JURNAL SYNTAX ADMIRATION

Volume 4, No. 7, July 2023
p-ISSN 2722-7782 | e-ISSN 2722-5356
DOI: https://doi.org/ 10.46799/jsa.v4i7.654

CLASSIFICATION OF CONGESTION IN JAKARTA USING KNN, NAÏVE BAYES AND DECISION TREE METHODS

Sri Rahayu, Bayu Rimbi Asmoro, Ery Rinaldi
Faculty of Information Technology, Budi Luhur University
2111600322@student.budiluhur.ac.id, 2111600124@student.budiluhur.ac.id, 2111600223@student.budiluhur.ac.id

Abstract

Congestion has now become a problem that occurs in almost all big cities in Indonesia. The problem of traffic jams generally occurs in areas with high intensity of activity and land use. Given the increasing level of congestion that is happening, the capital city of DKI Jakarta is one of the most densely populated cities with high population activity. Population activities are also offset by the use of transportation. Both by public and private vehicles. Traffic jams are one of the problems that are still unsolved. West Palmerah Street is one of the roads with quite a lot of traffic jams. To prove it, he did some simple research. The method used is descriptive method, where the research begins with collecting the data needed at this time through several surveys. And the calculation is done by looking for the degree of saturation (DS) and vehicle speed at three checkpoints, the DS at the pre-market checkpoint is 0.89, the DS at the market checkpoint is 1.05, and the DS at the market checkpoint is 0.89. Then the movement speed was also obtained at the pre-market observation point of 32.05 km/hour, at the market review point of 27.5975 km/hour, and at the post-market observation point of 33.35 km/hour. The results prove that there is indeed a traffic delay in front of the market. This figure is due to the large number of angkots that stop and the narrowing of the traffic lane in front of the market due to the presence of street vendors and motorbikes stopping on the sidewalks with buying and selling activities on the sidewalks. Therefore, it is necessary to apply the best operational solutions to improve traffic flow on these roads.

Keywords : Congestion in Jakarta, Classification, K-nearest neighbors, Naïve Bayes, Decision Tree.

Abstract

Congestion has now become a problem that occurs in almost all big cities in Indonesia. The problem of traffic jams generally occurs in areas with high intensity of activity and land use. Given the increasing level of congestion that is happening, the capital city of DKI Jakarta is one of the most densely populated cities with high population activity. Population activities are also offset by the use of transportation. Both by public and private vehicles. Traffic jams are one of the problems that are still unsolved. West Palmerah Street is one of the roads with quite a lot of traffic jams. To prove it, he did some simple research. The method used is a descriptive method, where the research begins by collecting the data needed at this time through several surveys. And the calculation is done by looking for the degree of saturation (DS) and vehicle speed at three checkpoints, the DS at the pre-market checkpoint is 0.89, the DS at the market checkpoint is 1.05, and the DS at the market checkpoint is 0.89. Then the movement speed was also obtained at the pre-market observation point of 32.05 km/hour, at the market review point of 27.5975 km/hour, and at the post-market observation point of 33.35 km/hour. The results prove that there is indeed a traffic delay in front of the market. This figure is due to the large number of angkots that stop and the narrowing of the traffic lane in front of the market due to the presence of street vendors and motorbikes stopping on the sidewalks with buying and selling activities on the sidewalks. Therefore, it is necessary to apply the best operational solutions to improve traffic flow on these roads.

Keywords: *Congestion in Jakarta, Classification, K-nearest neighbors, Naïve Bayes, Decision Tree.*

INTRODUCTION

Data mining is a method of determining certain patterns from a large amount of data. Data mining has many techniques, one of which is a classification technique. Classification is a data learning technique for generating value predictions from a series of attributes (Wahyuningsih & Utari, 2018) . Classification is widely used to predict classes on certain labels by classifying data (building models) based on training sets and values (class labels) when classifying certain attributes. Classification is divided into five categories based on differences in mathematical concepts, namely statistical based, distance based, decision tree based, neural network based, and rule based. Classification has many algorithms, but in this study using decision tree, KNN and Naïve Bayes algorithms (Sartika & Sensuse, 2017) . Of the three algorithms, the decision tree is one of the most commonly used methods, especially in data classification.

In case studies of sentiment analysis of BPJS service users using the KNN, Naïve Bayes and Decision Tree methods it proves that the Decision Tree method has a high level of accuracy in data classification (Puspita & Widodo, 2021) . In a comparative case study of the K-Nearest Neighbor Data Mining Method with Naïve Bayes for the classification of Congestion in Jakarta, the KNN method is proven to have high accuracy compared to Naïve Bayes (Rahman et al., 2018) . Compared to the Naïve Bayes method, this method rarely has a high level of accuracy, so this study will compare the three algorithms based on their level of accuracy, which method is the best for classification.

Based on the existing problems, specifically to compare the three decision tree methods, KNN and Naïve Bayes, a study was carried out with the title " Classification of Congestion in Jakarta Using the KNN, Naïve Bayes and Decision Tree Methods " using the rapid method. Mining software to find the highest accuracy value of the three methods that will be implemented in data classification is a comparative analysis of traffic jam accuracy using KNN, naive Bayes and decision tree classification data. The purpose of this study is to compare the three best methods used in the classification of congestion with maximum accuracy results.

A study that discusses the Naïve Bayes, KNN and Decision Tree methods for sentiment analysis of traffic jams with the problem of traffic conditions in the city of Jakarta which are so dense and congestion is increasing, that residents who want to work need more comfortable transportation (Riadi & Kom , 2017) . This research uses social media Twitter to get random data for up to 127 dates. Using the Naive Bayes Classifier, KNN and Decision Tree methods with several stages, namely emoticon conversion, cleaning, case stacking, tokenization and stemming (Romadloni et al., 2019) . The results obtained with the decision tree method have the highest accuracy compared to KNN and Naïve Bayes, where the decision tree has 100% accuracy, 100% accuracy, 100% sensitivity and 100% specificity. The KNN method has 80% accuracy, 100% accuracy, 50% sensitivity, 100% specificity, and the Naive Bayes method has 80% accuracy, 66.67% accuracy, 100% sensitivity and 66.67% specificity.

Research on the classification of traffic jams uses a comparison of the K-Nearest Neighbor and Naïve Bayes data mining methods. Monitoring and processing of the surrounding environment, including water resources, is necessary to create traffic jams that comply with congestion standards (Rahman et al., 2018) . The accuracy results are 82.42% for K-Nearest Neighbor and Naïve Bayes of 70.32%, it can be concluded that KNearest Neighbor is the best method for determining congestion.

In the research on sentiment analysis of BPJS users using the KNN, Decision Tree and Naïve Bayes methods, discussing people who use BPJS services, which often raises pros and cons, for this reason, data mining sentiment analysis research was carried out on BPJS.

Twitter users with 1,000 entries are filtered down to 903 due to duplicate data. Implement the KNN, Decision Tree and Naïve Bayes methods to compare the level of accuracy of the three methods used (Puspita & Widodo, 2021) . This study used rapid miner software version 9.9, where the results obtained were that the KNN method had an accuracy rate of 95.58%, a decision tree was 96.13% and the Naive Bayes method was 89.14%, so it can be concluded that the best method for decision making decision tree is used.

Data Mining is the process of obtaining information to obtain new information (Harahap, 2019) . The research conducted this time uses data mining techniques that implement the K-nearest Neighbors, Naïve Bayes and Decision Tree methods to compare the results of the maximum accuracy of the three methods used. Data mining is a data source and use operation that is used to find relationships or patterns from large data sets to obtain new information (Cahyanti et al., 2020) .

The K-Nearest Neighbor algorithm is a classification method for a dataset based on previously classified training data (Siregar et al., 2019) . The KNN classification algorithm is a method for classifying objects based on training data that has the shortest distance (Romadloni et al., 2019) . The working principle of the KNN algorithm is to determine and find the shortest distance to the nearest neighbor value in the training data with the data to be tested. The best k value for this algorithm depends on the data value, where usually a high k value reduces the effect of errors or noise on the classification process, but creates suboptimal boundaries between classifications (Sukmana et al., 2020) . This research will carry out a computational process to obtain accurate data results using the KNN method. The formula for finding the distance using the Euclidean formula:

$$\mathbf{d_i} = \sqrt{\sum_{i=1}^{p} (\mathbf{x_{2i}} - \mathbf{x_{1i}})^2}$$

where x1 is sample data; d is distance; x2 is test data; p is the data dimension, i is the data variable.

Naive Bayes Classifier is a data mining method for data classification. The operation of the Naive Bayes Classifier method uses probabilistic calculations. Naive Bayes is one of the algorithms included in the classification technique (Zulfauzi & Alamsyah, 2020) . The basic concept of Naive Bayes uses the Bayes theorem, which is a theorem used in statistics that is used to calculate probabilities. The Naive Bayes Classifier calculates the probability of one class from each group of attributes and determines the most optimal class (Lestari et al., 2021) . The Naive Bayes classifier function calculates and looks for the highest probability value to classify test data into the correct category. A simple probability

prediction technique based on the application of the Bayes theorem or Bayes rule is a technique implemented in the Naïve Bayes algorithm. Naive Bayes Formula:

$$P(H|X) = \frac{P(H)P(X|H)}{P(X)}$$

where X is data with unknown class; H is the hypothesis that data X is class specific; P(H|X) is the probability of the hypothesis H under condition X ; P(H) is the probability of the hypothesis H (prior probability); P(X|H) is the probability of X based on the conditions in hypothesis H; P(X) is the probability of X

The data classification process can use several methods, one of which is a decision tree. The decision tree is one of the commonly used algorithms for decision making (Pamuji & Ramadhan, 2021) . The decision tree is an algorithm that is good for classification or prediction (Muningsih, 2022) . The Decision Tree Model is in the form of a tree which consists of several parts, namely the root node, internal node, and terminal node. The root node from searching query data and the internal node that reaches the end node is the classification process in this decision tree method. The concept of entropy to be used to determine which attribute in the decision tree to split, the higher the sample entropy, the less pure the sample is. The formula for calculating sample entropy is:

$$Entropy(S) = - P_1 \log_2 P1 - P2 \log_2 P_2$$

where p1, p2, p3 ….. , pn respectively represent class 1, class 2,….. class n proportions in the output.

METHODOLOGY

In this study several stages were used which are presented in the form of Figure 1 Research Stages.

Gambar 1. Tahapan Penelitian

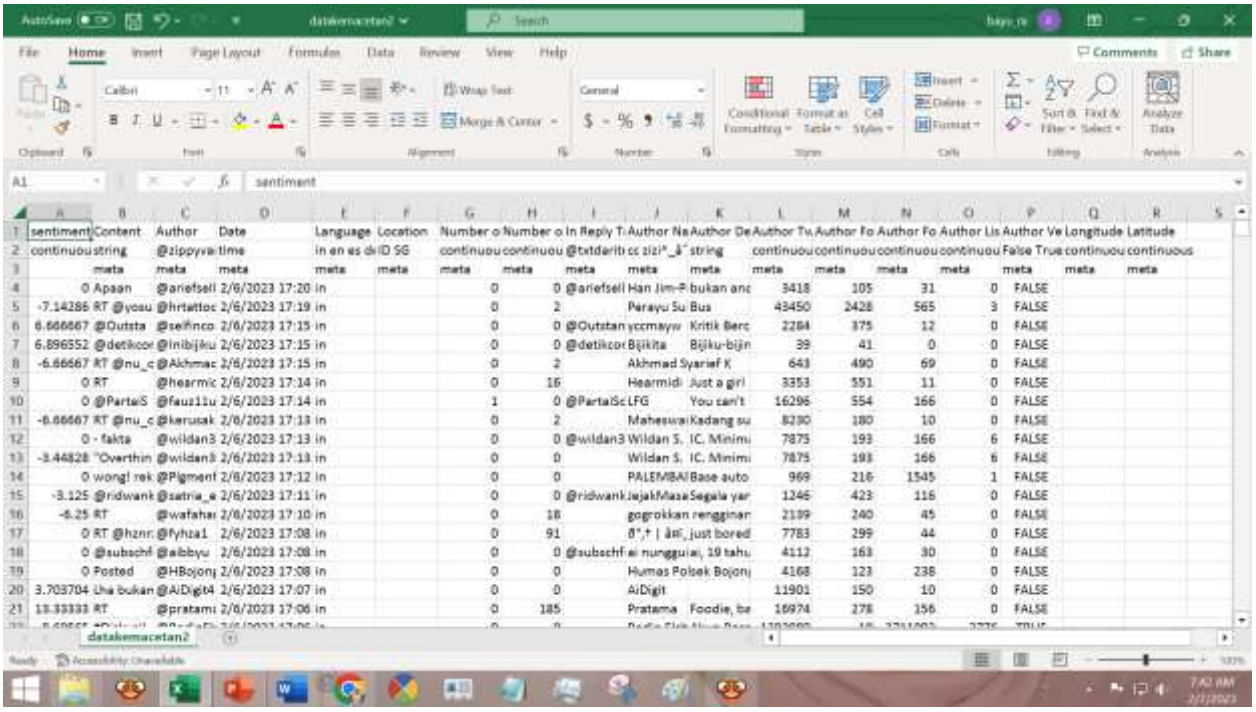
The first stage of this research begins with mining data on Twitter using Orange Software and of course the Twitter website. The second stage is the study of literature as a collection of information relating to the preparation of the final project. Collecting information to support this research in the form of journals, books, references and other reliable sources. Not spared from discussions and consultations, as well as research methods during the preparation of this diploma thesis, discussions and consultations with supervisors and various experts in this field. The data processing process at Rapid Miner includes several steps, starting from data sets, pre-processing, data separation into training data and data testing, model fitting/classification, prediction/model application, and the resulting process. The data processing carried out will produce a result or result that will be discussed and produce a conclusion in the research process carried out.

RESULTS AND DISCUSSION

Datasets

In this study, the overloaded csv data type dataset was used for the classification process as well as to compare the results of the accuracy of the three methods used, namely Naive Bayes, Decision Tree and KNN. The results of the data obtained in Table 1.

Table 1
Traffic jam dataset on Twitter



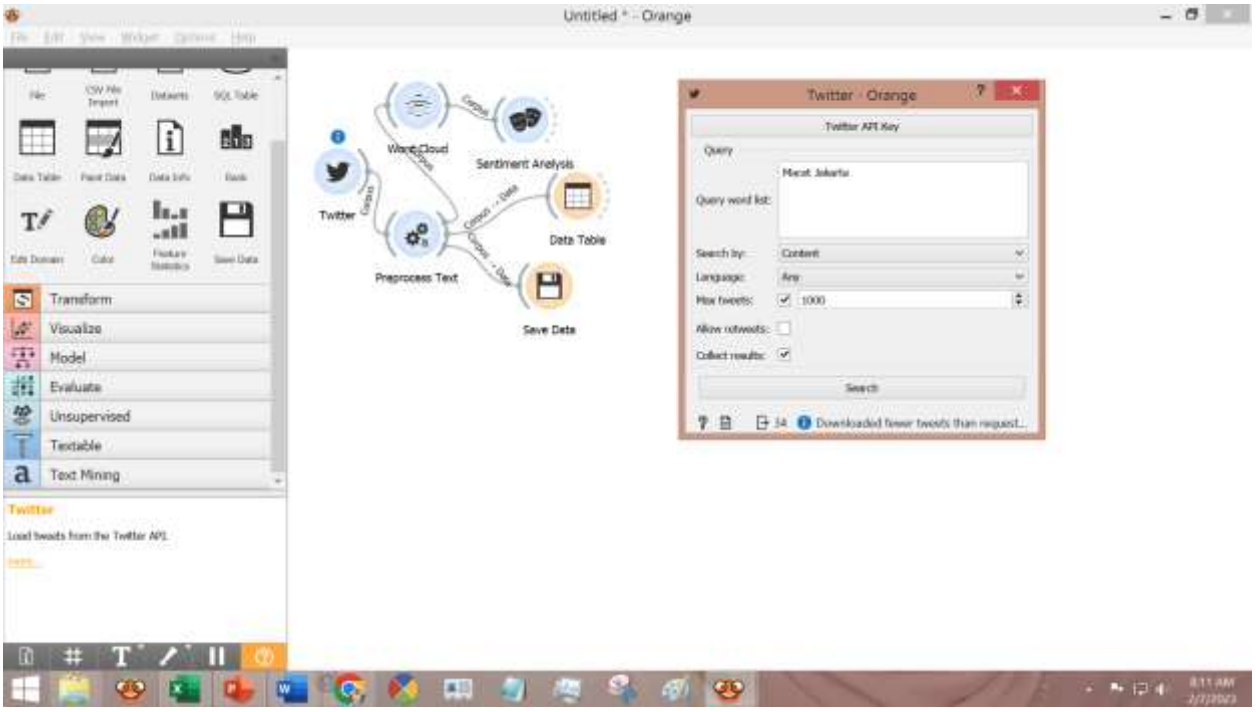
sentiment	Content	Author	Date	Language	Location	Number o	Number o	In Reply T	Author Na	Author De	Author Tw	Author Fo	Author Fo	Author Lis	Author Ve	Longitude	Latitude
continuou	string	@zippyva	time	in en es de	ID SG	continuou	continuou	@txtdarlb	cc zizi^_â´	string	continuou	continuou	continuou	continuou	False True	continuou	continuous
meta	meta	meta	meta	meta	meta	meta	meta	meta	meta	meta	meta	meta	meta	meta	meta	meta	meta
0	Apaan	@ariefsell	2/6/2023 17:20	in		0	0	@ariefsell	Han Jim-P	bukan anc	3418	105	31	0	FALSE		
-7.14286	RT @yosu	@hrtattoc	2/6/2023 17:19	in		0	2		Perayu Su	Bus	43450	2428	565	3	FALSE		
6.666667	@Outsta	@selfinco	2/6/2023 17:15	in		0	0	@Outstan	yccmayw	Kritik Berc	2284	375	12	0	FALSE		
6.896552	@detikcor	@inibijiku	2/6/2023 17:15	in		0	0	@detikcor	Bijikita	Bijiku-bijin	39	41	0	0	FALSE		
-6.66667	RT @nu_c	@Akhmac	2/6/2023 17:15	in		0	2		Akhmad	Syarief K	643	490	69	0	FALSE		
0	RT	@hearmic	2/6/2023 17:14	in		0	16		Hearmidi	Just a girl	3353	551	11	0	FALSE		
0	@PartaiS	@fauz11u	2/6/2023 17:14	in		1	0	@PartaiSc	LFG	You can't	16296	554	166	0	FALSE		
-6.66667	RT @nu_c	@kerusak	2/6/2023 17:13	in		0	2		Maheswai	Kadang su	8230	180	10	0	FALSE		
0	- fakta	@wildan3	2/6/2023 17:13	in		0	0	@wildan3	Wildan S.	IC. Minimi	7875	193	166	6	FALSE		
-3.44828	"Overthin	@wildan3	2/6/2023 17:13	in		0	0		Wildan S.	IC. Minimi	7875	193	166	6	FALSE		
0	wong! rek	@Pigment	2/6/2023 17:12	in		0	0		PALEMBAI	Base auto	969	216	1545	1	FALSE		
-3.125	@ridwank	@satria_e	2/6/2023 17:11	in		0	0	@ridwank	iejakMasa	Segala yar	1246	423	116	0	FALSE		
-6.25	RT	@wafahai	2/6/2023 17:10	in		0	18		gogrokkan	rengginar	2139	240	45	0	FALSE		
0	RT @hznr	@fyhza1	2/6/2023 17:08	in		0	91		ð",† | âñ,	just bored	7783	299	44	0	FALSE		
0	@subschf	@aibbyu	2/6/2023 17:08	in		0	0	@subschf	ai	nungguiai, 19 tahu	4112	163	30	0	FALSE		
0	Posted	@HBojonj	2/6/2023 17:08	in		0	0		Humas Polsek	Bojonj	4168	123	238	0	FALSE		
3.703704	Lha bukan	@AiDigit4	2/6/2023 17:07	in		0	0		AiDigit		11901	150	10	0	FALSE		
13.33333	RT	@pratami	2/6/2023 17:06	in		0	185		Pratama	Foodie, ba	16974	278	156	0	FALSE		

Pre-processing and Labeling

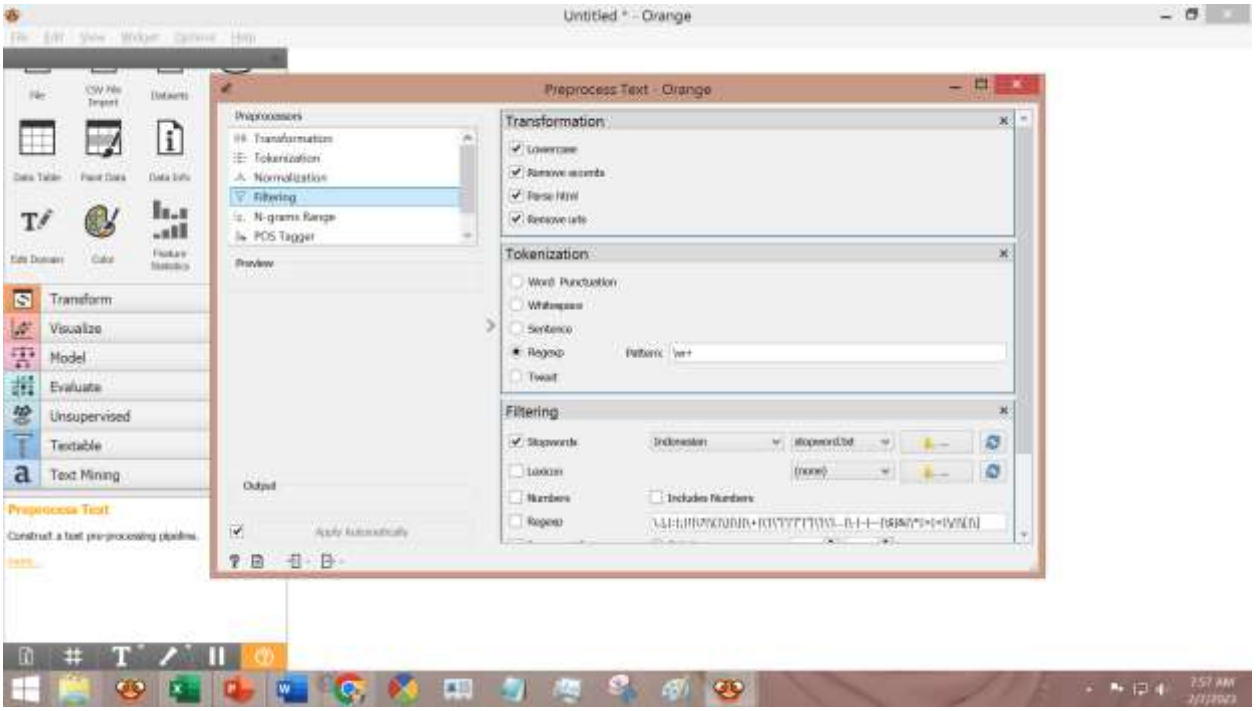
The data obtained in this study need to be processed first. Knowing the nature of the textual data previously collected, the data labeling process was carried out. The attribute identified in this study is pitability, an attribute that indicates whether bottlenecks can be overcome. The labeling process can be done by setting the color on the label to facilitate the research process. Several pre-processing methods are used, namely data validation to obtain good data with proper accuracy, to review the type of data obtained, and to identify data so as to achieve a maximum level of accuracy. Make inconsistent data consistent by replacing all missing operators. Data validation identifies and eliminates data that is not used, as well as inconsistent data and missing data, where raw data becomes data that is ready to be processed and can be analyzed through data cleaning and data

filtering processes. in the data validation process (Teak, 2021) . This study uses data integration and transformation methods to increase the accuracy of the three methods used. The Reduce Data Size and Decretize methods are used to remove duplicate data using the delete duplicate operator. The initial data condition of 1,000 becomes clean data through a process of data validation, data integration and transformation, as well as data size reduction and discretization so that the data can be analyzed to obtain new data information.

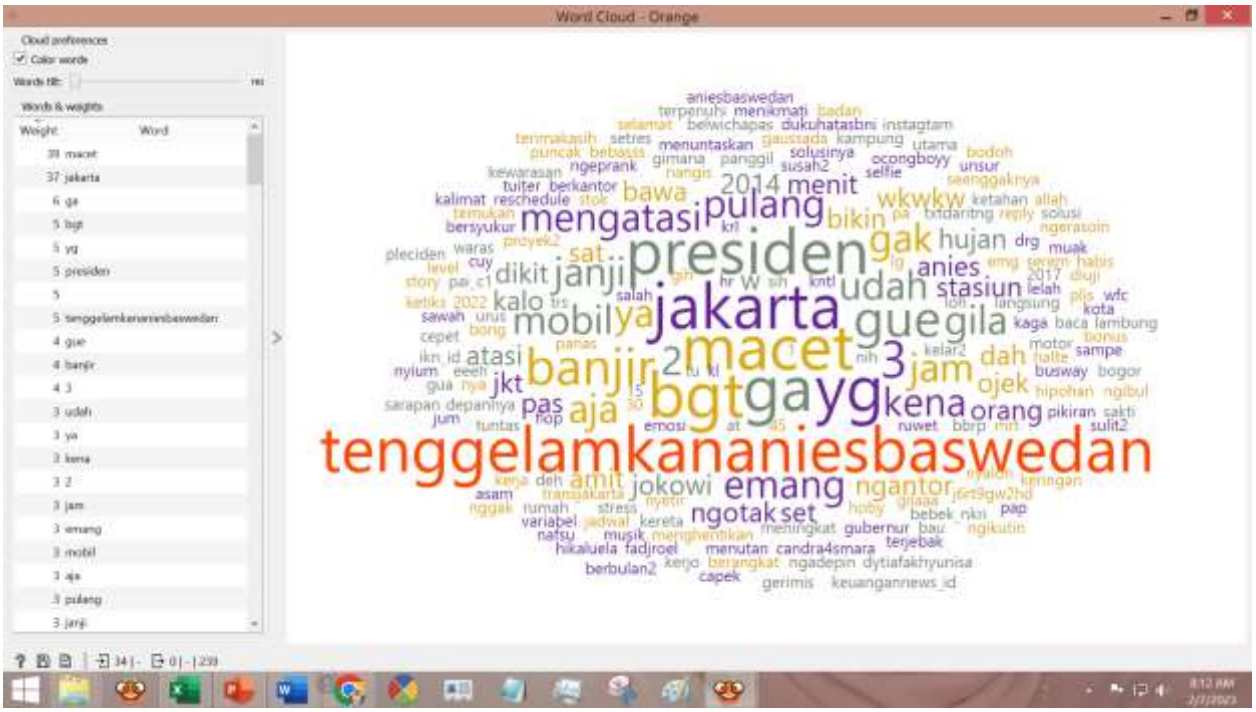
Keyword Determination in Orange: Jakarta Traffic jams



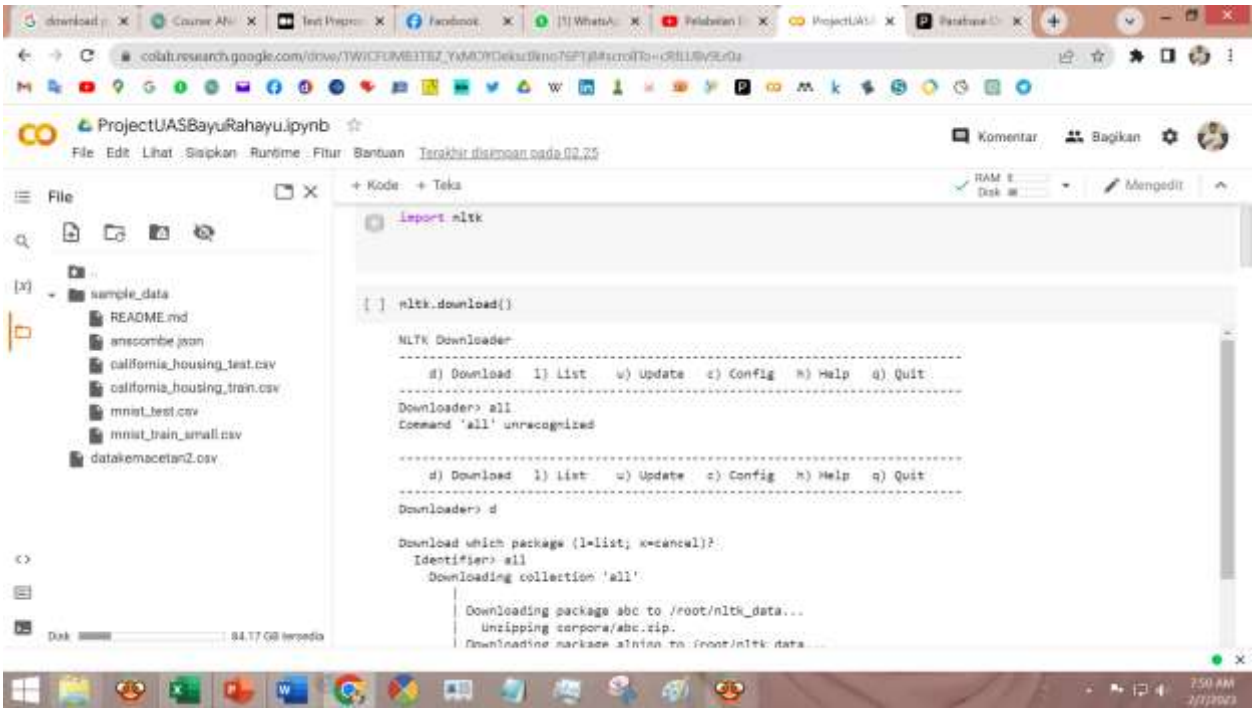
Process Preprocessing Data



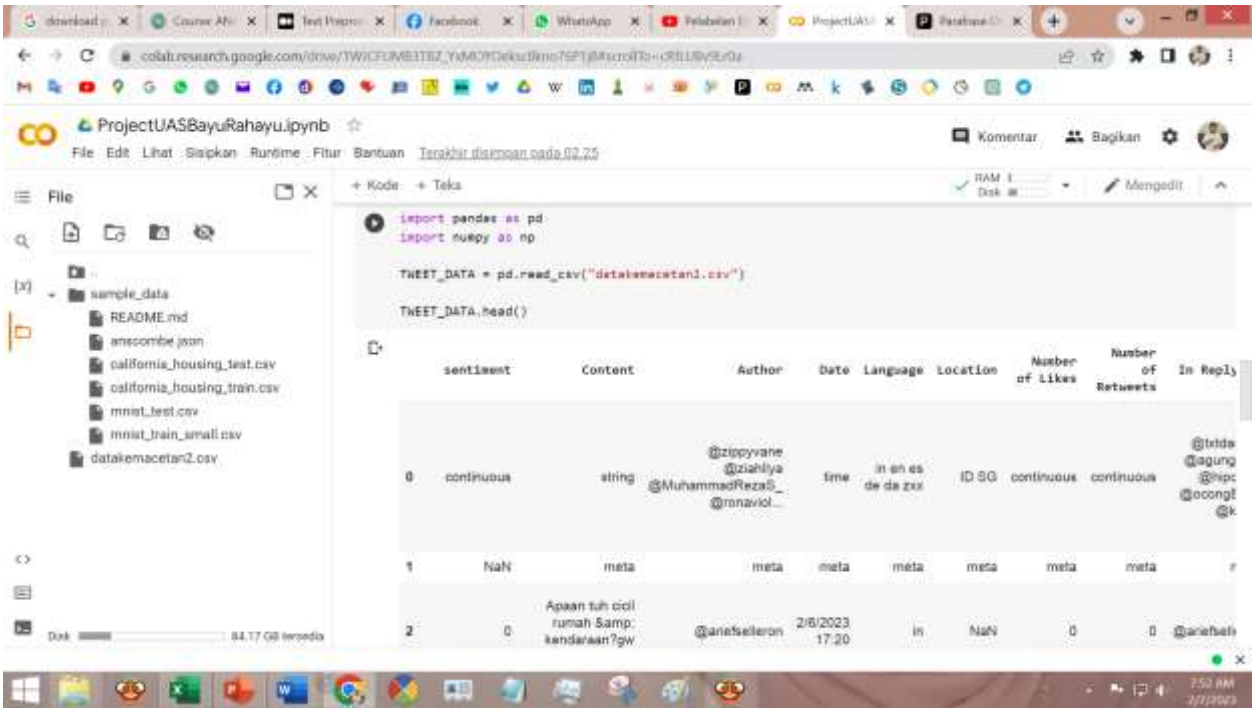
Word Cloud



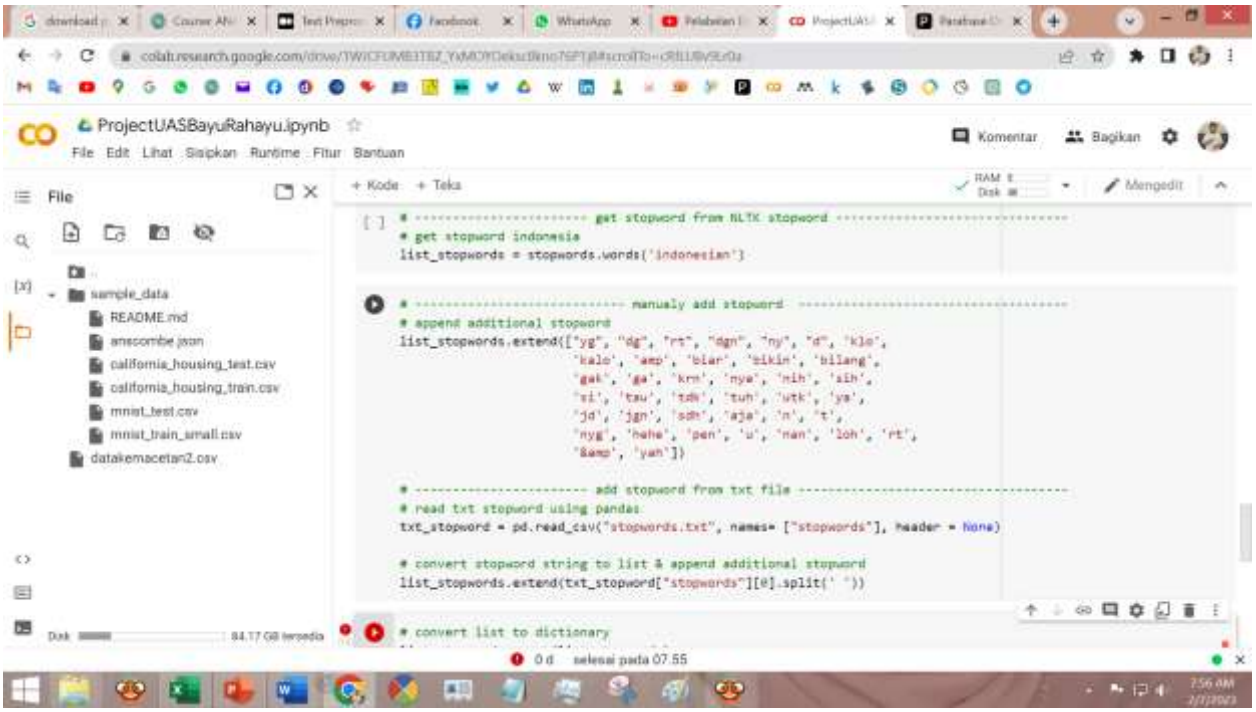
NLTK process in Google Colabs



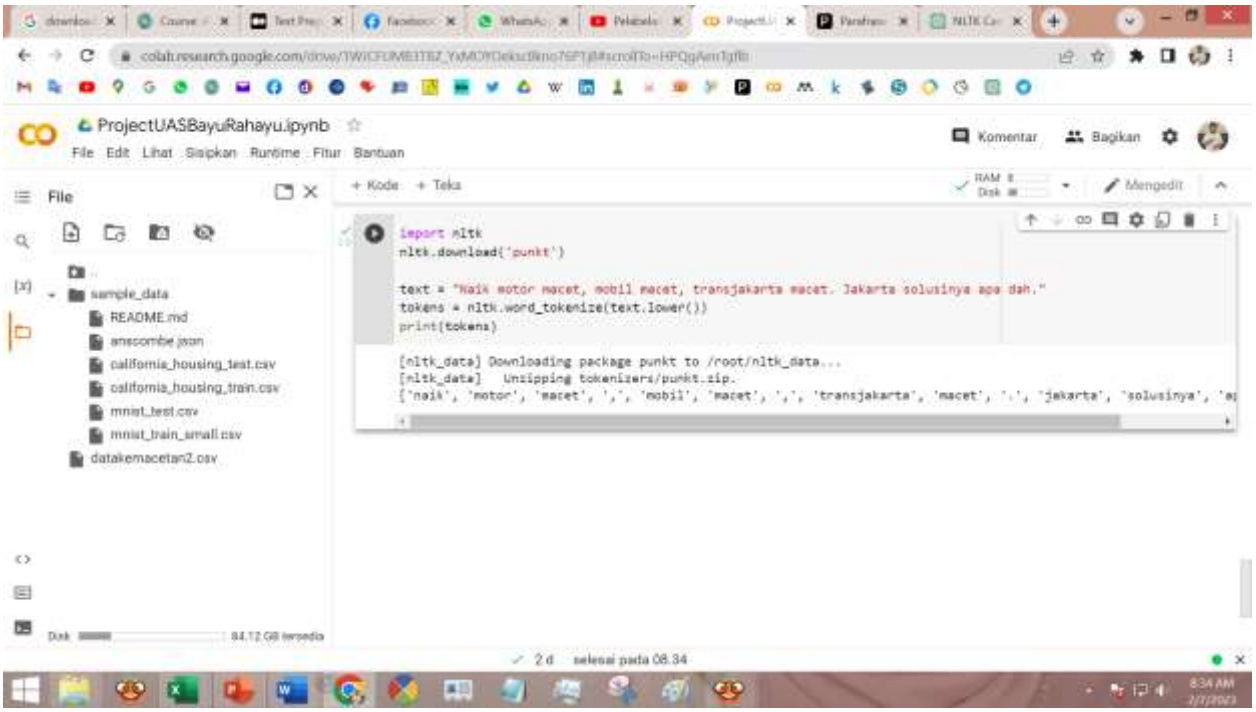
Data Upload Process Using Pandas file *.csv



Stopword process



Case Folding Process



Accuracy Measurement with Confusion Matrix

Confusion Matrix is a classification method based on the results of the classification that has been done, where the accuracy of the classification affects the performance of the classification. The confusion matrix provides comparative information on the classification results carried out by the system (model) with the actual classification results (Fikri et al., 2020) .

The confusion matrix describes the performance of the classification model on a set of test data whose true values are known. Confusion Matrix is used to calculate accuracy.

Confusion Matrix

Kelas	Terklarifikasi Positif	Terklarifikasi Negatif
Positif	TP (*True Positive*)	FN (*False Negative*)
Negatif	FP (*False Positive*)	TN (*True Negative*)

Confusion Matrix performance can be measured using the TP, FP, FN, and TN values. True Positive is positive data that is predicted to be correct. True Negative is negative data that is predicted to be true.

Calculating accuracy using the equation

$$accuracy = \frac{\text{TP+TN}}{\text{TP+TN}+FN+FP}$$

Naive Bayes Algorithm Accuracy Results

Confusion Matrix Naïve. Bayes

***accuracy*: 63.60%**

	true 0	*true* 1	*class precision*
pred. 0	788	430	64.70%
pred. 1	100	138	57.98%
class recall	88.74%	24.30%	

The accuracy result is 63.60%, with class precision for pred. zero (pred. negative) is 64.70% and pred one (pred.positive) is 57.98%. Accuracy results are obtained using equation 4, where the true positive values are 788, true negatives are 138, false negatives are 430, and false positives are 100. Accuracy results can be proven by:

$$accuracy = \frac{788 + 138}{788 + 138 + 430 + 100} = 63.60\%$$

Performance Vektor:

Tabel 4. *Performance Vektor Naïve Bayes*

PerformanceVector:
accuracy: 63.60%
ConfusionMatrix:

True	0	1
0	788	430
1	100	138

Performance Vector itself is a form of description of the table of analysis results obtained in the research conducted. The True Positive value is 788, which is a positive data value which means that water is safe to drink and is predicted to have the correct value. The False Positive value is 100, where the data is negative (water is not drinkable) but is predicted as positive data. The False Negative value is 430, positive data but predicted as negative data. The True Negative value is 138, which is negative data that is predicted to be true.

Decision Tree Algorithm Accuracy Results

Confusion Matrix Decision Tree

***accuracy*: 80.84%**

	true 0	*true* 1	*class precision*
pred. 0	817	208	79.71%
pred. 1	71	360	83.53%
class recall	92.00%	63.38%	

The accuracy result is 80.84%, with class precision for pred. zero (pred. negative) is 79.71% and pred one (pred.positive) is 83.53%. The accuracy results are obtained using equation 4, where the true positive values are 817, true negatives are 360, false negatives are 208, and false positives are 71.

$$accuracy = \frac{817 + 360}{817 + 360 + 208 + 71} = 80.84\%$$

Performance Vector

Tabel 6. Hasil *Performance Vektor Decision Tree*

PerformanceVector:
accuracy: 80.84%
ConfusionMatrix:

True	0	1
0	817	208
1	71	360

Performance Vector itself is a form of description of the table of analysis results obtained in the research conducted. The True Positive value is 817, which is a positive data value which means that water is safe to drink and is predicted to have the correct value. The False Positive value is 71, where the data is negative, but it is predicted as positive data. The False Negative value is 208, positive data but predicted as negative data. The True Negative value is 360, which is negative data that is predicted to be true.

Accuracy results of the K-nearest neighbors algorithm

Confusion Matrix KNN

accuracy: 86.88%

	true 0	*true* 1	*class precision*
pred. 0	836	139	85.74%
pred. 1	52	429	89.19%
class recall	94.14%	75.53%	

Accuracy results were obtained at 86.88%, where the class precision for pred. zero (pred. negative) is 85.74% and pred one (pred.positive) is 89.19%. The accuracy results are obtained using equation 4, where the true positive values are 836, true negatives are 429, false negatives are 139, and false positives are 52.

$$accuracy = \frac{836 + 429}{836 + 429 + 139 + 52} = 86.88\%$$

Performance Vector:

Tabel 8. Hasil *Performance Vektor KNN*

PerformanceVector:
accuracy: 86.88%
ConfusionMatrix:

True	0	1
0	836	139
1	52	429

Performance Vector is a form of description of the table of analysis results obtained in the research conducted. The True Positive (TP) value has a value of 836, which is a positive data value. The False Positive value is 52, where the data is negative (water is not drinkable) but is predicted as positive data. The False Negative value is 139, positive data but predicted as negative data. The True Negative value is 429, which is negative data that is predicted to be true.

The data classification process uses several operators to carry out classification methods, including CSV reading, data partitioning, model application, and performance. Classification methods such as KNN, Naïve Bayes and Decision Tree. These operators have their respective functions, the CSV read function is to import CSV data that has been obtained, in CSV read mode the preprocessing method is carried out, where the preprocessing function is to display imported data sets, whether there are inconsistent data or missing values. The Split data operator works by taking a set of examples as input and sending a subset of the sample sets through its output port. To use the classification method, use the model features. Performance is used to display the accuracy of all types of classification methods.

Accuracy Results

Comparison of Accuracy Results

Naïve-Bayes	*K-nearest neighbors*	*Decision tree*
63.60%	86.88%	80.84%

Comparative analysis of Water Quality accuracy using data from classification results with K-nearest neighbors, Naïve Bayes, and Decision Tree shows that K-nearest neighbors is the method that produces the highest level of accuracy, namely 86.88% for the classification of quality data used in this study, while Naïve -Bayes is 63.60% and Decision tree is 80.84%.

Taxonomy Table

No	Writer	Research Title	Method	Results
1.	Adi Kusuma, Agung Nugroho, 2021	Sentiment Analysis on Twitter of the Increase in Basic Electricity Rates Using the Naïve Bayes Method	Naïve Bayes	This study attempts to analyze sentiment to see public perception of the issue of increasing basic electricity rates on Twitter social media using the Baïve Bayes method by classifying sentiments into positive, negative and neutral. From the results of research that has been done, it can be seen that the most negative sentiment is formed around 60% in response to the issue of increasing the basic electricity rate.

2	Rani Nooraeni, Aulia Fikri Fadhilah, Heny Dwi, Siti Fatimatul, Suciarti Pertiwi, Yulianus Ronaldias, 2020	Twitter Data Sentiment Analysis Regarding the Issue of the KPK Bill Using the Support Vector Machine (SVM) Method	Support Vector Machine (SVM)	From the original data classification model, training or testing, the percentage of responses in the form of negative sentiment related to the KPK Bill issue was 60.9 percent greater than the percentage of positive sentiment of 39.1 percent. The performance of the SVM model in classifying sentiment is quite good because it has an accuracy, sensitivity and specificity value of 81.32 percent, 71.47 percent and 87.64 percent, respectively.
3	Dianati Duei Putri 1 , Persistent Forda Nama2 , Wahyu Eko Sulistiono, 2022	ANALYSIS OF THE PERFORMANCE SENTIMENT OF THE COUNCIL OF REPRESENTATIVES (DPR) ON TWITTER USING THE NAIVE BAYES CLASSIFIER METHOD	NAIVE BAYES CLASSIFIER	This research uses 1546 data tweets. The results of this study found that the DPR received 95 positive tweets with a polarity of 0.75 or 75% positive sentiment, 693 neutral tweets with a polarity of 0.79 or 79% neutral sentiment and 758 negative tweets with a polarity of 0.82 or 82% negative sentiment with an accuracy score of 0.8 or 80%. based on testing data as much as 20%.

4	Amelia Syahadati1) , Novert Cyril Lengkong2) , Ouditiana Safitri3), Septriyan Machsus4) , Yongki Ramanda Putra5) , Rani Nooraeni	SENTIMENT ANALYSIS OF PSBB IMPLEMENTATION IN DKI JAKARTA AND ITS IMPACT ON JCI MOVEMENTS	JCI, Twitter, Sentiment Analysis	1) conduct an analysis of public sentiment regarding PSBB DKI Jakarta volume II; 2) see the impact of this sentiment on the JCI movement; 3) compare the results of several classification methods, namely logistic regression, k-nearest neighbor, random forest, and naïve Bayes. Scraping Twitter data for the period September 8 - October 9 was carried out using Orange and RStudio software. Furthermore, sentiment analysis with Orange classifies sentiment into positive and negative groups.
5	Puji Nurmawati1, Endang Supriyati2, Tri Listyorini	SENTIMENT ANALYSIS OF KPOP FANS ON TWITTER SOCIAL MEDIA USING NAIVE BAYES (CASE STUDY OF BTS GROUP FANS)	NAIVE BAYES	From the analysis carried out using the Naïve Bayes classification algorithm, there are negative sentiment polarities of 34.2%, 58.5% neutral, and 7.3% positive. Of the 1000 data taken according to the polarity results of the tweets, 342 were negative according to the polarity results. With an accuracy rate of 75%. From this research it is hoped that it can assist in the process of sentiment analysis and is appropriate in overcoming existing problems.
6	Primandani Arsi* 1 , Retno Waluyo	SENTIMENT ANALYSIS OF INDONESIAN CAPITAL REMOVAL DISCOURSE USING SUPPORT VECTOR MACHINE (SVM) ALGORITHM	SUPPORT VECTOR MACHINE (SVM) ALGORITHM	In this study, it is proposed that the Support Vector Machine (SVM) method be applied to tweets on the topic of moving the Indonesian capital city for the purpose of classifying sentiment classes on Twitter social media. Technical classification is done by classifying

				into 2 classes namely positive and negative. Based on the results of tests carried out on tweets on the sentiment of moving the capital city from social media Twitter, as many as 1,236 tweets (404 positive and 832 negative) using SVM obtained accuracy = 96.68%, precision = 95.82%, recall = 94.04% and AUC = 0.979.
7	Angelina Puput Giovani1), Ardiansyah2), Tuti Haryanti3), Laela Kurniawati4) , Windu Gata	ANALYSIS OF SENTIMENT OF GURU APPLICATION ON TWITTER USING CLASSIFICATION ALGORITHM	CLASSIFICATION ALGORITHM	This study compares the NB, SVM, K-NN methods without using feature selection with the NB, SVM, K-NN methods that use feature selection and compares the Area Under Curve (AUC) values of these methods to find out the most optimal algorithm. The test results show that the best optimization application in this model is the SVM-based PSO algorithm with an accuracy value of 78.55% and an AUC of 0.853. This research succeeded in obtaining the best and most effective algorithm for classifying positive comments and negative comments related to the Ruang Guru application.
8	Afif Nor Yusuf 1 , Endang Supriyati 2 , Tri Listyorini	Sentiment Analysis Regarding Indihome Service Providers Based on Customer Opinions Through Social Media Twitter with the Naïve Bayes Classifier Method	Naïve Bayes Classifier	The results of the Naïve Bayes method are very good. To test the level of accuracy of the system in classifying opinions, so that the test obtains classification results. The results of the classification obtain an average yield of 74.5%. The more training data that is similar to the

				testing data, the better the classification results will be.
9	Yan Wategulis Syaifudin 1, Rizki Andi Irawan	IMPLEMENTATION OF CLUSTERING ANALYSIS AND TWITTER DATA SENTIMENT ON BEACH TOURISM OPINIONS USING K-MEANS METHOD	K-MEANS	The accuracy of the classification using the Support Vector Machine algorithm is 74.39%. Furthermore, opinion data from the questionnaire was added to classify beaches based on the availability of resources, facilities, access, community readiness, market potential and tourism position. In the process of grouping this data, the K-Means method is used.
10	Imam Kurniawan 1, Ajib Susanto	Implementation of the K-Means and Naïve Bayes Classifier Methods for Sentiment Analysis for the 2019 Presidential Election (Pilpres)	K-Means and Naïve Bayes Classifier	The purpose of this study is to obtain an analysis of text documents to obtain positive or negative sentiments. The method used is K-Means for clustering the training data and the Naive Bayes classifier for classifying the testing data. The results of this weighting are in the form of positive and negative sentiments. The data was taken from Twitter regarding the 2019 presidential election as many as 500 tweet data. From the test results of 100 and 150 test data obtained an

				average accuracy of 93.35% and an error rate of 6.66%.
11	Sigit Suryono, Em a Utami, Em ha Taufiq Luthfi	SENTIMENT CLASSIFICATION IN TW ITTER WITH NAIVE BAYES CLASSIFIER	NAIVE BAYES CLASSIFIER	From the results of the 3 trials, the accuracy rate in the first trial was 64.95%, second 66.36% and third 66.79%. Other results obtained from the classification process were positive sentiment 28% negative sentiment 20% and neutral sentiment 52%. Based on the results of the sentiment class percentage, neutral sentiment is the most common sentiment when it comes to the topic of President Joko Widodo and his government.
12	Tati Mardiana1; Hafiz Syahreva2; Tuslaela	COMPARISON OF CLASSIFICATION METHODS ON FRANCHISING BUSINESS SENTIMENT ANALYSIS BASED ON TWITTER DATA	Sentiment, Python, Twitter, Comparison.	The test results with the confusion matrix obtained an accuracy value of 83% for Neural Network, 52% for K-Nearest Neighbor, 83% for Support Vector Machine, and 81% for Decision Tree. This research shows that the Support Vector Machine and Neural Network methods are the best for classifying positive and negative comments related to franchising.

13	Dedi Darwis 1, Eka Shintya Pratiwi 2, A. Ferico Octaviansyah Pasaribu	APPLICATION OF SVM ALGORITHM FOR SENTIMENT ANALYSIS ON CORRUPTION ERADICATION COMMISSION TWITTER DATA OF THE REPUBLIC OF INDONESIA	SVM ALGORITHM	This research produced 1890 data and 3846 terms/words from the preprocessing results and then calculated the value of the appearance of the word for labeling which resulted in positive, negative and neutral sentiments. Based on the test results generated, the application of the SVM method produces an accuracy value of 82% and produces sentiment with a greater negative label with a total of 77%, 8% positive label and 25% neutral label.
14	Fira Fathonah1), Asti Herliana	The Application of Sentiment Analysis Text Mining Regarding the Covid - 19 Vaccine Using the Naïve Bayes Method	Naïve Bayes	Naïve Bayes is considered to have good potential in classifying documents compared to other classification methods in terms of accuracy and efficiency. Based on the results of testing 100 training data which were then re-selected using data crawling techniques into 34 data, it was found that sentiment analysis from Twitter users for the COVID-19 vaccine This obtained an accuracy percentage of 100%
15	Ragil Dimas Himawan # 1 , Eliyani	Comparison of the Accuracy of Tweet Sentiment Analysis for the Provincial Government of DKI	Support Vector Machine, Naïve Bayes, Random Forest Classifier	The data obtained is 14208 lines by querying tweets containing the word or mentioning the username @dkijakarta, which will be grouped by sentiment class, namely negative,

		Jakarta during the Pandemic Period		neutral and positive using the TF-IDF Vectorizer for word weighting and classification using several methods, namely random. forest classifier with 75.81% accuracy, naive Bayes algorithm with 75.22% accuracy, and support vector machine algorithm 77.58%. A sentiment analysis process was carried out on tweets with the percentage of negative, neutral and positive results, respectively, namely, 8.8%, 83.6%, 7.6%.

CONCLUSION

The purpose of this study is to find out the results of the accuracy comparison of the research methods used, namely K-nearest neighbor, Naïve Bayes and Decision Tree. Judging from Class Recall and Class Precision, the method that provides the highest level of precision is the decision tree which is equal to 86.88%. The Decision Tree and KNN classification methods in this study were used quite well because they produced an accuracy rate above 80%, but other methods can be used to obtain maximum accuracy results for further research.

BLIBLIOGRAPHY

Cahyanti, D., Rahmayani, A., & Husniar, SA (2020). Analysis of the performance of the Knn method on the dataset of patients with breast cancer. *Indonesian Journal of Data and Science* , *1* (2), 39–43.

Fikri, MI, Sabrila, TS, & Azhar, Y. (2020). Comparison of the Naïve Bayes method and the support vector machine for Twitter sentiment analysis. *SMATIKA Jurnal: STIKI Informatika Jurnal* , *10* (02), 71–76.

Hope, PN (2019). Implementation of Data Mining in Predicting Sales Transactions Using the Apriori Algorithm (Case Study of PT. Arma Anugerah Abadi Branch of Sei Rampah). *MATICS: Journal of Computer Science and Information Technology (Journal of Computer Science and Information Technology)* , *11* (2), 46–50.

Jati, NP (2021). *S Integration of Kansei Engineering and Kano Based on Natural Language Processing (Nlp) to Support the Development of Service Products in Borobudur Temple Tourism* .

Lestari, UI, Nadhiroh, AY, & Novia, C. (2021). Application of the K-Nearest Neighbor Method for a Decision Support System for the Identification of Diabetes Mellitus. *JATISI (Journal of Informatics Engineering and Information Systems)* , *8* (4), 2071–2082.

Muningsih, E. (2022). Combination of K-Means and Decision Tree Methods with Comparison of Criteria and Split Data. *Teknoinfo Journal* , *16* (1), 113–118.

Pamuji, FY, & Ramadhan, VP (2021). Comparison of Random Forest and Decision Tree Algorithms for Predicting Immunotherapy Success. *Journal of Information Technology and Management* , *7* (1), 46–50.

Puspita, R., & Widodo, A. (2021). Comparison of the KNN, Decision Tree, and Naïve Bayes Methods on the Sentiment Analysis of BPJS Service Users. *J. Inform. Univ. Pamulang* , *5* (4), 646.

Rahman, MA, Hidayat, N., & Supianto, AA (2018). Comparison of K-Nearest Neighbor and Naïve Bayes Data Mining Methods for Clean Water Quality Classification (Case Study of PDAM Tirta Kencana, Jombang Regency). *Journal of Information Technology*

Development and Computer Science , *2* (12), 6346–6353.

Riadi, I., & Kom, M. (2017). *Analysis of Digital Evidence of Cyberbullying in Social Networks Using the Naïve Bayes Classifier (NBC)* .

Romadloni, NT, Santoso, I., & Budilaksono, S. (2019). Comparison of the Naive Bayes, Knn and Decision Tree Methods on Sentiment Analysis of KRL Commuter Line Transportation. *IKRA-ITH Informatics: Journal of Computers and Informatics* , *3* (2), 1–9.

Sartika, D., & Sensuse, DI (2017). Comparison of the Naive Bayes, Nearest Neighbor, and Decision Tree classification algorithms in case studies of clothing pattern selection decision making. *JATISI (Journal of Informatics Engineering and Information Systems)* , *3* (2), 151–161.

Siregar, RRA, Siregar, ZU, & Arianto, R. (2019). Classification of Sentiment Analysis on Training Participants' Comments Using the K-Nearest Neighbor Method. *The Flash* , *8* (1), 81–92.

Sukmana, RN, Abdurrahman, A., & Wicaksono, Y. (2020). Implementation of K-Nearest Neighbor to Determine Sales Predictions: (Case Study: Pt Maksiplus Utama Indonesia). *Journal of Information and Communication Technology* , *9* (2), 31–37.

Wahyuningsih, S., & Utari, DR (2018). Comparison of the K-Nearest Neighbor, Naive Bayes and Decision Tree Methods for Predicting Creditworthiness. *Information Systems National Conference (KNSI) 2018* .

Zulfauzi, Z., & Alamsyah, MN (2020). Application of the Naive Bayes Algorithm for Predicting New Student Admissions Case Study at Bina Insan University, Faculty of Computers. *Journal of Information Technology Mura* , *12* (2), 156–165.

Journal of Syntax Admiration