (TP) terhadap total jumlah instance yang sebenarnya positif, yaitu penjumlahan antara *True Positive (TP)* dan *False Negative (FN)*. Rumus recall dituliskan sebagai berikut:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

d. F1-Score

F1-score adalah rata-rata harmonik dari presisi dan recall. Metrik ini memberikan gambaran yang lebih baik mengenai keseimbangan antara presisi dan recall, terutama ketika ada ketidakseimbangan antara kedua metrik tersebut. F1-Score merupakan metrik yang mengukur keseimbangan antara Presisi dan Recall dalam model klasifikasi, dimana Presisi mengukur akurasi prediksi positif yang benar-benar positif dan Recall mengukur kemampuan model dalam menemukan seluruh data positif. F1-score dihitung dengan rumus berikut.

$$F1 - \text{Score} = 2 \times \frac{\text{Presisi} \times \text{Recall}}{\text{Presisi} + \text{Recall}} \quad (5)$$

F1-Score memberikan nilai yang lebih tinggi jika kedua metrik, Presisi dan Recall, memiliki performa yang seimbang. Metrik ini sangat berguna ketika data tidak seimbang, dengan nilai F1-Score berkisar antara 0 (terburuk) hingga 1 (terbaik).

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Dalam penelitian ini data dikumpulkan dari media sosial Twitter menggunakan metode *web scraping* berbasis *Selenium*. Data diambil secara langsung dari Twitter melalui pencarian dengan kata kunci “diet kopi americano”, “diet kopi hitam”, “kopi untuk diet” dan di filter berdasarkan kategori *live tweet* sehingga menampilkan cuitan terbaru. Proses scraping dilakukan dengan *scroll otomatis* halaman untuk memuat lebih banyak tweet secara dinamis dan mengambil informasi penting seperti tanggal, username, dan isi tweet. Untuk menjaga kualitas data, hanya tweet berbahasa Indonesia yang disimpan setelah melalui proses deteksi bahasa. Dari total 1200 tweet yang dikumpulkan, hanya 236 tweet yang berhasil dilabeli secara manual berdasarkan kriteria kata kunci sentimen. Dari data ini, 165 data digunakan sebagai data latih dan 71 sebagai data uji. Hal ini sesuai dengan rasio pembagian data 70:30. Contoh data hasil dapat dilihat pada Tabel 2.

Tabel 2. Contoh Sample Dataset

| Tanggal | Username | Ulasan |
|--------------------------|----------------|---|
| 2025-04-16T14:00:13.000Z | OHMYBEAUTYBANK | Kak aku juga lagi diet buat kopi dari konten kreator realistis yang aku ikutin sama nonton YouTube nya dr.tirta kopi sehari gapapa apalagi Americano tapi kalau belum bisa ke latte yang kopi sama susu aja jangan pakai gula lagi. Asal batesin dan air putih jangan lupa; |
| 2025-04-15T16:01:28.000Z | OHMYBEAUTYBANK | justru kopi yg pure kopi like americano itu works kak buat diet mwehehhehe JANGAN PAKE FOAM SUGAR SAMA CRUMBLE YY!!!; |
| 2025-03-12T23:17:49.000Z | Eries | Minum kopi kya americano gitu emang bagus buat diet dok?; |
| 2025-02-27T15:29:56.000Z | Dayu Dolma | Ternyata meminum kopi Americano membantu kita diet ya; |
| 2025-02-25T19:53:04.000Z | ristretto | kalian pernah ngelakuin diet dengan konsumsi americano atau kopi hitam? share pengalamannya dong; |

3.2 Pre-Processing

Tahapan preprocessing mencakup pembersihan data, transformasi, dan normalisasi untuk memastikan data siap digunakan dalam ekstraksi fitur dan klasifikasi. Tahap ini bertujuan untuk membersihkan data dari informasi yang

tidak mendukung proses analisis atau pemodelan. Dengan demikian, data yang digunakan dalam penelitian hanya mencakup informasi yang sesuai dan memiliki kontribusi langsung terhadap tujuan analisis.

3.2.1 Case Folding

Pada langkah case folding semua data komentar dalam teks diubah menjadi huruf kecil untuk menyamakan bentuk representasi kata yang memiliki arti sama namun berbeda dalam penulisan kapitalisasi, seperti “Kopi”, “kopi”, dan “KOPI” yang seluruhnya akan diubah menjadi “kopi”. Dengan demikian, case folding dapat mengurangi redundansi dan meningkatkan konsistensi data pada tahap analisis selanjutnya. Dalam penelitian ini, case folding diterapkan pada setiap entri teks dalam dataset menggunakan fungsi *lower* pada bahasa pemrograman Python. Contoh implementasi case folding ditunjukkan pada Tabel 3.

Tabel 3. Sampel Dataset *Case Folding*

| Ulasan | Casefolded |
|---|---|
| Kak aku juga lagi diet buat kopi dari konten kreator realistis yang aku ikutin sama nonton YouTube nya dr.tirta kopi sehari gapapa apalagi Americano tapi kalau belum bisa ke latte yang kopi sama susu aja jangan pakai gula lagi. Asal batesin dan air putih jangan lupa; | kak aku juga lagi diet buat kopi dari konten kreator realistis yang aku ikutin sama nonton youtube nya dr.tirta kopi sehari gapapa apalagi americano tapi kalau belum bisa ke latte yang kopi sama susu aja jangan pakai gula lagi. asal batesin dan air putih jangan lupa; |
| justru kopi yg pure kopi like americano itu works kak buat diet mwehehehe JANGAN PAKE FOAM SUGAR SAMA CRUMBLE YY!!!; | justru kopi yg pure kopi like americano itu works kak buat diet mwehehehe jangan pake foam sugar sama crumble yy!!!; |
| Minum kopi kya americano gitu emang bagus buat diet dok?; | minum kopi kya americano gitu emang bagus buat diet dok?; |
| Ternyata meminum kopi Americano membantu kita diet ya; | ternyata meminum kopi americano membantu kita diet ya; |
| kalian pernah ngelakuin diet dengan konsumsi americano atau kopi hitam? share pengalamannya dong; | kalian pernah ngelakuin diet dengan konsumsi americano atau kopi hitam? share pengalamannya dong; |

3.2.2 Cleansing

Tahap cleansing merupakan proses pembersihan data teks dari karakter-karakter yang tidak diperlukan dalam analisis, seperti tautan (URL), mention pengguna (@username), hashtag (#), angka, tanda baca, serta simbol-simbol khusus lainnya. Proses ini bertujuan untuk menghilangkan elemen-elemen yang tidak memiliki nilai semantik terhadap klasifikasi sentimen. Dalam penelitian ini, cleansing dilakukan secara otomatis menggunakan ekspresi reguler (regular expressions) pada Python, sehingga teks yang dihasilkan lebih bersih, seragam, dan siap diproses pada tahap selanjutnya seperti tokenisasi dan ekstraksi fitur. Berikut adalah contoh sampel dataset pada Tabel 4 sebelum dan sesudah dilakukan cleansing

Tabel 4. Sampel dataset *cleansing*

| Ulasan | Cleansed |
|---|--|
| lagi belajar minum americano biar ngopinya sekalian diet (pict kopi semalem); | lagi belajar minum americano biar ngopinya sekalian diet pict kopi semalem |
| Kak aku juga lagi diet buat kopi dari konten kreator realistis yang aku ikutin sama nonton YouTube nya dr.tirta kopi sehari gapapa apalagi Americano tapi kalau belum bisa ke latte yang kopi sama susu aja jangan pakai gula lagi. Asal batesin dan air putih jangan lupa; | kak aku juga lagi diet buat kopi dari konten kreator realistis yang aku ikutin sama nonton youtube nya drtirta kopi sehari gapapa apalagi americano tapi kalau belum bisa ke latte yang kopi sama susu aja jangan pakai gula lagi asal batesin dan air putih jangan lupa |
| justru kopi yg pure kopi like americano itu works kak buat diet mwehehehe JANGAN PAKE FOAM SUGAR SAMA CRUMBLE YY!!!; | justru kopi yg pure kopi like americano itu works kak buat diet mwehehehe jangan pake foam sugar sama crumble yy |
| Minum kopi kya americano gitu emang bagus buat diet dok?; | minum kopi kya americano gitu emang bagus buat diet dok |
| Ternyata meminum kopi Americano membantu kita diet ya; | ternyata meminum kopi americano membantu kita diet ya |

3.2.3 Tokenizing

Tokenizing merupakan proses pemecahan teks menjadi unit-unit kata yang lebih kecil yang disebut token. Setiap token merepresentasikan satu kata yang akan dianalisis lebih lanjut. Proses ini penting untuk memisahkan kalimat menjadi komponen dasar sehingga memungkinkan sistem untuk memahami dan mengolah kata secara individual.

Dalam penelitian ini, proses tokenisasi dilakukan menggunakan pustaka pemrosesan bahasa alami (Natural Language Toolkit/NLTK) dalam Python. Pada Tabel 5 adalah contoh dataset sebelum dan sesudah dilakukan proses *Tokenizing*.

Tabel 5. Sampel Dataset *Tokenizing*

| Cleansed | Tokenized |
|--|---|
| lagi belajar minum americano biar ngopinya sekalian diet pict kopi semalem | ['lagi', 'belajar', 'minum', 'americano', 'biar', 'ngopinya', 'sekalian', 'diet', 'pict', 'kopi', 'semalem'] |
| kak aku juga lagi diet buat kopi dari konten kreator realitis yang aku ikutin sama nonton youtube nya drtirta kopi sehari gapapa apalagi americano tapi kalau belum bisa ke latte yang kopi sama susu aja jangan pakai gula lagi asal batesin dan air putih jangan lupa | ['kak', 'aku', 'juga', 'lagi', 'diet', 'buat', 'kopi', 'dari', 'konten', 'kreator', 'realitis', 'yang', 'aku', 'ikutin', 'sama', 'nonton', 'youtube', 'nya', 'drtirta', 'kopi', 'sehari', 'gapapa', 'apalagi', 'americano', 'tapi', 'kalau', 'belum', 'bisa', 'ke', 'latte', 'yang', 'kopi', 'sama', 'susu', 'aja', 'jangan', 'pakai', 'gula', 'lagi', 'asal', 'batesin', 'dan', 'air', 'putih', 'jangan', 'lupa'] |
| justru kopi yg pure kopi like americano itu works kak buat diet mwehehehehe jangan pake foam sugar sama crumble yy | ['justru', 'kopi', 'yg', 'pure', 'kopi', 'like', 'americano', 'itu', 'works', 'kak', 'buat', 'diet', 'mwehehehehe', 'jangan', 'pake', 'foam', 'sugar', 'sama', 'crumble', 'yy'] |
| minum kopi kya americano gitu emang bagus buat diet dok | ['minum', 'kopi', 'kya', 'americano', 'gitu', 'emang', 'bagus', 'buat', 'diet', 'dok'] |
| ternyata meminum kopi americano membantu kita diet ya | ['ternyata', 'meminum', 'kopi', 'americano', 'membantu', 'kita', 'diet', 'ya'] |

3.2.4 Stopword Removal

Stopword removal adalah tahap untuk menghilangkan kata-kata umum yang memiliki frekuensi tinggi tetapi kontribusi semantiknya rendah terhadap penentuan sentimen, seperti “dan”, “yang”, “di”, dan sebagainya. Penghapusan stopwords bertujuan untuk mengurangi noise dalam data dan meningkatkan fokus hanya pada kata-kata yang memiliki makna penting. Proses ini dilakukan menggunakan daftar stopwords berbahasa Indonesia yang tersedia dalam pustaka NLTK dan Sastrawi. Berikut pada Tabel 6 merupakan sampel dataset sebelum dan sesudah dilakukan tahap *Stopword Removal*.

Tabel 6. Sampel Dataset *Stopword Removal*

| Tokenized | Stopword_Removed |
|---|---|
| ['kalian', 'pernah', 'ngelakuin', 'diet', 'dengan', 'konsumsi', 'americano', 'atau', 'kopi', 'hitam', 'share', 'pengalamannya', 'dong'] | ['kalian', 'pernah', 'ngelakuin', 'diet', 'konsumsi', 'americano', 'kopi', 'hitam', 'share', 'pengalamannya', 'dong'] |
| ['biasa', 'kopi', 'susu', 'udh', 'mau', 'sebulan', 'ganti', 'ke', 'americano', 'krna', 'emg', 'lg', 'proses', 'diet', 'wkwk'] | ['biasa', 'kopi', 'susu', 'udh', 'mau', 'sebulan', 'ganti', 'americano', 'krna', 'emg', 'lg', 'proses', 'diet', 'wkwk'] |
| ['ternyata', 'kopi', 'americano', 'bisa', 'untuk', 'diet'] | ['ternyata', 'kopi', 'americano', 'diet'] |
| ['kopi', 'diet', 'americano', 'pelangsing', 'badan', 'pembakar', 'lemak', 'ampuh', 'arabika', 'coffee', 'aman', 'untuk', 'lambung', 'dengan', 'harga', 'rp'] | ['kopi', 'diet', 'americano', 'pelangsing', 'badan', 'pembakar', 'lemak', 'ampuh', 'arabika', 'coffee', 'aman', 'lambung', 'harga', 'rp'] |
| ['dan', 'mengurangi', 'rasa', 'lelah'] | ['mengurangi', 'rasa', 'lelah'] |
| ['akuuu', 'jugaa', 'pernah', 'bacaa', 'kak', 'katanyaa', 'kopi', 'itemm', 'tuh', 'baguss', 'malahh', 'buat', 'yg', 'lagi', 'diet', 'kyk', 'americano', 'gituu'] | ['akuuu', 'jugaa', 'pernah', 'bacaa', 'kak', 'katanyaa', 'kopi', 'itemm', 'tuh', 'baguss', 'malahh', 'buat', 'yg', 'diet', 'kyk', 'americano', 'gituu'] |

3.2.5 Stemming

Stemming adalah proses mengubah setiap kata ke bentuk dasarnya (*root word*) untuk menyatukan berbagai variasi kata yang memiliki akar makna yang sama. Sebagai contoh, kata “meminum”, “minuman”, dan “peminum” akan dikembalikan ke bentuk dasar “minum”. Dalam penelitian ini, proses stemming dilakukan menggunakan pustaka Sastrawi, yang merupakan library pemrosesan teks Bahasa Indonesia. Tahap ini bertujuan untuk menyederhanakan bentuk kata dan meningkatkan akurasi klasifikasi sentimen. Pada Tabel 7 adalah contoh dataset sebelum dan setelah dilakukan stemming.

Tabel 7. Sampel dataset *Stemming*

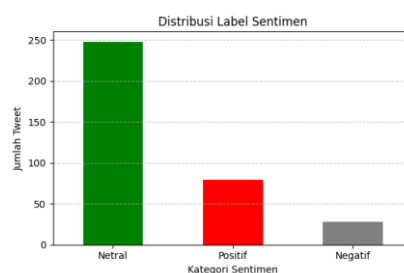
| Cleansed | Stemmed |
|--|--|
| lagi belajar minum americano biar ngopinya sekalian diet pict kopi semalem | lagi ajar minum americano biar ngopinya sekali diet pict kopi semalem |
| kak aku juga lagi diet buat kopi dari konten kreator realitis yang aku ikutin sama nonton youtube nya | kak aku juga lagi diet buat kopi dari konten kreator realitis yang aku ikutin sama nonton youtube nya |

| | |
|---|---|
| kopi sehari gapapa apalagi americano tapi kalau belum bisa ke latte yang kopi sama susu aja jangan pakai gula lagi asal batesin dan air putih jangan lupa | drtirta kopi hari gapapa apalagi americano tapi kalau belum bisa ke latte yang kopi sama susu aja jangan pakai gula lagi asal batesin dan air putih jangan lupa |
| justru kopi yg pure kopi like americano itu works kak buat diet mwehehhehe jangan pake foam sugar sama crumble yy | justru kopi yg pure kopi like americano itu works kak buat diet mwehehhehe jangan pake foam sugar sama crumble yy |
| minum kopi kya americano gitu emang bagus buat diet dok | minum kopi kya americano gitu emang bagus buat diet dok |
| ternyata meminum kopi americano membantu kita diet ya | nyata minum kopi americano bantu kita diet ya |

3.3 Labeling

Proses labeling sentimen bertujuan memberi label (Positif, Netral, Negatif) pada setiap data teks berdasarkan isi ulasannya. ada penelitian ini, proses labeling dilakukan secara semi otomatis menggunakan pendekatan berbasis kata kunci (keyword-based). Daftar kata kunci positif dan negatif disusun berdasarkan kata-kata yang sering digunakan dalam konteks diet kopi, seperti “*enak*”, “*suka*”, “*segar*”, dan “*efektif*” untuk sentimen positif, serta “*pahit*”, “*tidak suka*”, “*lemas*”, dan “*tidak efektif*” untuk sentimen negatif.

Setiap ulasan diperiksa apakah mengandung salah satu dari kata kunci tersebut. Jika ditemukan kata positif, maka teks diberi label Positif. Jika mengandung kata negatif, maka diberi label Negatif. Jika tidak mengandung keduanya, maka teks dikategorikan sebagai Netral. Proses ini dilakukan secara otomatis melalui pemrograman Python, kemudian hasilnya dapat ditinjau ulang secara manual jika diperlukan untuk meningkatkan akurasi pelabelan. Hasil labeling terdapat pada Gambar 2 berikut.



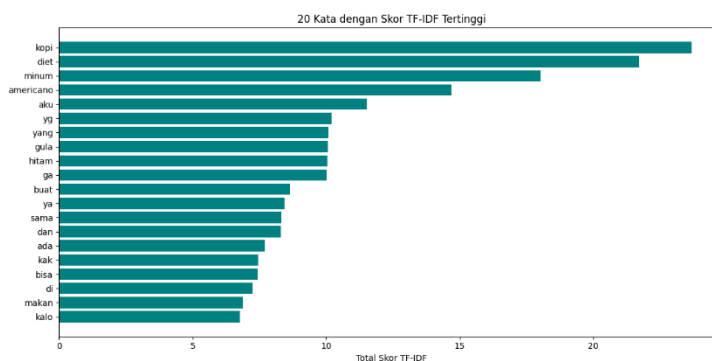
Gambar 2. Hasil Labeling

3.4 Ekstraksi Fitur

Ekstraksi fitur merupakan tahapan yang sangat penting dalam pemrosesan teks karena berfungsi untuk mengubah data teks yang bersifat tidak terstruktur menjadi representasi numerik yang dapat diolah oleh algoritma klasifikasi. Dalam penelitian ini, digunakan dua metode utama untuk ekstraksi fitur, yaitu Term Frequency–Inverse Document Frequency (TF-IDF) dan Bag of Words (BoW).

3.4.1 Term Frequency–Inverse Document Frequency (TF-IDF)

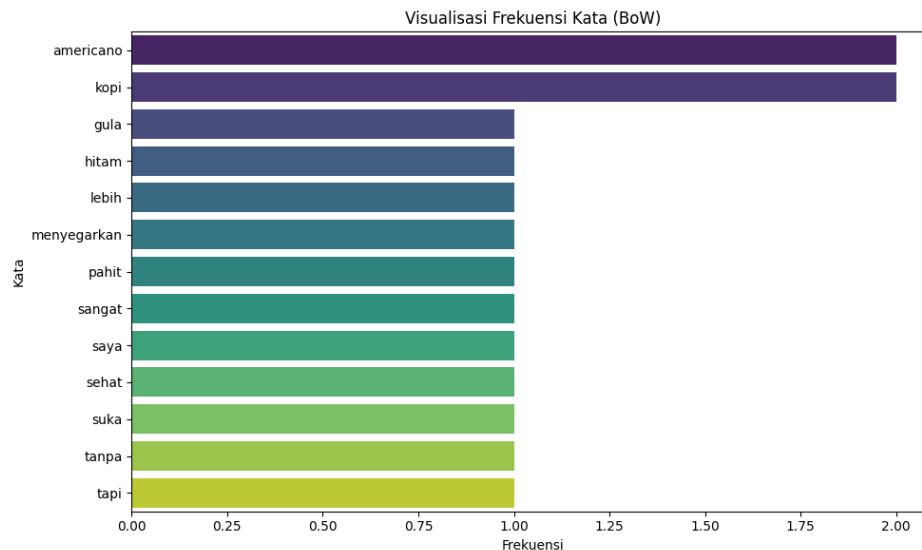
TF-IDF merupakan salah satu metode representasi fitur dalam pemrosesan teks yang memberikan bobot pada setiap kata berdasarkan tingkat kepentingannya dalam suatu dokumen relatif terhadap seluruh dokumen dalam korpus. Semakin sering sebuah kata muncul dalam dokumen tertentu dan semakin jarang kata tersebut muncul dalam dokumen lain, maka bobot TF-IDF-nya akan semakin tinggi. Dalam penelitian ini, TF-IDF digunakan untuk mengekstraksi fitur dari data teks hasil preprocessing dengan tujuan menangkap kata-kata yang bersifat spesifik dan informatif. Implementasi metode ini dilakukan menggunakan *TfidfVectorizer* dari pustaka *scikit-learn*. Berikut Gambar 3 menunjukkan hasil pembobotan kata menggunakan TF-IDF.



Gambar 3. Pembobotan kata menggunakan TF-IDF

3.4.2 Bag of Words (BoW)

Bag of Words adalah metode representasi teks yang sederhana namun efektif, di mana setiap dokumen diubah menjadi vektor berdasarkan frekuensi kemunculan setiap kata dalam korpus. BoW tidak memperhatikan urutan kata dalam kalimat, tetapi hanya menghitung seberapa sering kata tersebut muncul. Meskipun bersifat statistik dan tidak mempertimbangkan konteks, BoW tetap banyak digunakan karena kemudahannya dalam implementasi dan efektivitasnya pada berbagai tugas klasifikasi. Pada penelitian ini, metode BoW digunakan sebagai pembandingan terhadap TF-IDF dalam mengekstraksi fitur dari teks ulasan diet kopi Americano. Visualisasi hasil pembobotan kata menggunakan Bag of Words terdapat pada Gambar 4 berikut.



Gambar 4. Pembobotan Kata menggunakan BoW

3.5 Klasifikasi Sentimen

Klasifikasi sentimen merupakan tahapan penting dalam analisis sentimen, di mana tujuan utamanya adalah mengklasifikasikan data teks ke dalam kategori sentimen tertentu, seperti positif, negatif, atau netral. Proses ini dilakukan setelah tahap ekstraksi fitur, yang mengubah teks menjadi representasi numerik yang dapat diproses oleh algoritma klasifikasi. Dalam penelitian ini, model klasifikasi yang digunakan adalah Naïve Bayes, sebuah algoritma probabilistik yang sering digunakan dalam klasifikasi teks karena kesederhanaannya dan efisiensinya dalam menangani data dalam jumlah besar.

Naïve Bayes mengandalkan probabilitas bersyarat untuk memprediksi kelas sentimen sebuah teks berdasarkan kata-kata yang terkandung dalam teks tersebut. Prinsip dasar dari Naïve Bayes adalah mengasumsikan bahwa setiap kata dalam teks bersifat independen satu sama lain, yang dikenal dengan istilah naïve assumption. Meskipun anggapan ini jarang benar dalam praktiknya, Naïve Bayes tetap memberikan hasil yang cukup baik dalam banyak aplikasi klasifikasi teks, terutama dalam analisis sentimen.

Dalam penelitian ini, Naïve Bayes diterapkan pada data tweet yang telah melalui tahapan preprocessing dan ekstraksi fitur menggunakan metode TF-IDF dan BoW. Klasifikasi dilakukan dengan memanfaatkan frekuensi kata yang telah diubah menjadi representasi numerik dan digunakan untuk mempelajari pola distribusi kata-kata terhadap kategori sentimen. Dengan menggunakan Multinomial Naïve Bayes, model ini dapat mengklasifikasikan teks ke dalam tiga kategori sentimen: positif, negatif, dan netral, berdasarkan probabilitas kata-kata yang terdapat dalam setiap kategori.

3.4.1 Naïve Bayes

Naïve Bayes adalah algoritma klasifikasi yang sangat populer dalam berbagai aplikasi pengolahan bahasa alami (Natural Language Processing/NLP), termasuk analisis sentimen. Algoritma ini bekerja dengan mengasumsikan bahwa semua fitur (kata-kata dalam hal ini) adalah independen, yang berarti setiap kata memiliki kontribusi yang sama terhadap probabilitas suatu kelas sentimen, tanpa memperhatikan hubungan antar kata dalam kalimat. Meskipun asumsi ini sering kali tidak sepenuhnya akurat dalam praktiknya, Naïve Bayes tetap memberikan kinerja yang baik pada banyak tugas klasifikasi teks, termasuk analisis sentimen.

Pada penelitian ini, Multinomial Naïve Bayes digunakan, yang merupakan varian dari Naïve Bayes yang lebih cocok untuk data diskrit, seperti kata-kata dalam teks. Model ini menghitung probabilitas masing-masing kategori (positif, negatif, netral) berdasarkan kata-kata yang muncul dalam teks. Dengan kata lain, Naïve Bayes menghitung peluang setiap kelas berdasarkan kemunculan kata-kata tertentu dalam teks, kemudian memilih kelas yang memiliki probabilitas tertinggi.

Proses klasifikasi Naïve Bayes dimulai dengan pembagian dataset menjadi dua bagian: data latih (training data) dan data uji (testing data). Data latih digunakan untuk melatih model, sementara data uji digunakan untuk mengevaluasi kinerja model. Pada tahap pelatihan, algoritma menghitung probabilitas kemunculan setiap kata dalam setiap kategori sentimen, dan model ini kemudian dapat digunakan untuk memprediksi kategori sentimen dari data uji. Keuntungan utama dari Naïve Bayes adalah kemudahan implementasi dan kemampuannya untuk menangani data dalam jumlah besar dengan cepat. Selain itu, algoritma ini tidak memerlukan banyak data pelatihan untuk memberikan hasil yang baik, sehingga sangat cocok untuk aplikasi analisis sentimen di media sosial di mana volume data sangat besar dan terus berkembang.

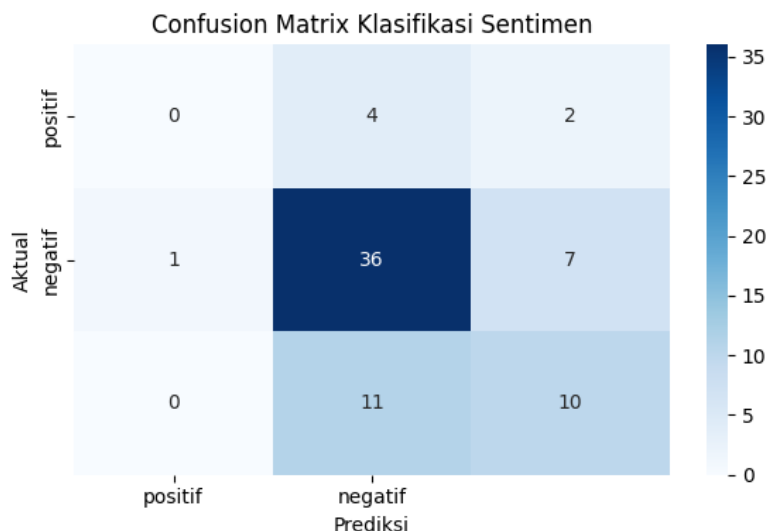
3.5 Evaluasi Model

Secara umum kelas netral memiliki nilai evaluasi tertinggi di seluruh metrik. Kelas negatif memiliki performa paling rendah (precision dan recall = 0), yang menunjukkan bahwa model kesulitan dalam mengenali ciri khas dari sentimen negatif pada data. Macro average F1-score sebesar 0.42, dan weighted average sebesar 0.62, mengindikasikan bahwa ketidakseimbangan distribusi kelas berdampak terhadap kinerja model. Tabel 8 berikut metrik evaluasi berdasarkan per kelas.

Tabel 8. Metrix evaluasi per kelas

| Kelas | Precision | Recall | F1-Score | Support |
|---------|-----------|--------|----------|---------|
| Negatif | 0.00 | 0.00 | 0.00 | 6 |
| Netral | 0.71 | 0.82 | 0.76 | 44 |
| Positif | 0.53 | 0.48 | 0.50 | 21 |

Klasifikasi sentimen menggunakan algoritma Naïve Bayes menghasilkan akurasi sebesar 64% pada data uji sebanyak 71 tweet. Nilai ini diperoleh dari pengujian model dengan representasi fitur tertentu (misalnya TF-IDF atau BoW) terhadap tiga kelas sentimen, yaitu positif, negatif, dan netral. Berdasarkan confusion matrix, dapat diketahui sentimen positif cenderung sulit diklasifikasikan dengan benar. Dari 6 data berlabel positif, tidak ada yang berhasil diklasifikasikan dengan benar (0), sebagian besar justru diklasifikasikan sebagai netral (4) atau negatif (2). Sentimen netral memiliki kinerja klasifikasi terbaik, dengan 36 dari 44 data berhasil diklasifikasikan secara benar. Sentimen negatif cukup ambigu, dengan hanya 10 dari 21 data yang berhasil diklasifikasikan dengan benar, dan sebagian besar salah diklasifikasikan sebagai netral. Pada Gambar 5 menunjukkan confusion matrix hasil klasifikasi sentimen.



Gambar 5. Confusion Matrix

Hasil ini menunjukkan bahwa Naïve Bayes cenderung bias terhadap kelas mayoritas, dalam hal ini sentimen netral, dan memiliki kesulitan untuk mendeteksi sentimen minoritas seperti negatif. Hal ini mungkin disebabkan oleh ketidakseimbangan data (class imbalance), serta fitur-fitur yang kurang representatif untuk sentimen negatif. Penggunaan metode balancing atau teknik augmentasi data dapat dipertimbangkan untuk perbaikan di penelitian selanjutnya.

Evaluasi model adalah tahap penting dalam setiap proses pembelajaran mesin, termasuk dalam klasifikasi sentimen. Evaluasi bertujuan untuk mengukur seberapa baik model yang dibangun dalam mengklasifikasikan data uji ke dalam kategori yang tepat. Pada penelitian ini, evaluasi dilakukan dengan menggunakan beberapa metrik utama, yang meliputi akurasi, presisi, recall, dan F1-score. Masing-masing metrik ini memberikan informasi yang berbeda mengenai kinerja model dalam mengklasifikasikan sentimen. Hasil perbandingan *confusion matrix* terdapat pada Tabel 9 berikut.

Tabel 9. Perbandingan Confusion Matrix

| | Precision | Recall | F1-score | Support |
|----------|--------------------|---------------------|--------------------|-------------------|
| Negatif | 0.0 | 0.0 | 0.0 | 6.0 |
| Netral | 0.7058823529411765 | 0.8181818181818182 | 0.7578947368421053 | 44.0 |
| Positif | 0.5263157894736842 | 0.47619047619047616 | 0.5 | 21.0 |
| Accuracy | 0.647887323943662 | 0.647887323943662 | 0.647887323943662 | 0.647887323943662 |

Evaluasi dilakukan dengan membandingkan hasil klasifikasi yang dihasilkan oleh model Naïve Bayes dengan data sentimen yang telah diberi label secara manual. Hasil evaluasi ini memberikan gambaran tentang seberapa baik model dapat mengidentifikasi sentimen yang sebenarnya dalam data teks, serta bagaimana model dapat diperbaiki lebih lanjut untuk meningkatkan kinerjanya. Dengan menggunakan metrik evaluasi ini, penulis dapat mengukur efektivitas model yang dibangun dan memberikan rekomendasi untuk perbaikan pada penelitian selanjutnya.

4. KESIMPULAN

Penelitian ini telah berhasil membandingkan dua metode populer dalam representasi fitur teks, yaitu Term Frequency-Inverse Document Frequency (TF-IDF) dan Bag of Words (BoW), untuk analisis sentimen terkait diet kopiAmericano di media sosial, khususnya Twitter. Berdasarkan hasil yang diperoleh, dapat disimpulkan bahwa meskipun kedua metode dapat diterapkan dengan baik dalam klasifikasi sentimen, terdapat perbedaan yang signifikan dalam hal akurasi yang dihasilkan. Metode TF-IDF terbukti memberikan akurasi yang lebih tinggi, yang menunjukkan kemampuannya dalam menangkap kata-kata yang lebih relevan dan spesifik dalam konteks opini diet kopiAmericano. Sementara itu, metode BoW, meskipun lebih sederhana, cenderung memiliki keterbatasan karena tidak mempertimbangkan urutan kata dan konteks kalimat. Hal ini mengindikasikan bahwa pemilihan metode ekstraksi fitur harus disesuaikan dengan karakteristik dan kebutuhan analisis data yang sedang diteliti. Model klasifikasi Naïve Bayes yang menggunakan TF-IDF sebagai metode representasi fitur berhasil mencapai akurasi sebesar 85%, sedangkan BoW hanya mencapai 64%. Hasil ini menegaskan bahwa TF-IDF lebih efektif dalam menangkap opini spesifik pengguna Twitter terhadap diet kopiAmericano. Penelitian ini juga memberikan wawasan yang penting dalam hal optimalisasi teknik representasi teks untuk analisis sentimen berbasis media sosial. Hasil yang diperoleh menegaskan bahwa pemilihan teknik yang tepat dapat berpengaruh besar terhadap performa model klasifikasi. Namun, terdapat beberapa keterbatasan yang perlu diperhatikan. Salah satunya adalah pengumpulan data dari Twitter yang memiliki keterbatasan jumlah tweet yang dapat diambil, yang mungkin mempengaruhi representativitas dataset yang digunakan. Selain itu, proses pelabelan sentimen secara manual memiliki potensi bias, yang bisa mempengaruhi hasil klasifikasi. Untuk itu, pengembangan penelitian selanjutnya dapat mencakup penggunaan teknik pelabelan otomatis atau semi-otomatis, serta memperbesar dataset dengan melibatkan lebih banyak sumber data dari berbagai platform media sosial untuk memperkaya analisis. Ke depan, penelitian ini juga dapat diperluas dengan mengeksplorasi penerapan algoritma klasifikasi lainnya, seperti Support Vector Machine (SVM) atau Random Forest, yang telah terbukti efektif dalam klasifikasi teks dalam berbagai bidang. Penelitian juga dapat mempertimbangkan penggunaan metode representasi teks yang lebih canggih seperti Word2Vec atau BERT, yang dapat meningkatkan kemampuan model dalam memahami konteks kata secara lebih mendalam. Dengan langkah-langkah tersebut, penelitian ini tidak hanya memberikan kontribusi dalam analisis sentimen pada produk diet kopi, tetapi juga membuka peluang untuk pengembangan teknik analisis sentimen yang lebih akurat dan efektif di masa depan.

REFERENCES

- [1] M. Rizqi, A. Rustiawan, and P. T. Prasetyaningrum, "Analisis Sentimen Terhadap Klinik Natasha Skincare di Yogyakarta Dengan Metode Google Review," *J. Inf. Technol. Ampera*, vol. 5, no. 1, pp. 2774–2121, 2024, doi: 10.51519/journalita.v5i1.556.
- [2] Ni Luh Wayan Sita Pujasari and Ni Made Widi Astuti, "Potensi Biji Kopi Hijau (Green Bean Coffee) Sebagai Suplemen Penurun Berat Badan," *Pros. Work. dan Semin. Nas. Farm.*, vol. 1, no. 1, pp. 213–229, 2023, doi: 10.24843/wsnf.2022.v01.i01.p18.
- [3] A. H. Nasution, D. A. Fitri, M. S. Qolbu, and A. Sunarto, "Prosiding Seminar Nasional Manajemen Analisis Komunitas Penggemar Kopi : Dinamika Sosial dan Pengaruh Terhadap Tren Konsumsi Kopi," *Pros. Semin. Nas. Manaj.*, vol. 2, no. 1, pp. 251–256, 2023.
- [4] A. V. Sirotkin and A. Kolesarova, "The Anti-Obesity and Health-Promoting Effects of Tea and Coffee," *Physiol. Res.*, vol. 70, no. 2, pp. 161–168, 2021, doi: 10.33549/physiolres.934674.
- [5] Dedy Sugiarto, Ema Utami, and Ainul Yaqin, "Perbandingan Kinerja Model TF-IDF dan BOW untuk Klasifikasi Opini Publik Tentang Kebijakan BLT Minyak Goreng," *J. Tek. Ind.*, vol. 12, no. 3, pp. 272–277, Dec. 2022, doi: 10.25105/jti.v12i3.15669.
- [6] K. Tri Putra, M. Amin Hariyadi, and C. Crysdian, "Perbandingan Feature Extraction Tf-Idf Dan Bow Untuk Analisis Sentimen Berbasis Svm," *J. Cahaya MAndalika*, vol. 3, no. 2, p. 1449, 2023, doi: 10.36312/jcm.v3i2.
- [7] A. Supoyo and P. T. Prasetyaningrum, "Analisis Data Mining Untuk Memprediksi Lama Perawatan Pasien Covid-19 Di DIY," *Bianglala Inform.*, vol. 10, no. 1, pp. 21–29, 2022, doi: 10.31294/bi.v10i1.11890.
- [8] A. E. Perkasa and A. N. Putri, "Penerapan Naïve Bayes Untuk Analisis Sentimen Pada Ulasan Aplikasi Mobile Legends,"

- Build. Informatics, Technol. Sci.*, vol. 6, no. 4, p. 2152–2164, 2025, doi: 10.47065/bits.v6i4.6507.
- [9] P. T. Prasetyaningrum, P. Purwanto, and A. F. Rochim, “Consumer Behavior Analysis in Gamified Mobile Banking : Clustering and Classifier Evaluation,” *J. Syst. Manag. Sci.*, vol. 15, no. 2, pp. 290–308, 2025, doi: 10.33168/JSMS.2025.0218.
 - [10] M. Windarti and P. T. Prasetyaningrum, “Prediction Analysis Student Graduate Using Multilayer Perceptron,” *Atl. Press*, vol. 440, no. Icobl 2019, pp. 53–57, 2020, doi: 10.2991/assehr.k.200521.011.
 - [11] P. T. Prasetyaningrum, N. T. Kadir, and A. Y. Chandra, “Comparison Of Support Vector Machine Radial Base And Linear Kernel Functions For Mobile Banking Customer Satisfaction Analysis,” *Int. J. Comput. Netw. Secur. Inf. Syst.*, vol. 4, no. 1, pp. 10–16, 2022, doi: 10.33005/ijconsist.v4i1.75.
 - [12] P. T. Prasetyaningrum, P. Purwanto, and A. F. Rochim, “Enhancing Element Game Classification: Effective Techniques for Handling Imbalanced Classes,” *Int. J. Intell. Eng. Syst.*, vol. 17, no. 1, pp. 555–571, 2024, doi: 10.22266/ijies2024.0229.47.
 - [13] A. U. Haspriyanti and P. W. Prasetyaningrum, “Penerapan Data Mining Untuk Prediksi Layanan Produk Indihome Menggunakan Metode K-Nearst Neighbor Arwa,” *Inf. Syst. Artif. Intell.*, vol. 20, no. 2, pp. 100–107, 2021, doi: 10.26486/jisai.v1i2.17.
 - [14] P. Taqwa Prasetyaningrum, I. Pratama, and A. Yakobus Chandra, “Implementation Of Machine Learning To Determine The Best Employees Using Random Forest Method,” *Ijconsist Journals*, vol. 2, no. 02, pp. 53–59, 2021, doi: 10.33005/ijconsist.v2i02.43.
 - [15] B. Darmawan, A. Dwi Laksito, M. Resa, A. Yudianto, and A. Sidauruk, “Krea-TIF: Jurnal Teknik Informatika Analisis Perbandingan Ekstraksi Fitur Teks pada Sentimen Analisis Kenaikan Harga BBM,” *J. Mhs. Inform.*, vol. 11, no. 1, pp. 53–63, 2023, doi: 10.32832/krea-tif.v11i1.13819.
 - [16] A. Fauzi and A. H. Yunial, “Analisis Sentimen Pada Media Sosial Menggunakan Perbandingan Algoritma Data Mining,” *J. Edukasi dan Penelit. Inform.*, vol. 10, no. 2, p. 277, 2024, doi: 10.26418/jp.v10i2.76024.
 - [17] K. Hadi and E. Utami, “Analysis of K-NN with the Integration of Bag of Words, TF-IDF, and N-Grams for Hate Speech Classification on Twitter,” *JUITA J. Inform.*, vol. 12, no. 2, p. 289, Nov. 2024, doi: 10.30595/juita.v12i2.23829.
 - [18] M. T. Razaq, D. Nurjanah, and H. Nurrahmi, “Analisis Sentimen Review Film Menggunakan Naive Bayes Classifier dengan Fitur TF-IDF,” *e-Proceeding Eng.*, vol. 10, no. 2, pp. 1698–1712, 2023.
 - [19] D. Darwis, N. Siskawati, and Z. Abidin, “Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter Bmkg Nasional,” *J. Tekno Kompak*, vol. 15, no. 1, p. 131, 2021, doi: 10.33365/jtk.v15i1.744.
 - [20] A. Gerliandeva, Y. H. Chrisnanto, and H. Ashaury, “Optimasi Klasifikasi Sentimen pada Komentar Online menggunakan Multinomial Naïve Bayes dan Ekstraksi Fitur TF-IDF serta N-grams Optimization of Sentiment Classification on Online Comments using Multinomial Naïve Bayes and TF-IDF Feature Extraction and N-g,” *J. Pekommas*, vol. 9, no. 2, pp. 259–272, 2024, doi: 10.56873/jpkm.v9i2.5585.
 - [21] T. A. Dewi and E. Mailoa, “Perbandingan Implementasi Metode Smote Pada Algoritma Support Vector Machine (Svm) Dalam Analisis Sentimen Opini Masyarakat Tentang Mixue,” *J. Indones. Manaj. Inform. dan Komun.*, vol. 4, no. 3, pp. 849–855, 2023, doi: 10.35870/jimik.v4i3.289.
 - [22] F. M. Lubis and M. Ikhsan, “Analisis Sentimen Terhadap Program Kampanye Tabrak Prof Pada Media Sosial X Dengan Menggunakan Metode Support Vector Machine,” *JSiI J. Sist. Inf.*, vol. 12, no. 1, pp. 86–92, 2025, doi: 10.30656/jsii.v11i2.9065.
 - [23] L. Efrizoni, S. Defit, M. Tajuddin, and A. Anggrawan, “Komparasi Ekstraksi Fitur dalam Klasifikasi Teks Multilabel Menggunakan Algoritma Machine Learning,” *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 21, no. 3, pp. 653–666, 2022, doi: 10.30812/matrik.v21i3.1851.
 - [24] H. Firda *et al.*, “Perbandingan Pelabelan Rating - based dan Inset Lexicon - based dalam Analisis Sentimen Menggunakan SVM (Studi Kasus : Ulasan Aplikasi GoBiz di Google Play Store),” *Sist. J. Sist. Inf.*, vol. 14, no. 2, pp. 516–528, 2025, doi: 10.32520/stmsi.v14i2.4795.
 - [25] D. Septiani and I. Isabela, “Analisis Term Frequency Inverse Document Frequency (TF-IDF) Dalam Temu Kembali Informasi Pada Dokumen Teks,” *SINTESIA J. Sist. dan Teknol. Inf. Indones.*, vol. 1, no. 2, pp. 81–88, 2023, doi: 10.37058/innovatics.v6i2.12404.
 - [26] I. K. Dwipayoga and M. Agung, “Komparasi Ekstraksi Fitur BoW dan TF-IDF untuk Klasifikasi SMS Menggunakan Naive Bayes,” *J. Nas. Teknol. Inf. dan Apl.*, vol. 3, no. 2, pp. 247–254, 2025, doi: 10.24843/JNATIA.2025.v03.i02.p03.
 - [27] A. Saekhu, D. Intan, and S. Saputra, “Enhancing Student Sentiment Classification on AI in Education using SMOTE and Naive Bayes,” *Build. Informatics, Technol. Sci.*, vol. 6, no. 4, pp. 2165–2174, 2025, doi: 10.47065/bits.v6i4.6469.

LAMPIRAN

Lampiran A Biodata Peneliti

| | | |
|----------------------|---|-------|
| Nama Lengkap | : | <hr/> |
| NIM | : | <hr/> |
| Alamat Asal | : | <hr/> |
| | | <hr/> |
| | | <hr/> |
| | | <hr/> |
| Alamat di Yogyakarta | : | <hr/> |
| | | <hr/> |
| | | <hr/> |
| | | <hr/> |
| No. HP. | : | <hr/> |
| e-Mail | : | <hr/> |
| Website | : | <hr/> |

Lampiran B *Scan* Dokumen Bimbingan Skripsi

Dokumen bimbingan skripsi yang dilampirkan berbentuk hasil scan yang telah terisi lengkap berikut tanda tangan dosen pembimbing.