International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Performance Measurement of Confusion Matrix Accuracy in Sentiment Analysis with Decision Trees, Naïve Bayes, K-Nearest Neighbor Methods Using Rapidminer

Amrin Fakhruddin Jauhari. S. Kom[1], Dr. Anggraeni Ridwan, SKom., MMSI [2]
[1]Department of Computer Since, Gunadarma University, Jakarta, Indonesia, 16424.
[2]Department of Computer Since, Gunadarma University, Jakarta, Indonesia, 16424.

Abstract**—***Sentiment analysis is a technique for analyzing many opinions or views of others, speed and ease in disseminating information, encouraging a lot of research on sentiment analysis with various classification techniques such as, Naive Bayes, K-Nearest Neighbor, and Decision Tree, encouraging authors to do research on performance measurement tests of the algorithm, using the standard confusion matrix measurement method to determine the value of accuracy, recall, and precision. in sentiment analysis modeling using 10,590 data samples, with a comparison of 8,472 training data, and 2,118 test data, and using the Naive Bayes, KNN, and decision tree methods, assisted with the rapitminer application. the result of this research is a performance measurement information on the Naive Bayes, KNN and decision tree methods presented through graphs and tables, using modules contained in the rapitminer application.*

Keywords**—** *Sentiment analysis, Decision tree, KNN, Naive Bayes, Rapidminer.*

I. INTRODUCTION

Text mining is a fairly new and interesting science from a field of computer science, text mining solves excess information by combining techniques from data mining, machine learning, information search, knowledge management and Natural Language Processing (NLP), the application of text mining has a purpose to extract and find information that is still hidden from forecasting statistical patterns of a text data, both structured and unstructured so as to produce new knowledge [1].

Research conducted by Ridi Ferdina [2]. About the Indonesian Dataset for Sentiment Analysis: This dataset, which is comprised of tweets from social media platform Twitter, is produced as an Indonesian-language datashet publication called Indonesian-general-sentiment-analysis-datashet. The resulting data amounted to 10,806 tweets and has been labeled with sentiment polarity, which is positive, negative, and neutral. Sentiment data is taken through Twitter social media randomly, so it can be used in sentiment analysis modeling with general topics.

With so much research on sentiment analysis using a variety of objects and methods, the author wants to know how to measure the performance of the confusion matrix, the accuracy, recall and precision of the Naive Bayes algorithm, KNN, and decision tree, using datahset objects whose topics are general in nature, so that the data which is tested free from various interests of others and ensures the dataset has been selected with a good linguist. these measurements are aided by using the existing framework in the rapitminer application.

II. REALETED WORK

Research from sentiment analysis has done quite a lot including Nanang Ruhyana [3] in his journal entitled Sentiment Analysis on the application of odd / even number plate systems on Twitter with the naïve Bayes classification method.

Omman Somantri and Dairoh [4] also carried out additional research using different techniques, titled Sentiment analysis assessment of Tegal city tourist destination based on text mining. They came to the conclusion that the Decision Tree and Naïve Bayes method could be used as a model for the sentiment analysis on the assessment tourist destination.

Furthermore, research by Akhmad Devianto and M. Didik R.W. [5] with the title application of sentiment analysis on twitter users using the K-Nearest Neighbor method. Who conducts research on candidate pairs in the 2017 DKI elections.

Based on these studies, there are various methods that can be used, Oluwakemi CA et al [6] also said that there are many methods or Machine Learning Aligorithms (MLA) that can be used to classify a text, one of which is the Decision trees, Naïve Bayes (NB), and K-Nearest Neighbor (KNN).

III. SENTIMENT ANALYSIS PROCESS

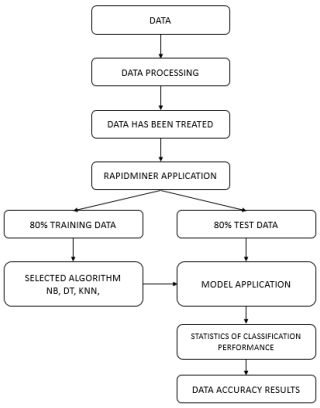


Fig.1. Problem Solving Framework

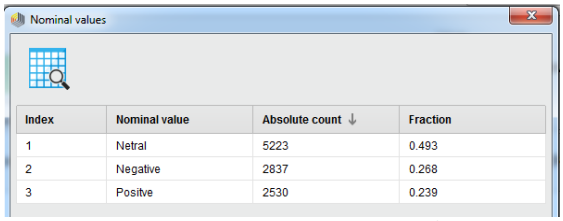
Amrin Fakhruddin Jauhari and Dr. Anggraeni Ridwan, "Performance Measurement of Confusion Matrix Accuracy in Sentiment Analysis with Decision Trees, Naïve Bayes, K-Nearest Neighbor methods Using Rapidminer," *International Research Journal of Advanced Engineering and Science*, Volume 8, Issue 4, pp. 123-127, 2023.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

This type of research conducted in this study is a comparative study by comparing different objects, namely the decision tree algorithm, Naïve Bayes (NB), and K-Nearest Neighbor (KNN) using the Rapidminer tool. In this comparative experiment based on the renewal of the problem updates that can be seen in Figure 1.

A. Understanding Data

At this stage there are several things done to understand the data to be processed. The data taken is a datashet that has been made by previous researchers which can be downloaded at the github.com website address: https:// github.com/ ridife/ dataset-idsa. There are (5,223) tweets labeled Neutral, (2,837) tweets labeled Negative, and (2,530) tweets labeled Positive, as attached in Figure 2.



Index	Nominal value	Absolute count	Fraction
1	Netral	5223	0.493
2	Negative	2837	0.268
3	Positve	2530	0.239

Fig.2. The Nominal Amount Distribution Of Tweets

The adjusted data was divided in half with a comparison of 80% of training data, and 20% of test data. Data that has been divided, named into test data, and training data. Data that has been successfully imported will be displayed with a table along with a description of the number of sample data. 8,472 for training data, and 2,118 for test data. as in figure 3

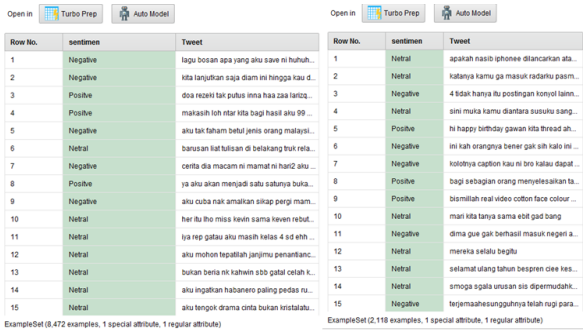


Fig.3. Comparison of Training Data Samples And Test Data

B. Modeling

Modeling is done using the Rapidminer Framework, using modules such as retrive, Apply Model, Performance classification, and algorithm modules for the classification methods of Naïve Bayes, Decision trees, and K-Nearest Neighbor (KNN). To obtain predictive opportunities for trial results, experiments using other methods are also used to see the accuracy of each method as a comparison.

At this stage, modeling is done to arrange each of the module modules in rapitminer so that it can enter, process and produce the desired information. For modeling sentiment analysis using the Naïve Bayes algorithm, Decision trees, and K-Nearst Neighbor and produce an evaluation to check the performance of each algorithm with a standard Confusion Matrix measurement.

After successfully designing a basic model for measuring sentiment analysis performance, separately the author included a dataset that has been divided 8: 2 for training data and test data, input training data on retrive module 1, and test data on retrive module (2). Then select the Naïve Bayes algorithm module to test the performance of the Naïve Bayes algorithm as shown in Figure 4.

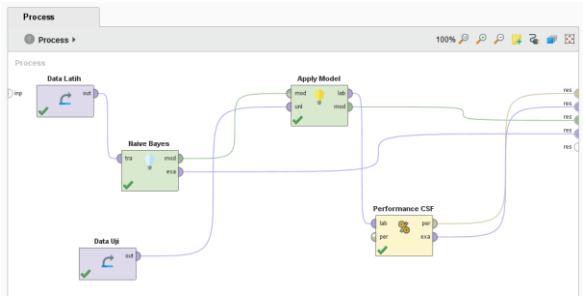


Fig.4. Naive Bayes Algorithm Framework Model

For the K-Nearest Neighbor classification model. A slight modification was made to the algorithm module as in Figure 5

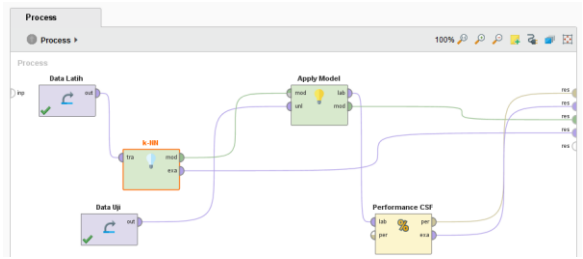


Fig.5. K-Nearest Neighbor Algorithm Framework Model

In the KNN model, there are different property settings in the parameter menu, for parameters for the KNN algorithm module, it can adjust the K value in the KNN algorithm, the K value will be tested from a range of 1-8 iterations or until the accuracy value produces the same value. as in figure 6.

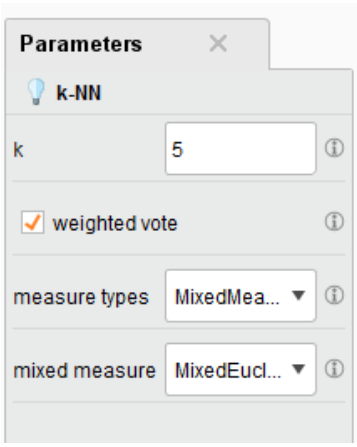


Fig.6. Menu Parameters in KNN Algorithm.

For the Decision Tree classification model, it can be made as shown in Figure 7.

Amrin Fakhruddin Jauhari and Dr. Anggraeni Ridwan, "Performance Measurement of Confusion Matrix Accuracy in Sentiment Analysis with Decision Trees, Naïve Bayes, K-Nearest Neighbor methods Using Rapidminer," *International Research Journal of Advanced Engineering and Science*, Volume 8, Issue 4, pp. 123-127, 2023.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

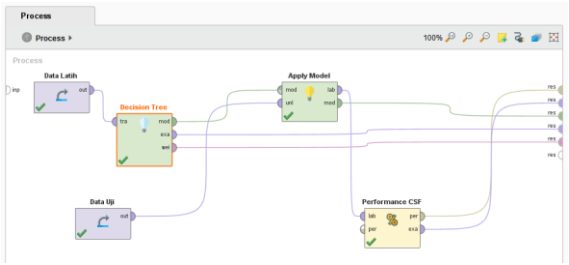


Fig.7. Decision Tree Algorithm Framework Model

In the sentiment analysis model using the Decission tree module, no changes are made to the parameters model property, because it does not affect any value in the test results. The value of the test results is not affected because the parameter values used in the data set only contain 2 attributes and do not contain decision or comparison parameters in the dataset, Therefore, the decision tree computation does not generate any decision trees; rather, it just forecasts measurement accuracy.

C. Testing and Evaluation.

During the testing and assessment phases of this project, K-Nearst Neighbor, Decision trees, and Naïve Bayes algorithm modules were used with rapidminer tools. using the retrive module to use datashets that have been divided and imported into the rapitminer application, using the Apply model module to apply the results of training data processed with algoirtma and test data to be compared, then measured using the Performance classification module to produce the required evaluation measurements.

For evaluation measurements, using a standard measurement that is Testing the Confusion Matrix. Confusion Matrix measurement method is a method commonly used to calculate the accuracy of data mining. Information regarding classes that a classification system properly predicts is contained in the confusion matrix. In this method there are three parameters to be calculated, namely Recall, Precision and Accuracy.

At this stage, running the Framework model that was previously designed using the Naïve Bayes method will produce output like Figure 8.

Table View Plot View

accuracy: 51.46%

	true Netral	true Negative	true Positve	class precision
pred. Netral	1067	528	490	51.18%
pred. Negative	2	9	0	81.82%
pred. Positve	5	3	14	63.64%
class recall	99.35%	1.67%	2.78%	

Fig.8. The Results of Measurements Using The Naive Bayes Method

from figure 8. It can be seen the value in the Naïve Bayes table, has an accuracy value or closeness between the predicted value and the actual value of 51.46%, with a neutral prediction accuracy rate of 51.18%, negative prediction of 81.82%, and positive prediction of 63.64%, while the ability of the system to rediscover very high information or recall for neutral sentiment of 99.35%, while negative 1.67%, and positive 2.78%.

At this stage 2, running the Framework model that was previously designed using the KNN method will produce output like Figure 9.

Table View Plot View

accuracy: 26.11%

	true Netral	true Negative	true Positve	class precision
pred. Netral	5	3	2	50.00%
pred. Negative	1064	534	488	25.60%
pred. Positve	5	3	14	63.64%
class recall	0.47%	98.89%	2.78%	

Fig.9. The Results of Measurements Using The KNN Method With K=1

Using the weight value K = 1, K-Nearst Neighbor model, with 8,472 samples with 1 dimension representing class, Negative, Positive and Neutral. In figure 9, the confidence value is seen, because the range of confidence is only close to 1, while for other sentiments it is 0. In order to obtain an accuracy level that is constant, the K value will be used in the subsequent iteration to modify the difference in confidence value.

Testing continues to be done by increasing the iteration value of K to get the value of accuracy, precision, and recall that does not change (fixed value), the test is summarized in Figure 10 and 11.

Accuracy : 21.11%	Classification Error : 73.89%		Kappa : 0.007	
K-Nearst Neighbour K=1	True Netral	True Negative	True Positve	Class Precision
pred. Netral	5	3	2	50.00%
pred. Negative	1064	534	488	25.60%
pred. Positve	5	3	14	63.64%
class recall	0.47%	98.89%	2.78%	

Accuracy : 24.36%	Classification Error : 75.64%		Kappa : 0.006	
K-Nearst Neighbour K=2	True Netral	True Negative	True Positve	Class Precision
pred. Netral	5	3	2	50.00%
pred. Negative	2	9	0	81.82%
pred. Positve	1067	528	502	23.94%
class recall	0.47%	1.67%	99.60%	

Accuracy : 51.46%	Classification Error : 48.54%		Kappa : 0.023	
K-Nearst Neighbour K=3	True Netral	True Negative	True Positve	Class Precision
pred. Netral	1067	528	490	51.18%
pred. Negative	2	9	0	81.82%
pred. Positve	5	3	14	63.64%
class recall	99.35%	1.67%	2.78%	

Accuracy : 51.32%	Classification Error : 48.68%		Kappa : 0.023	
K-Nearst Neighbour K=4	True Netral	True Negative	True Positve	Class Precision
pred. Netral	1062	525	488	51.18%
pred. Negative	2	9	0	81.82%
pred. Positve	10	6	16	50.00%
class recall	98.88%	1.67%	3.17%	

Fig.10. The iteration measurement table values K = 1 to 4.

Accuracy : 51.37%	Classification Error : 48.63%		Kappa : 0.024	
K-Nearst Neighbour K=5	True Netral	True Negative	True Positve	Class Precision
pred. Netral	1062	525	488	51.18%
pred. Negative	7	12	2	57.14%
pred. Positve	5	3	14	63.64%
class recall	98.88%	2.22%	2.78%	

Accuracy : 65.12%	Classification Error : 34.88%		Kappa : 0.468	
K-Nearst Neighbour K=6	True Netral	True Negative	True Positve	Class Precision
pred. Netral	5	3	2	50.00%
pred. Negative	2	9	0	81.82%
pred. Positve	5	3	14	63.64%
class recall	41.67%	60.00%	87.50%	

Accuracy : 60.47%	Classification Error : 39.53%		Kappa : 0.382	
K-Nearst Neighbour K=7	True Netral	True Negative	True Positve	Class Precision
pred. Netral	0	0	0	0.00%
pred. Negative	7	12	2	57.14%
pred. Positve	5	3	14	63.64%
class recall	0.00%	80.00%	87.50%	

Accuracy : 60.47%	Classification Error : 39.53%		Kappa : 0.382	
K-Nearst Neighbour K=8	True Netral	True Negative	True Positve	Class Precision
pred. Netral	0	0	0	0.00%
pred. Negative	7	12	2	57.14%
pred. Positve	5	3	14	63.64%
class recall	0.00%	80.00%	87.50%	

Fig.11. The iteration measurement table values K = 5 to 8.

Amrin Fakhruddin Jauhari and Dr. Anggraeni Ridwan, "Performance Measurement of Confusion Matrix Accuracy in Sentiment Analysis with Decision Trees, Naïve Bayes, K-Nearest Neighbor methods Using Rapidminer," *International Research Journal of Advanced Engineering and Science*, Volume 8, Issue 4, pp. 123-127, 2023.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

From the results of iteration in figures 10 and 11, it can be seen the best K value to find the greatest accuracy using the value of K = 6. Which produces the largest accuracy value of 65.12%, provided the machine can only predict a little sentiment, because the value of the matching confidence is very little, thus ignoring the value of the unsuitable sentiment confidence.

accuracy: 65.12%

	true Netral	true Negative	true Positve	class precision
pred. Netral	5	3	2	50.00%
pred. Negative	2	9	0	81.82%
pred. Positve	5	3	14	63.64%
class recall	41.67%	60.00%	87.50%	

Fig.12. The Results of Measurements Using The KNN Method With K=6

Processing the third test data and training data using the Decision trees method Prediction sentiment uses the Decision Trees method, which produces (2,118) neutral predictions, (0) positive predictions, and (0) negative predictions. From processing the test data and training data using the Decision Tree method produces the accuracy contained in the figure 13.

Table View Plot View

accuracy: 50.71%

	true Netral	true Negative	true Positve	class precision
pred. Netral	1074	540	504	50.71%
pred. Negative	0	0	0	0.00%
pred. Positve	0	0	0	0.00%
class recall	100.00%	0.00%	0.00%	

Fig.13. The Results of Measurements Using The Decision Tree Method

In the third trial the classification method for classifying and predicting sentiment analysis results in a comparison of the data presented in the form of a table figure 14.

Accuracy : 51.46%	Classification Error : 48.54%		Kappa : 0.023	
Naïve Bayes	true Netral	true Negative	true Positve	class precision
pred. Netral	1067	528	490	51.18%
pred. Negative	2	9	0	81.82%
pred. Positve	5	3	14	63.64%
class recall	99.35%	1.67%	2.78%	

Accuracy : 21.11%	Classification Error : 73.89%		Kappa : 0.007	
K-Nearst Neighbour	true Netral	true Negative	true Positve	class precision
pred. Netral	5	3	2	50.00%
pred. Negative	1064	534	488	25.60%
pred. Positve	5	3	14	63.64%
class recall	0.47%	98.89%	2.78%	

Accuracy : 50.71%	Classification Error : 49.29%		Kappa : 0.000	
Decision trees	true Netral	true Negative	true Positve	class precision
pred. Netral	1074	540	504	50.71%
pred. Negative	0	0	0	0.00%
pred. Positve	0	0	0	0.00%
class recall	100.00%	0.00%	0.00%	

Fig.14. Comparison of Accuracy, Class Recall, and Class precision.

IV. CONCLUSION

Based on the results of research conducted by testing three times for the Decision Trees method, and K-Nearst Neighbor and Naïve Bayes, on the two attributes, sentiment attributes which have a polynominal type, and have rules as labels. And the tweet attribute that has a polynominal data type. By using Indonesian-general-sentiment-analysis-data. As test data and training data in the performance analysis sentiment analysis using rapidminer has been successfully carried out.

The use of Indonesian-general-sentiment-analysisis-dataseht. In testing the measurement of accuracy in sentiment analysis using the rapitminer tool, it can be used to measure the performance results of sentiment analysis for the Decision Trees method, and K-Nearst Neighbor and Naïve Bayes. Because all three methods can be run with polynominal data types that can measure three positive, negative, and neutral label values. Whereas for other methods such as Neural Network (NN), and Support Vector Machine (SVM), testing cannot be done because for these methods the sentiment attribute must be of binominal data type for which only has two values, such as positive and negative.

From the results of experimental measurements by the Decision Trees, K-Nearest Neighbor, and Naïve Bayes methods. Generate measurements of performance accuracy on methods :

A. Naïve Bayes Method

With an accuracy of 51.46%, the closest match between the expected and real values, and recall rates of 99.35% for true neutral, 1.67% for true negative, and 2.78% for true positive, Naïve Bayes has the highest accuracy rate of any algorithm. While for the measurement of the level of accuracy between the information requested by the user with the answer given by the system (Precision) in the form, neutral prediction 51.18%, negative prediction 81.82%, and positive prediction 63.64%. with the number of wrong prediction percentage (classification Error) 48.54% and the calculation of the correct prediction because it takes into account the correct prediction, which happened by chance (Kappa) of 0.023.

B. K-Nearest Neighbor Method

K-Nearest Neighbor with a weight of K = 1 has a closeness level between the predicted value and the actual value (Accuracy) to the lowest of 21.11%, with the success rate of the system in finding back an information (Recall) True Neutral 0.47%, true negative 98.89%, and true positive 2.78%. while for the measurement of the level of accuracy between the information requested by the user with the answer given by the system (Precision) in the form, neutral prediction 50.00%, negative prediction 25.60%, and positive prediction 63.64%. With the number of wrong prediction percentage (classification Error) 73.89%. Although the computation of the accurate forecast considers the accurate forecast, which occurred by chance (Kappa) of 0.07.

While in the 6th iteration KNN has a closeness level between the predicted value and the actual value (Accuracy) which is quite high at 65.12%, with (Recall) True Neutral 41.67%, true negative 60.00%, and true positive 87.50%.

Amrin Fakhruddin Jauhari and Dr. Anggraeni Ridwan, "Performance Measurement of Confusion Matrix Accuracy in Sentiment Analysis with Decision Trees, Naïve Bayes, K-Nearest Neighbor methods Using Rapidminer," *International Research Journal of Advanced Engineering and Science*, Volume 8, Issue 4, pp. 123-127, 2023.

Regarding the measurement (precision), it is 50.00% for neutral predictions, 81.82% for negative predictions, and 63.64% for positive predictions. With the number of wrong prediction percentage (classification Error) 34.88%. In contrast, the Kappa prediction calculation is 0.07.

C. Decision Tree Method

Deccision Tree has a level of closeness between the predicted value and the actual value (Accuracy) that is 50.71%, with the success rate of the system in finding back an information (Recall) True Neutral 100.00%, true negative 0.00%, and true positive 0.00%. while for the measurement of the level of accuracy between the information requested by the user with the answer given by the system (Precision) in the form, neutral prediction 50.71%, negative prediction 0.00%, and positive prediction 0.00%. With the number of wrong prediction percentage (classification Error) 49.29%. While the calculation of the correct prediction because it takes into account the correct prediction, which happened by accident (Kappa) of 0.00

ACKNOWLEDGMENT

I am very grateful for the time, direction, motivation, and guidance given by Dr. Anggraeni Ridwan, SKom., MMSI. for helps the smoothness of the author in completing this research.

I also thank all my colleagues in the information systems laboratory, for providing the most informed information in this study.

REFERENCES

[1] Ruchi Rautela., Priyanka Dilip Huilgol dan Sunit Pravin Kajarekar. “Text mining : A comprehensive Survey”. International Journal of recent scientific research. Vol.9 issue, 5(G), PP.2659-26962. 2018.

[2] Ridi Ferdiana, Fahim Jatmiko, Desi Dwi Purwanti, Artmita Sekar Tri Ayu, Wiliam Fajar Dicka. “Datashet Indonesia Untuk Analisis Sentimen”. JNTETI, Vol.8, No. 4. 2019.

[3] Nanang Ruhyana. “Analisis Sentimen Terhadap Penerapan Sistem Plat Nomor Ganjil/Genap pada twitter dengan metode klasifikasi Naïve bayes”. Jurnal IKRA-ITH Informatika vol.3, No. 1. 2019.

[4] Oman Somantri dan Dairoh. “Analisis sentiment penelitian tempat tujuan wisata kota tegal berbasis text mining”. JEPIN. Vol.5, No.2. 2019.

[5] Akhmad Deviyanto, M. Didik R. Wahyudi. “Penerapan analisis sentimen pada pengguna twitter menggunakan metode K-Nearest Neighbor”. JISKA. Vol.3, No.1. PP.1-13. 2018.

[6] Oluwakemi Christiana Abikoye, Samuel Oladeji Omokanye, Taye Oladele Aro. “Text Classification Using Data Mining Techniques: A Review”. Computing and information system journal. Vol.22, No. 22. 2018.