Available at https://journalenrichment.com/index.php/jr/

Enrichment: Journal of Multidisciplinary Research and Development

Covid-19 Sentiment Analysis Using Random Forest Classification

Salsa Safira Nur Syamsi[1*], Asep Id Hadiana[2], Yulison H. Chrisnanto[3]

[1*, 2, 3] Universitas Jenderal Achmad Yani, Indonesia

*Email: salsasafira300301@gmail.com, asep.hadiana@lecture.unjani.ac.id, y.chrisnanto@lecturer.unjani.ac.id

ARTICLE INFO

Keywords:
Analisis Sentimen
Covid-19
Random Forest
Twitter

ABSTRACT

The spread of the COVID-19 pandemic has reached a significant global scale, changing the dynamics of people's lives around the world. Social media platforms such as Twitter have become important channels for individuals to share experiences, voice opinions, and participate in discussions related to this pandemic. Sentiment analysis emerged as an important approach to reveal changes in people's attitudes and emotions in facing this challenge. This research involves analyzing sentiment during the COVID-19 pandemic to understand the feelings, attitudes, and views of the community after the peak phase of the pandemic. This study refers to previous findings which show that the Random Forest Algorithm provides the highest accuracy in this analysis. Through testing with the Random Forest Algorithm method, model accuracy testing is carried out using a confusion matrix and comparing test data and training data in a ratio of 80:20. Test results show that this model achieves an accuracy rate of 91%, providing a more comprehensive view of changes in public sentiment during the COVID-19 pandemic.

INTRODUCTION

The spread of COVID-19 has reached global pandemic levels, affecting the lives of people all over the world, During this period, social media and online news platforms especially Twitter became important channels for individuals to experience, express opinions, and discuss the pandemic (Keahlian et al., 2021).

During the COVID-19 pandemic phase, there were changes in public sentiment towards situations related to the pandemic, such as government policies, and various social and economic impacts, Sentiment analysis makes it possible to identify positive, negative, or neutral changes, sentiment analysis during the COVID-19 pandemic aims to understand changes in people's emotions, attitudes and opinions after the peak of the pandemic (Harahap, n.d.). One method that can be used to carry out sentiment analysis is using the Random Forest algorithm (Bayu Baskoro et al., n.d.). The Random Forest algorithm is an effective machine learning method in carrying out text-based sentiment analysis because it can handle a large number of text features and can classify text effectively (Fadiyah Basar et al., 2022).

Based on previous research conducted by M. Aldean (2022), in their research, the Random Forest classification method was used with a case study of the Sinovac vaccine. This research states that this method is used as an approach for analysis. 1500 tweets with data using the TF-IDF algorithm by balancing the data using SMOTE then trained using the Random Forest algorithm and validated using the K-Fold Cross Validation and Confusion Matrix techniques to evaluate model performance (Amardita et al., 2022). The research results show that public sentiment towards Sinovac Vaccination is positive (Dwiki et al., 2021). The model built can predict the sentiment of a tweet with an accuracy of up to 79%. Apart from that, the Precision value reaches 85%, Recall reaches 90%, and the F1 Score reaches 88%. This indicates that the model has good abilities in predicting sentiment from tweet data related to Sinovac vaccination

Research by Villavicencio et al. (2021) examines sentiment analysis for the COVID-19 vaccine in the Philippines by applying the Naive Bayes and TF-IDF methods. Researchers used a dataset containing 993 sentences from tweets. This study classified sentiment into three main labels: positive, neutral, and negative. The evaluation results show that the developed model has an accuracy rate of 81.77%, indicating its ability to

recognize and classify diverse sentiments regarding COVID-19 vaccination in the Philippines. By combining these analytical techniques, this research provides valuable insight into the public's views on COVID-19 vaccination in the country.

Therefore, based on the background described previously, there have been many previous studies that have studied sentiment analysis, but there has been no research that specifically focuses on sentiment analysis during the COVID-19 pandemic on the Twitter platform using the TF-IDF and Random Forest methods (Agarwal et al., 2011). as well as supporting algorithms such as pre-processing several experiments and Hyperparameter Optimization so that people can find out opinions, views and thoughts on public views related to topics related to the COVID-19 pandemic to categorize people's positive, negative and neutral comments on social media. Twitter, using models in data mining.

METHOD

Research method is a method used to conduct research. In this research there are 6 process stages, namely the first stage of data collection, then the second stage of text preprocessing, followed by the data labeling stage then tf-idf, then the classification modeling stage with random forest, and finally the confusion matrix.

Data Collection

The initial stage is data collection, which is carried out through the "Crawling" technique of data from Twitter social media using the keyword "after COVID-19". The amount of crawling data collected was 2,000 tweets. The crawling process is carried out between January 1, 2021, and December 30, 2022.

Text Preprocessing

Text preprocessing is the process of cleaning text so that the data can meet the requirements for execution. Text preprocessing is important because the state of the text affects the accuracy of the results (Larasati et al., 2022). The purpose of preprocessing is so that the data obtained will be more structured so that it is easier to process the data.

Data labeling

At this stage, sentiment labeling was carried out on opinion tweets during the COVID-19 pandemic using a dictionary with positive, negative, and neutral labels using a lexicon. The Lexicon dictionary applied in this research is the InSet Lexicon dictionary, which was taken from previous research conducted by Doni Winarso (Nooryuda Prasetya & Winarso, n.d.). The InSet Lexicon dictionary contains several words that cover both positive and negative sentiments and has a certain value weighted for each word. This lexicon dictionary includes 3609 words with positive sentiment and 6609 words with negative sentiment (Nooryuda Prasetya & Winarso, n.d.). In this research, several words were added related to the topic of the Covid-19 pandemic. Using lexicon-based, we were able to classify sentences according to the issues contained in the sentence. The purpose of data labeling is to provide information or labels on text data as positive, negative, or neutral.

TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is a weighting method for pre-processed datasets. In this research, text data is converted into numerical representation using N-grams consisting of Unigrams and Bigrams. The purpose of weighting words (terms) is to give weight to each word (term), especially in the text document that is being processed. The Term Frequency-Inverse Document Frequency (TF-IDF) method offers a simple approach but produces good results

RESULTS AND DISCUSSION

Implementation

In implementation, the software will be applied to the previously planned software design. The implementation process will be carried out using a personal computer to facilitate system testing. In this research, the software was built using the Python programming language with the help of Visual Studio Code tools and using the MySQL database as the backend, with the front end using HTML, JavaScript, and JQuery.

Interface Implementation

The interface system displays to make it easier for users to use the sentiment analysis system during the COVID-19 pandemic.

Implementation of the Dashboard Page

The dashboard page can be accessed by the user. When the user opens the sentiment analysis system website during the COVID-19 pandemic, the first thing that appears is the dashboard page. The dashboard page is shown in Figure 1.

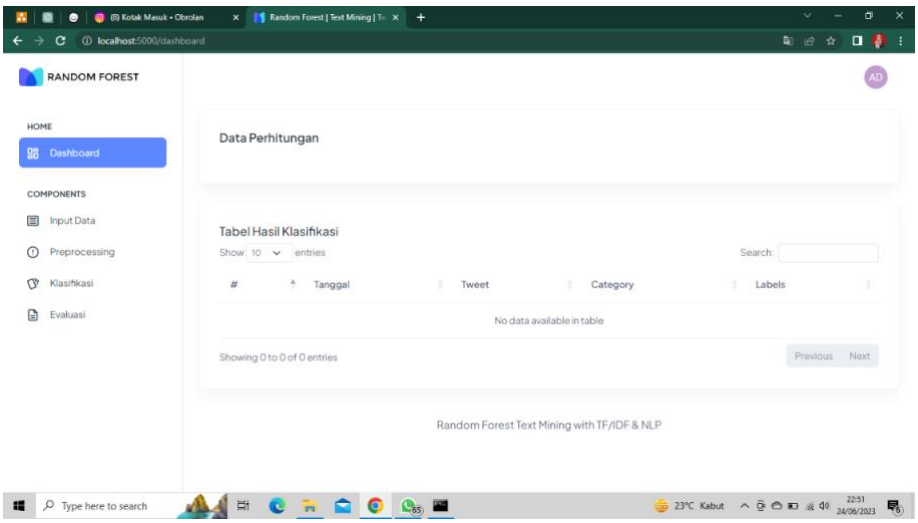


Figure 1. Implementation of the Dashboard Page

Implementation of Data Input Page

The data input page is accessed by the user. The data input page displays the table that will be imported into the system. The data input page is shown in Figure 2.

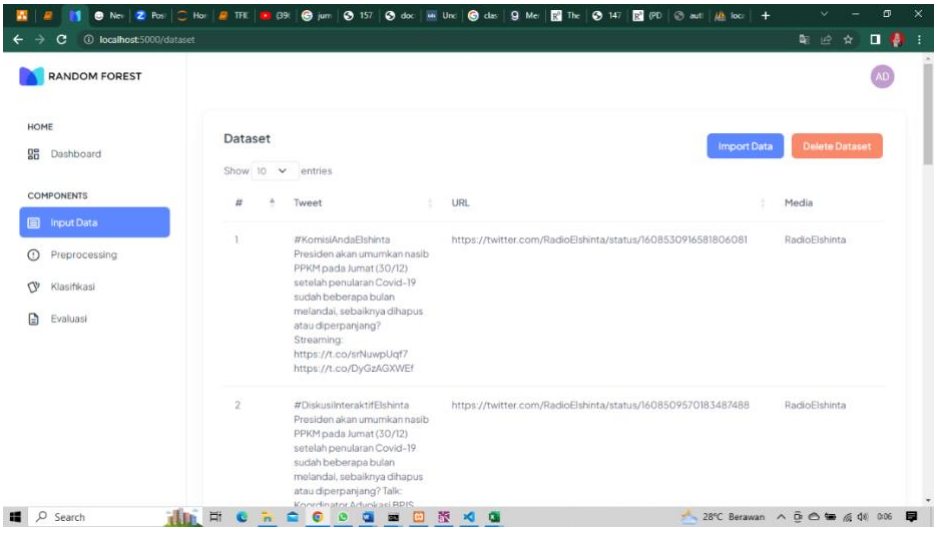


Figure 2. Implementation of the Data Input Page

Implementation of Preprocessing Pages

This preprocessing page can be accessed by the user. This preprocessing display contains tables that have been previously input, then this preprocessing can carry out processing and labeling. The preprocessing page is shown in Figure 3.

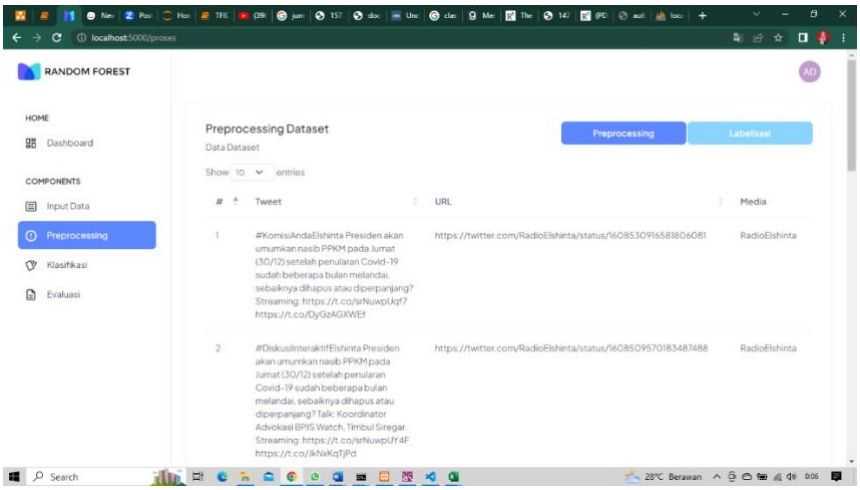


Figure 3. Preprocessing Page Implementation

Implementation of the Classification Page

This classification page can be accessed by users. This classification display contains tables that have been previously processed and labeled, then this classification can calculate the shape of the model. The classification page is shown in Figure 4.

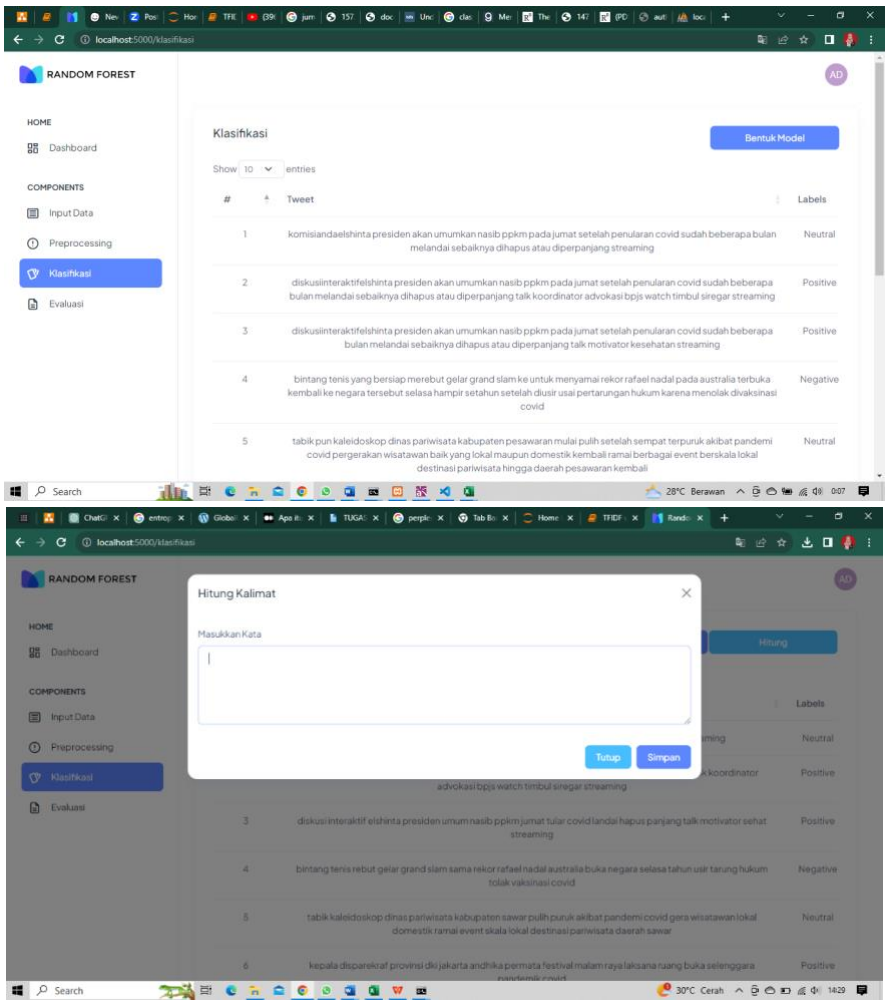


Figure 4. Implementation of the Classification Page

Implementation of the Evaluation Page

This evaluation page can be accessed by users. This evaluation display contains the number of confusion matrix calculations, training data, test data, and classification results that have been previously processed. The evaluation page can be seen in Figure 5.

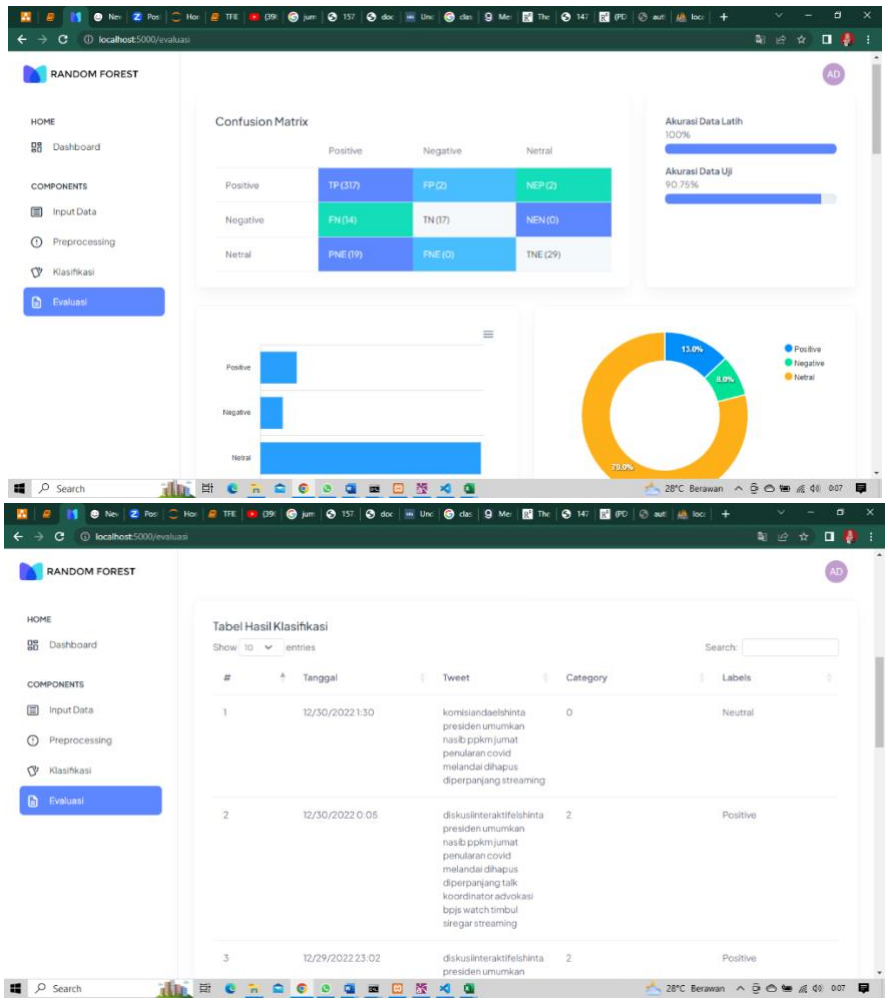


Figure 5. Evaluation Page implementation

Black Box Testing

Black Box testing is a test carried out to observe execution results through test data and check the functionality of the software being built.

Testing Stages

The steps taken when testing this software include the activities described below (D. Kurniawan & Kom, n.d.):

1. Determine the objectives of quality testing
2. Determine the category of quality test results
3. Design quality testing based on use case grouping
4. Implementation of quality testing
5. Conclusion and quality testing results.

Quality Testing Objectives

This stage discusses the objectives of quality testing of the software being built. The purpose of the test is explained in Table 1.

Table 1. Quality Testing objectives

No.	Use Case	Purpose
1.	Data Input	Testing actors, to carry out the data import process
2.	Preprocessing	Testing actors, to process data and label data

No.	Use Case	Purpose
3.	Classification	Carry out testing on actors, to see the results of data that has been preprocessed and labeled, then you can calculate classification in the form of a model
4.	Evaluation	Conduct testing on actors, to see the results of confusion matrix calculations, training data, test data, and classification results

Test Success Categories

The category for determining quality testing on software is divided into two categories, namely (T. A. Kurniawan, 2018):

1. Appropriate, If the quality and function of the software being tested are by the planning objectives and its use, then it is included in the Appropriate category.
2. Not Appropriate, If the quality and function of the software being tested is not by the planning objectives and its use, then it is included in the Unsuitable category.

Testing Scenarios

The purpose of designing quality testing of the software that has been built is to serve as a reference in carrying out quality testing of the software that has been built. The test scenario is shown in Table 2.

Table 2. Test Scenarios

Use Case	Feature Name	Test Code	Test Case
Input Data	Import Data	SS1.1	This feature functions to carry out the data import process
	Delete Dataset	SS1.2	This feature functions to delete datasets
Preprocessing	Preprocessing	SS2.1	This feature functions for processing case folding, cleansing, tokenizing, and stopword data.
	Labelisasi	SS2.2	This feature functions to label data that has been preprocessed
Classification	Model Shape	SS3	This feature functions to calculate the classification process, Confusion matrix, test data, and training data
Evaluation	Evaluation	SS4	This feature functions to display the results of the model shape calculations in the classification menu

Test Implementation

At this stage, testing the quality of the review system software that has been built during the Covid-19 pandemic. Testing is carried out by looking at the design references that have been made, and then adjusting the test results to the objectives to be achieved from the design that has been made (Paramitha et al., n.d.). The test implementation is shown in Table 3.

Table 3. Test Implementation

No	Test Code	System Response	Expected Results	Results
1	SS1.1 (Data Import)	Carry out the data import process	The system saves the dataset that has been imported	In accordance
	SS1.2 (Delete Dataset)	Delete the dataset.	The system deletes data that has been deleted.	In accordance
2	SS2.1 (Preprocessing)	Perform data processing case folding, cleansing, tokenizing, and stopword.	The system displays the preprocessing results in the classification menu	In accordance
	SS2.2 (Labelization)	Displays labeling results.	Displays labeling results in the classification menu.	In accordance
3	SS3 (Model Form)	Carry out the process of calculating classification, Confusion matrix, test data, and training data	The system processes the calculation results	In accordance
4	SS4 (Evaluation)	Displays the results of the calculation of the model form in the classification menu.	The system displays the results	In accordance

Furthermore, the conclusion of the black box testing that has been carried out is in Table 3 with a total of 4 functions, namely data input, preprocessing, classification, and evaluation results (Stt-pln et al., 2016). So the percentage of function suitability in the system can be calculated as follows:

Number of Test Codes = 6 Test Codes
Test Codes with Corresponding Results = 6 Test Codes
Test Codes with Inappropriate Results = 0 Test Codes
Percentage

$$Percentage = \frac{(number\ of\ test\ codes - non - matching\ test\ codes)}{(number\ of\ test\ codes)} \times 100\%$$

=((6-0))/((6)) ×1 00%

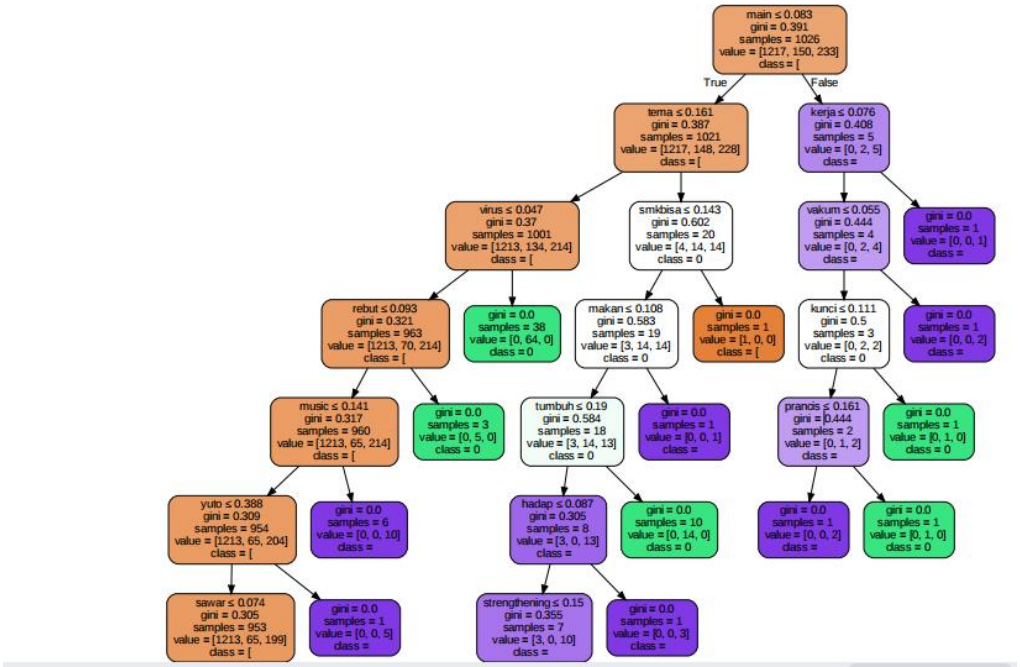
=100%

The results of the calculation of the system suitability function can be concluded that testing of the review software during the COVID-19 pandemic using black box testing has run according to the specifications that have been set with a percentage of 100%.

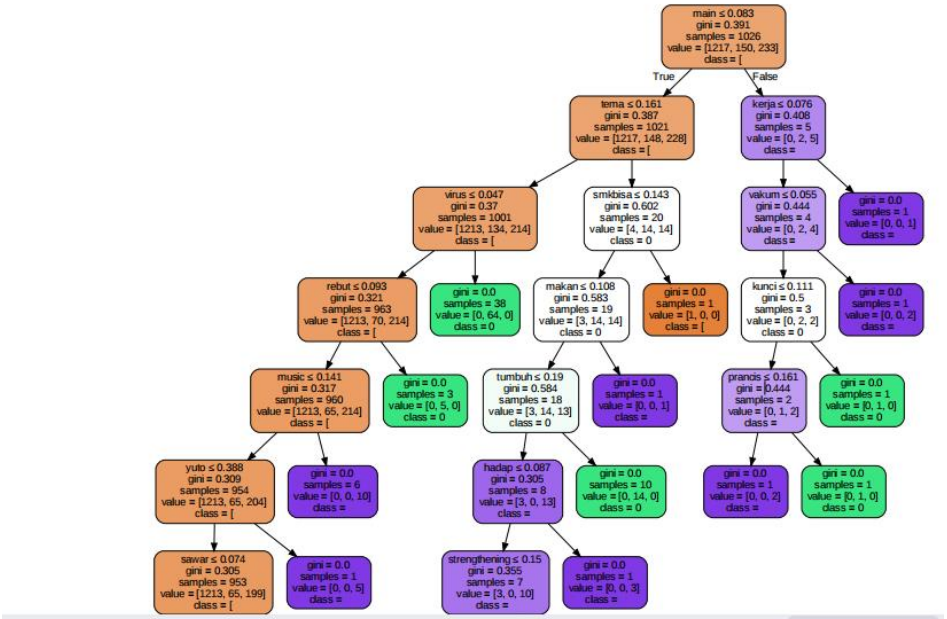
Random Forest Testing

This decision tree visualization can help to understand how the decision tree model makes decisions so that it can understand how data is separated based on certain attributes and how these rules guide the final decision. For example, in the picture below, out of 100 trees, only 5 trees were taken:

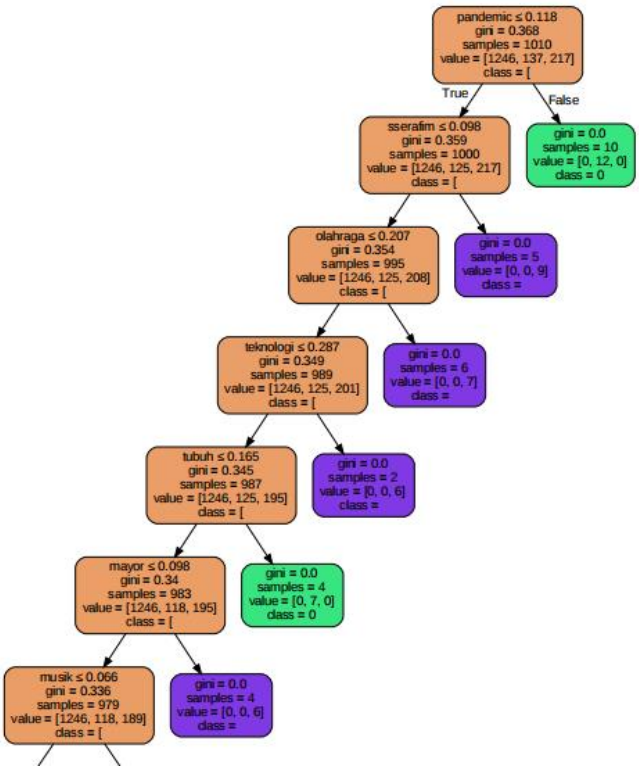
The first tree is a decision tree with a depth of 8.



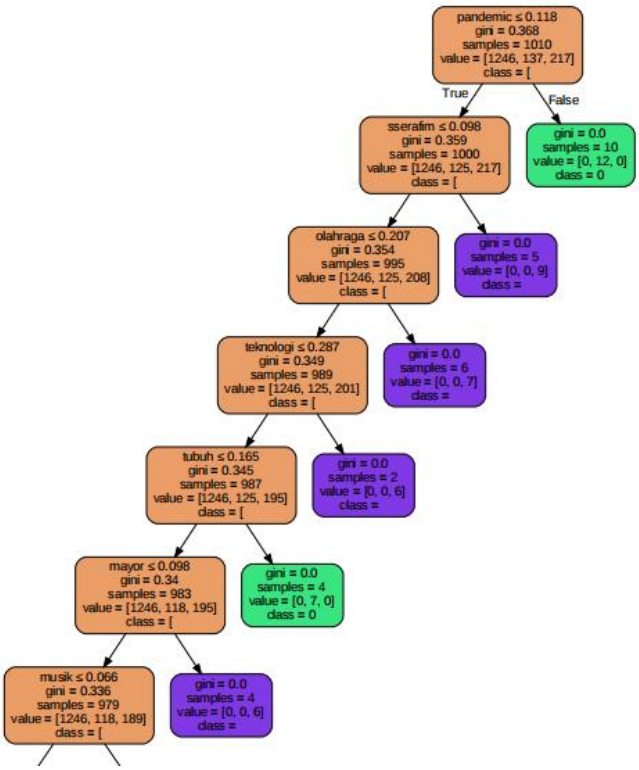
The first tree is a decision tree with a depth of 7.



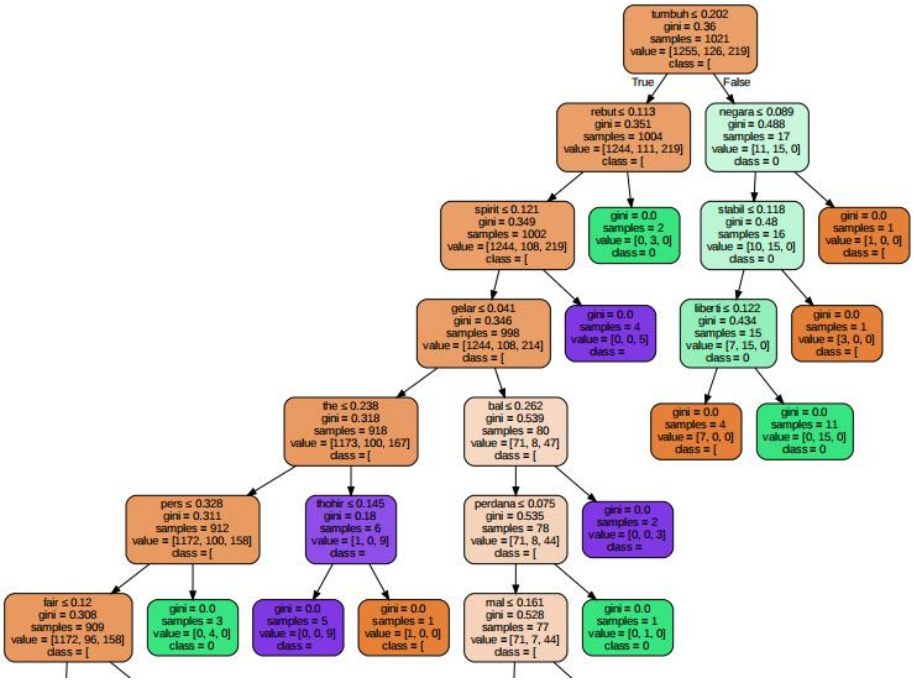
The first tree is a decision tree with a depth of 7.



The first tree is a decision tree with a depth of 7.



The first tree is a decision tree with a depth of 7.



Model Testing with Confusion Matrix

At this stage, model testing is carried out using the Confusion Matrix to test the quality of the model that has been designed by looking at accuracy metrics. This test uses the Confusion Matrix method to compare model predictions with actual tables in the test data. Testing. The system tested has 3 classes, namely Positive, Negative and Neutral. This test was carried out using test data with a dataset ratio of 80:20 from the total data of 2001 data.

Testing of classification results was carried out using test data with a dataset comparison of 8:20 from the total data of 2001. Data carried out using the Random Forest model was evaluated using the confusion matrix method. The value of the confusion matrix can be seen in Figure 6. Model Testing with Confusion Matrix below.

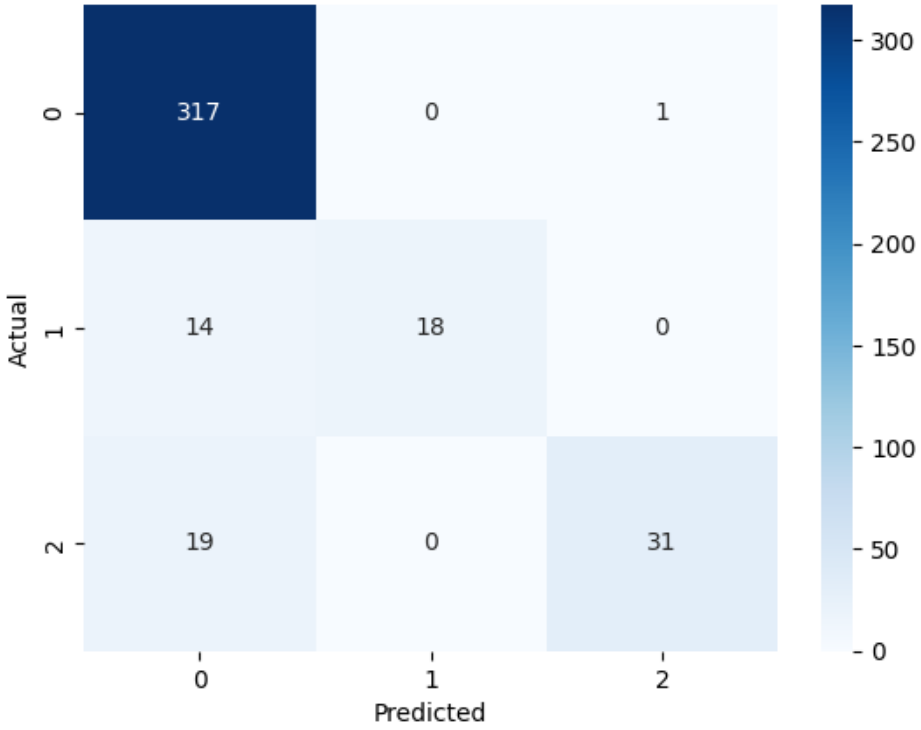


Figure 6 Model Testing with 80:20 Confusion Matrix

The values obtained in the confusion matrix can be used to calculate accuracy with the following formula.

$$Accuracy = \frac{Total\ True\ Positives}{Total\ Data} \times 100$$

Total True Positives can be calculated by adding up the True Positives for classes 0, 1 and 2.

$$Total\ True\ Positif = TP_0 + TP_1 + TP_2$$
$$Total\ True\ Positif = 317 + 18 + 31$$
$$Total\ True\ Positif = 366$$

Total data is obtained from the number of datasets used to test the model, namely 400 records so that accuracy can be calculated using the following calculations.

$$Accuracy = \frac{366}{400} \times 100$$
$$Accuracy = 0{,}91 \times 100$$
$$Accuracy = 91\ \%$$

Testing of classification results was carried out using test data with a dataset comparison of 60:40 from the total 2001 data. Data carried out using the Random Forest model was evaluated using the confusion matrix method. The value of the confusion matrix can be seen in Figure 7. Model Testing with Confusion Matrix below.

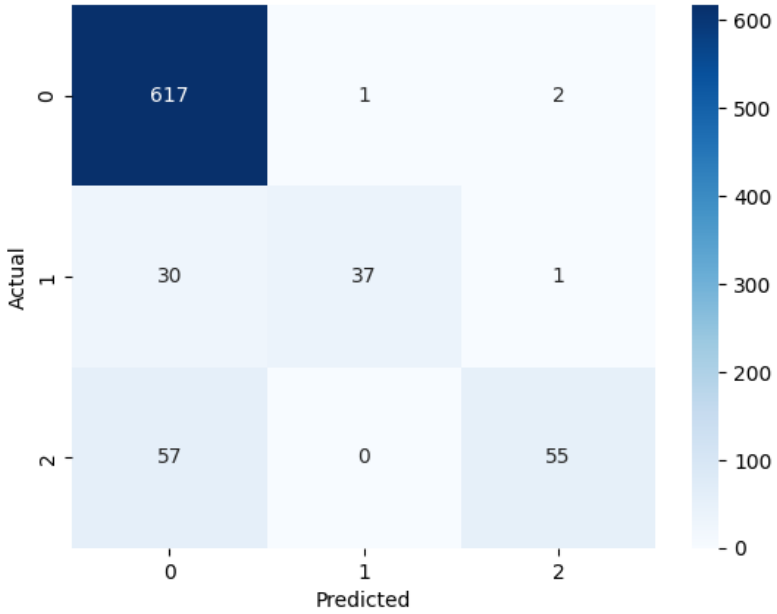


Figure 7. Model Testing with 60:40 Confusion Matrix

The values obtained in the confusion matrix can be used to calculate accuracy with the following formula.

$$Accuracy = \frac{Total\ True\ Positives}{Total\ Data} \times 100$$

Total True Positives can be calculated by adding up the True Positives for classes 0, 1 and 2.

$$Total\ True\ Positif = TP_0 + TP_1 + TP_2$$
$$Total\ True\ Positif = 617 + 37 + 55$$
$$Total\ True\ Positif = 709$$

Total data is obtained from the number of datasets used to test the model, namely 800 records so that accuracy can be calculated using the following calculations.

$$Accuracy = \frac{709}{800} \times 100$$
$$Accuracy = 0{,}8862 \times 100$$
$$Accuracy = 88{,}62\ \%$$

Testing of classification results was carried out using test data with a dataset comparison of 70:30 from the total 2001 data. Data carried out using the Random Forest model was evaluated using the confusion matrix method. The value of the confusion matrix can be seen in Figure 8. Model Testing with Confusion Matrix below.

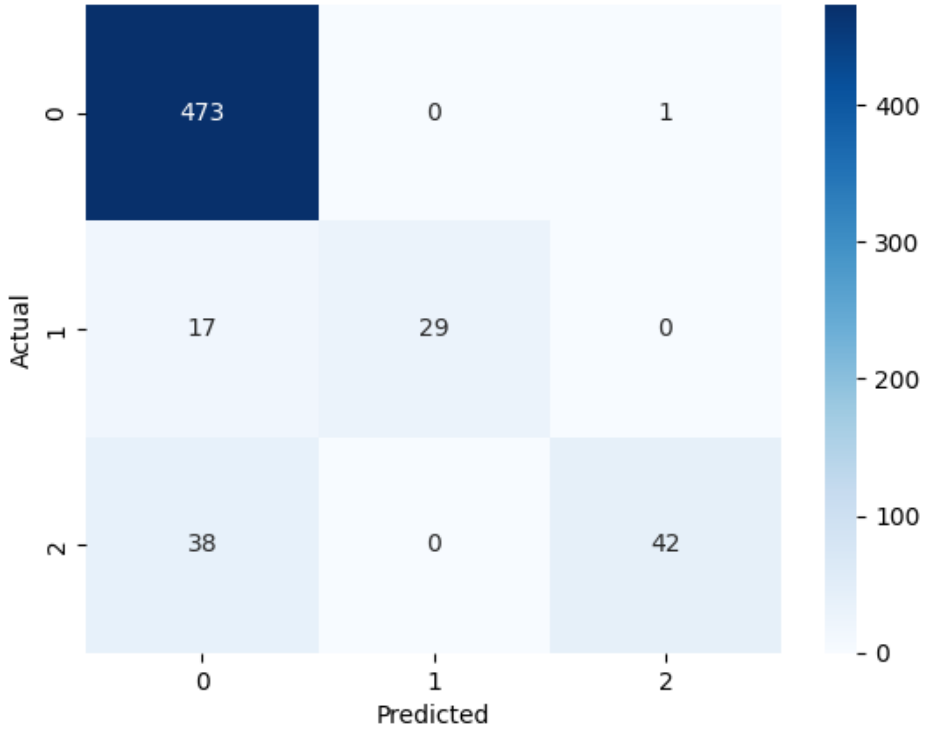


Figure 8. Model Testing with 70:30 Confusion Matrix

The values obtained in the confusion matrix can be used to calculate accuracy with the following formula.

$$Accuracy = \frac{Total\ True\ Positives}{Total\ Data} \times 100$$

Total True Positives can be calculated by adding up the True Positives for classes 0, 1 and 2.

$$Total\ True\ Positif = TP_0 + TP_1 + TP_2$$
$$Total\ True\ Positif = 473 + 29 + 42$$
$$Total\ True\ Positif = 544$$

Total data is obtained from the number of datasets used to test the model, namely 600 records so that accuracy can be calculated using the following calculations.

$$Accuracy = \frac{544}{600} \times 100$$
$$Accuracy = 0{,}906 \times 100$$
$$Accuracy = 90{,}6\ \%$$

Confusion Matrix Testing Conclusion

After testing the confusion matrix, it can be concluded that the results obtained to measure the performance of the Random Forest Classification method using 3 experiments on the confusion matrix with a ratio of 80:20 have an accuracy of 92%, then a ratio of 60:40 has an accuracy of 88%, and finally, the 70:30 ratio has 90% accuracy. So it can be concluded that the ratio of 80:20 is the highest accuracy value, namely 91%. The following are the conclusions from testing the method with the confusion matrix shown in Table 4.

Table 4. Conclusions from Confusion Matrix Testing

Class	Presisi	Recall	F1-Score
Negative	0.9080459770114943	0.9844236760124611	0.944693572496263
Neutral	0.9	0.5806451612903226	0.705882352941176
Positive	0.90625	0.6041666666666666	0.725

CONCLUSION

Based on the research results above, it can be concluded that the Random Forest Classification algorithm has been proven to provide a high level of accuracy in classifying public sentiment during the COVID-19 pandemic, namely 91%. This shows that the Random Forest model can be relied on to predict sentiment with a high level of accuracy. By using the Random Forest Classification Model, sentiment can be successfully classified into three main categories positive, negative, and neutral. This sentiment grouping provides well-structured information about people's feelings and facilitates understanding of their response to the pandemic situation.

REFERENCES

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). *Sentiment Analysis of Twitter Data*. Association for Computational Linguistics. http://www.webconfs.com/stop-words.php

Amardita, R. S., Adiwijaya, A., & Purbolaksono, M. D. (2022). Analisis Sentimen terhadap Ulasan Paris Van Java Resort Lifestyle Place di Kota Bandung Menggunakan Algoritma KNN. *JURIKOM (Jurnal Riset Komputer)*, *9*(1), 62. https://doi.org/10.30865/jurikom.v9i1.3793

Bayu Baskoro, B., Susanto, I., Khomsah, S., Informatika, P., Sains Data, P., Teknologi Telkom Purwokerto Jl Panjaitan, I. DI, & Tengah, J. (n.d.). *Analisis Sentimen Pelanggan Hotel di Purwokerto Menggunakan Metode Random Forest dan TF-IDF (Studi Kasus: Ulasan Pelanggan Pada Situs TRIPADVISOR)*. *3*(2), 21–029. https://doi.org/10.20895/INISTA.V3I2

Dwiki, A., Putra, A., & Juanita, S. (2021). *Analisis Sentimen Pada Ulasan Pengguna Aplikasi Bibit Dan Bareksa Dengan Algoritma KNN*. *8*(2). http://jurnal.mdp.ac.id

Fadiyah Basar, T., Ratnawati, D. E., & Arwani, I. (2022). *Analisis Sentimen Pengguna Twitter terhadap Pembayaran Cashless menggunakan Shopeepay dengan Algoritma Random Forest* (Vol. 6, Issue 3). http://j-ptiik.ub.ac.id

Harahap, I. (n.d.). *Informasi dan Teknologi Ilmiah (INTI)*.

Keahlian, K., Data, R., Luthfika Fairuz, A., Dias Ramadhani, R., Annisa, N., & Tanjung, F. (2021). *JURNAL DINDA Analisis Sentimen Masyarakat Terhadap COVID-19 Pada Media Sosial Twitter*. http://journal.ittelkom-pwt.ac.id/index.php/dinda

Kurniawan, D., & Kom, M. (n.d.). *USE CASE DIAGRAM*.

Kurniawan, T. A. (2018). Pemodelan Use Case (UML): Evaluasi Terhadap beberapa Kesalahan dalam Praktik. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *5*(1), 77. https://doi.org/10.25126/jtiik.201851610

Larasati, F. A., Ratnawati, D. E., & Hanggara, B. T. (2022). *Analisis Sentimen Ulasan Aplikasi Dana dengan Metode Random Forest* (Vol. 6, Issue 9). http://j-ptiik.ub.ac.id

M. Aldean, P. P. and N. S. N. (2022). Analisis Sentimen Masyarakat Terhadap Vaksinasi Covid-19 di Twitter Menggunakan Metode Random Forest Classifier (Studi Kasus: Vaksin Sinovac). *Journal of Informatics Information System Software Engineering and Applications (INISTA)*.

Nooryuda Prasetya, Y., & Winarso, D. (n.d.). *Doni Winarso 2)*. *11*(2).

Paramitha, A., Kom, S., & Kom, M. (n.d.). *Diagram Aktivitas (Activity Diagram) Pertemuan 4*.

Stt-pln, M., Informatika STT-PLN, D. T., Lingkar Luar Barat, J., & Kosambi, D. (2016). *Aplikasi Buku Digital Bidang Teknologi Informasi Berbasis Android Mobile Pada Perpustakaan Bppki Surabaya Badan Litbang Kementerian Kominfo [1]adithya Marhaendra Kusuma, [2]efy Yosrita Application Of Information Technology Digital Book Based On Android Mobile At Library Of Bppki Surabaya Badan Litbang Ministry Of Kominfo* (Vol. 5, Issue 2).

Villavicencio, C., Macrohon, J. J., Inbaraj, X. A., Jeng, J. H., & Hsieh, J. G. (2021). Twitter sentiment analysis towards COVID-19 vaccines in the Philippines using naïve bayes. *Information (Switzerland)*, *12*(5). https://doi.org/10.3390/info12050204